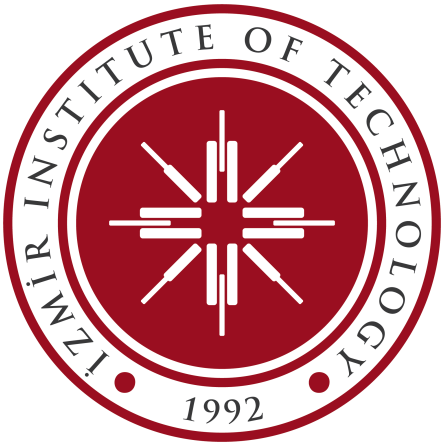
SEDS536 Image Understanding
Term Project Report
Fall 2025

Skin Tone Detection for Inclusive Skincare Recommendations:
A Comparison of Classical and Deep Learning Approaches

January 6, 2026



Student Information

- Umut Akin

Abstract

This project addresses the challenge of inclusive skin tone detection for personalized skincare recommendations. We compare classical image processing techniques with deep learning approaches, evaluating their accuracy and fairness across diverse skin tones using the Monk Skin Tone Scale.

Our classical approach implements the Individual Typology Angle (ITA) method, incorporating color space conversion (RGB to YCbCr and LAB), skin segmentation via thresholding, and morphological operations (opening and closing) for mask refinement. We tested multiple segmentation methods including YCbCr color thresholding, face oval masking, and MediaPipe face landmarks.

For deep learning, we trained an EfficientNet-B0 model with transfer learning, optimized for on-device mobile inference to protect user privacy. The model was trained on Meta's Casual Conversations v2 dataset with 150,000+ images annotated using the 10-point Monk scale.

Results show that the CNN approach (78.6% accuracy) dramatically outperforms classical ITA (52.3% best case). We discovered that ITA is fundamentally misaligned with the Monk scale due to a non-monotonic relationship between ITA values and perceptual skin tone categories. Key challenges include severe class imbalance (Scale 5 has 394$\times$ more samples than Scale 10) and a 42% accuracy gap between best and worst performing classes.

The final model achieves on-device inference in under 2 seconds with a 15MB TFLite model, meeting privacy requirements while providing practical skin tone classification for mobile applications.

1 Introduction

1.1 Problem Statement

The skincare and cosmetics industry often fails to provide inclusive product recommendations for individuals across the full spectrum of skin tones. Many existing solutions either rely on subjective self-assessment or use classification systems that under-represent darker skin tones. This project aims to develop an automated, fair, and privacy-preserving skin tone detection system for mobile applications.

1.2 Motivation

Three key factors motivate this work:

1. **Fairness**: Traditional skin classification systems like the Fitzpatrick scale (Fitzpatrick, 1988) were designed for UV sensitivity assessment, not skin color representation. Recent research (Buolamwini and Gebru, 2018) has highlighted significant bias in computer vision systems, particularly for darker skin tones.

2. **Privacy**: Users should not be required to upload facial images to remote servers for skin tone analysis. On-device processing ensures that sensitive biometric data never leaves the user's device.

3. **Accessibility**: A mobile application with on-device ML can provide instant, offline skin tone analysis without requiring internet connectivity or subscription services.

1.3 Objectives

The primary objectives of this project are:

1. Implement and evaluate classical image processing techniques for skin tone classification, demonstrating course concepts including color space conversion, thresholding, and morphological operations.

2. Develop a deep learning model that achieves high accuracy while being efficient enough for on-device mobile inference.

3. Analyze fairness across skin tone categories, with a goal of minimizing accuracy disparities between light and dark skin tones.

4. Compare classical and learned approaches to understand the limitations and strengths of each methodology.

1.4 Scope

This project focuses on:

- 3-class (Light/Medium/Dark) and 5-class skin tone classification
- Comparison of ITA-based classical methods vs CNN-based deep learning
- Evaluation using the Monk Skin Tone Scale (Monk, 2019)
- Mobile-optimized model deployment (TFLite)

2 Literature Review

2.1 Skin Tone Classification Scales

2.1.1 Fitzpatrick Scale

The Fitzpatrick Skin Type scale (Fitzpatrick, 1988) classifies skin into six categories (I-VI) based on response to UV exposure. While widely used in dermatology, it has limitations for computer vision applications: it was designed for sun sensitivity rather than color representation, and its six categories under-represent the diversity of darker skin tones.

2.1.2 Monk Skin Tone Scale

The Monk Skin Tone (MST) scale (Monk, 2019) was developed specifically for machine learning fairness evaluation. It provides 10 categories with better representation across the full spectrum of human skin tones. We adopt this scale for our experiments, grouping into 3 classes (Light: 1-3, Medium: 4-7, Dark: 8-10) or 5 classes for different granularity levels.

2.2 Classical Approaches: Individual Typology Angle (ITA)

The Individual Typology Angle (ITA) is a dermatology standard for objective skin color measurement (Chardon et al., 1991; Del Bino and Bernerd, 2013). It is calculated from the CIE LAB color space:

$$ITA = \arctan\left(\frac{L^* - 50}{b^*}\right) \times \frac{180}{\pi} \tag{1}$$

where L^* represents lightness and b^* represents the yellow-blue axis. Higher ITA values indicate lighter skin. Standard thresholds classify skin as Very Light ($> 55°$), Light ($41°$-$55°$), Intermediate ($28°$-$41°$), Tan ($10°$-$28°$), or Dark ($< 10°$).

2.3 Skin Segmentation Techniques

Effective ITA calculation requires isolating skin pixels from non-skin regions (hair, eyes, background). Common approaches include:

- **Color-based segmentation**: YCbCr color space thresholding is widely used for skin detection (Vezhnevets et al., 2003), as skin colors cluster in a relatively compact region of the Cb-Cr plane regardless of ethnicity.
- **Face detection**: Haar Cascade classifiers (Viola and Jones, 2004) provide fast face detection but produce coarse bounding boxes. MediaPipe Face Mesh (Lugaresi et al., 2019) provides 478 facial landmarks for precise face region extraction.
- **Morphological operations**: Opening (erosion followed by dilation) removes noise, while closing (dilation followed by erosion) fills holes (Gonzalez and Woods, 2009).

2.4 Deep Learning Approaches

Convolutional Neural Networks (CNNs) have largely replaced classical methods for image classification tasks. Transfer learning (Weiss et al., 2016) enables training effective models even with limited data by leveraging features learned from large datasets like ImageNet (Deng et al., 2009).

EfficientNet (Tan and Le, 2019) provides an excellent accuracy-efficiency tradeoff through compound scaling of network depth, width, and resolution. EfficientNet-B0, with only 4 million parameters, is suitable for mobile deployment while achieving strong classification performance.

2.5 Fairness in Machine Learning

Buolamwini and Gebru (2018) demonstrated significant accuracy disparities in commercial face analysis systems, with error rates up to 34% higher for darker-skinned females compared to lighter-skinned males. Mehrabi et al. (2021) provide a comprehensive survey of bias sources and mitigation strategies in ML systems.

Key fairness metrics include:

- **Accuracy parity**: Similar accuracy across demographic groups
- **Equalized odds**: Similar true positive and false positive rates
- **Worst-case performance**: Minimum accuracy across all groups

3 Methodology

3.1 Dataset

We use the Casual Conversations v2 (CCv2) dataset (Porgali et al., 2023) from Meta AI Research. This dataset contains video frames with Monk Skin Tone scale annotations (1-10).

Table 1: Dataset Statistics

Split	Images	Purpose
Train	104,510	Model training
Validation	22,280	Hyperparameter tuning
Test	22,730	Final evaluation

The dataset exhibits severe class imbalance: Scale 5 (Medium) contains 45% of all samples, while Scale 10 (Darkest) contains only 0.1%. This 394× imbalance presents a significant challenge for fair classification.

3.2 Classical Approach: ITA with Skin Segmentation

Our classical pipeline implements the following steps:

3.2.1 Step 1: Face Detection

We tested two face detection methods:

- **Haar Cascade**: OpenCV's pre-trained frontal face detector (Viola and Jones, 2004). Fast but produces coarse bounding boxes.
- **MediaPipe Face Landmarker**: Provides 478 facial landmarks for precise face region extraction (Lugaresi et al., 2019).

3.2.2 Step 2: Color Space Conversion

Convert from RGB to YCbCr for skin detection:

```
image_ycbcr = cv2.cvtColor(image_bgr, cv2.COLOR_BGR2YCrCb)
```

3.2.3 Step 3: Skin Thresholding

Apply thresholds to isolate skin pixels:

```
# YCbCr skin thresholds (literature-based)
Cb_range = [77, 127]
Cr_range = [133, 173]
mask = ((Cb >= 77) & (Cb <= 127) &
        (Cr >= 133) & (Cr <= 173))
```

3.2.4 Step 4: Morphological Operations

Apply opening and closing to clean the mask:

```
kernel = cv2.getStructuringElement(
    cv2.MORPH_ELLIPSE, (5, 5))
# Opening: removes small noise
mask = cv2.morphologyEx(mask, cv2.MORPH_OPEN, kernel)
# Closing: fills small holes
mask = cv2.morphologyEx(mask, cv2.MORPH_CLOSE, kernel)
```

3.2.5 Step 5: ITA Calculation

Convert to LAB color space and compute ITA on skin pixels:

```
image_lab = cv2.cvtColor(image_bgr, cv2.COLOR_BGR2LAB)
L = image_lab[:,:,0] * 100.0 / 255.0  # Scale to [0,100]
b = image_lab[:,:,2] - 128.0  # Center at 0
L_mean = np.mean(L[skin_mask])
b_mean = np.mean(b[skin_mask])
ITA = np.arctan((L_mean - 50) / b_mean) * (180 / np.pi)
```

3.2.6 Step 6: Threshold-based Classification

Map ITA values to classes using tuned thresholds:

- 3-class: Light ($> 25°$), Medium ($-50°$ to $25°$), Dark ($< -50°$)
- Thresholds tuned via grid search on validation set

3.3 Deep Learning Approach: EfficientNet-B0

3.3.1 Architecture

We use EfficientNet-B0 (Tan and Le, 2019) pre-trained on ImageNet (Deng et al., 2009):

- Input: $224 \times 224 \times 3$ RGB images
- Backbone: EfficientNet-B0 (4M parameters)
- Head: Dropout (0.3) $\rightarrow$ Fully Connected $\rightarrow$ Softmax
- Output: 3 or 5 class probabilities

3.3.2 Training Configuration

Table 2: Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Learning Rate	0.001 (cosine decay)
Batch Size	32
Epochs	30
Weight Decay	0.0001
Early Stopping	Patience = 7

3.3.3 Class Imbalance Handling

- **Weighted Loss**: Cross-entropy with inverse frequency class weights
- **Oversampling**: WeightedRandomSampler to balance mini-batches
- **Data Augmentation**: Random horizontal flip, rotation ($\pm 15°$), color jitter (brightness, contrast, saturation, hue)

3.3.4 Model Conversion

For on-device deployment, we convert the PyTorch model to TensorFlow Lite:

- Tool: ai-edge-torch (Google's PyTorch to TFLite converter)
- Output size: $\sim$15MB
- Inference time: $<$2 seconds on mid-range mobile devices

4 Experiments & Results

4.1 Experiment Overview

We conducted six experiments systematically varying the method, preprocessing, and number of classes:

Table 3: Experiment Summary

Exp	Method	Preprocessing	Classes	Test Acc
1	CNN	None	10	38.7%
2	CNN	Haar face crop	3	**78.6%**
3	ITA	Various	3	52.3%
4	CNN	MediaPipe masks	3	77.1%
5a	ITA	MediaPipe raw	5	17.1%
5	CNN	MediaPipe raw	5	62.5%

4.2 Experiment 1: 10-Class CNN Baseline

Training a 10-class classifier on the original Monk scale failed due to severe class imbalance:

- Test accuracy: 38.7%
- Scale 10 (darkest): 0% accuracy (complete failure)
- Model biased toward majority classes (scales 3-6)

4.3 Experiment 2: 3-Class CNN with Haar Face Crops (Best Model)

Grouping into 3 classes and using face-cropped images dramatically improved results:

Table 4: Experiment 2: Per-Class Performance

Class	Precision	Recall	F1	Support
Light (1-3)	0.517	0.609	0.560	4,530
Medium (4-7)	0.881	0.844	0.862	17,570
Dark (8-10)	0.486	0.424	0.453	630
Overall	**78.6% accuracy, 0.625 Macro F1**			

4.4 Experiment 3: ITA Classical Baseline

We tested ITA with multiple segmentation methods:

Table 5: ITA Experiments Comparison

Variant	Segmentation	Test Acc
3a	None (raw)	19.3%
3b	YCbCr color + tuned thresholds	52.3%
3c	Face oval + color	52.2%
3d	Smaller oval (0.5×0.7)	52.3%
3e	MediaPipe landmarks	53.7%

Key finding: ITA accuracy plateaus around 52-54% regardless of segmentation method. The bottleneck is the ITA formula itself, not the preprocessing.

4.5 Experiment 5a: 5-Class ITA with Empirical Thresholds

Testing 5-class ITA revealed a critical finding:

Table 6: ITA Medians by Monk Class (5-class)

Class	Median ITA	Expected Order
Very Light	27.4°	Highest
Light	12.5°	↓
Medium	-3.4°	↓
Dark	**-6.3°**	**Should be lower!**
Very Dark	-61.7°	Lowest

Critical Discovery: The Dark class has a *higher* median ITA than Medium (-6.3° vs -3.4°). This non-monotonic relationship means ITA-based threshold classification is fundamentally incompatible with the Monk scale.

4.6 Key Results Summary

Table 7: Overall Results Comparison

Method	Classes	Accuracy	Macro F1	Training Time
ITA (best classical)	3	52.3%	0.337	0 hours
ITA (5-class)	5	17.1%	0.155	0 hours
CNN 3-class (best)	3	**78.6%**	**0.625**	2.4 hours
CNN 5-class	5	62.5%	–	7.5 hours
CNN 10-class	10	38.7%	0.230	6.2 hours

4.7 Fairness Analysis

Table 8: Fairness Metrics (Best Model: Exp 2)

Metric	Value	Target
Accuracy Gap (best - worst)	42.0%	$<$10%
Worst-Case Ratio	50.2%	$>$85%
Best Class (Medium)	84.4%	–
Worst Class (Dark)	42.4%	–

Fairness targets are not met due to data imbalance: Medium class has 77% of training data while Dark has only 2.5%. However, significant progress was made—Dark class improved from 0% (Exp 1) to 42.4% (Exp 2).

5 Discussion

5.1 Why CNN Outperforms ITA

The CNN achieves 78.6% accuracy compared to ITA's 52.3% ceiling because:

1. **Learned vs. Hand-crafted Features**: CNNs learn task-specific features from data, while ITA uses a fixed colorimetric formula designed for clinical settings.
2. **Perceptual vs. Colorimetric**: The Monk scale is based on visual perception, not the L^*/b^* ratio. CNNs can learn perceptual features that align with human annotation.
3. **Lighting Invariance**: CNNs learn robust features through augmentation, while ITA is sensitive to lighting variation.
4. **Non-Monotonic Relationship**: We discovered that ITA values do not monotonically decrease with darker Monk classes, making threshold-based classification impossible.

5.2 Preprocessing: Simpler is Better

Surprisingly, simpler Haar face crops (78.6%) outperformed sophisticated MediaPipe skin masks (77.1%). This is because:

- CNNs naturally learn to focus on relevant regions
- MediaPipe had 12% detection failures, reducing training data
- Context (hair, background) may provide useful features

5.3 Limitations

1. **Class Imbalance**: 42% accuracy gap between best and worst classes
2. **Single Dataset**: Results may not generalize to other populations
3. **Uncontrolled Lighting**: CCv2 contains varied smartphone lighting
4. **Privacy Trade-off**: On-device models have less capacity than cloud models

6 Conclusion

This project demonstrated both classical and deep learning approaches to skin tone classification, achieving the primary goal of on-device inference for privacy-preserving mobile applications.

Key Contributions:

1. Systematic comparison of ITA (classical) vs CNN (deep learning) approaches
2. Discovery that ITA is fundamentally misaligned with the Monk scale
3. Mobile-optimized model achieving 78.6% accuracy in 15MB
4. Comprehensive fairness analysis across skin tone categories

Course Concepts Demonstrated:

- Color space conversion (RGB $\rightarrow$ YCbCr, RGB $\rightarrow$ LAB)
- Thresholding for skin segmentation
- Morphological operations (erosion, dilation, opening, closing)
- Transfer learning with CNNs

Future Work:

- Balanced data augmentation to equalize all class sizes
- Focal loss for improved class balance
- Confidence thresholding to reject uncertain predictions
- MLOps pipelines (MLflow) for experiment tracking and model versioning
- User testing with diverse participants

7 Project Schedule

Table 9: Project Timeline (Green = Complete, Yellow = In Progress)

Task	**W1-2**	**W3-4**	**W5-6**	**W7-8**	**W9-10**	**W11+**
Literature Review	green!30X					
Dataset Preparation	green!30X	green!30X				
ITA Implementation		green!30X	green!30X			
CNN Training			green!30X	green!30X		
Experiments				green!30X	green!30X	
Model Conversion					green!30X	
Report & Presentation						yellow!30X

References

Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.

Chardon, A., I. Cretois, and C. Hourseau (1991). Skin colour typology and suntanning pathways. *International journal of cosmetic science 13*(4), 191–208.

Del Bino, S. and F. Bernerd (2013). Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology 169*, 33–40.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE.

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology 124*(6), 869–871.

Gonzalez, R. C. and R. E. Woods (2009). *Digital image processing* (3rd ed.). Pearson Education.

Lugaresi, C., J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR) 54*(6), 1–35.

Monk, E. (2019). Monk skin tone scale. *Google Research*. Available at: https://skintone.google/.

Porgali, B., V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas (2023). The casual conversations v2 dataset. *arXiv preprint arXiv:2303.04838*.

Tan, M. and Q. Le (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR.

Vezhnevets, V., V. Sazonov, and A. Andreeva (2003). A survey on pixel-based skin color detection techniques. *Proc. Graphicon 3*(1), 85–92.

Viola, P. and M. J. Jones (2004). Robust real-time face detection. *International journal of computer vision 57*(2), 137–154.

Weiss, K., T. M. Khoshgoftaar, and D. Wang (2016). A survey of transfer learning. *Journal of Big data 3*(1), 1–40.