Enhancing Fairness in Machine Learning: Skin Tone Classification Using the Monk Skin Tone Scale

Vitor Pereira Matias and João Batista Neto
Institute of Mathematics and Computer Science
University of São Paulo
São Carlos, SP, Brazil
Email: vitorpmatias@usp.br, jbatista@usp.br

***Abstract*—In the machine learning era, unethical errors from poorly curated datasets are a pressing issue, especially in fields related to skin tone recognition in which imbalanced datasets lead to biased results. Developing a skin tone classification algorithm helps identify such imbalances. Existing methods range from classic computer vision pipelines to deep learning CNNs that typically employs controlled environment datasets with limited class diversity (two to six classes). Our work focuses on classifying skin tones using the 10-class Monk Skin Tone (MST) scale. To this end, we created the SkinTone in The Wild (STW) dataset by merging well-known face recognition datasets and labelling it according to the MST. This dataset comprises 39,605 images of 2,183 individuals, mostly captured in uncontrolled environments. To overcome this scenario, we evaluated different approaches which resulted in 74% accuracy and 92% off-by-one accuracy (OOAcc) with a RandomForest model, and 68% accuracy along with 86% OOAcc using a DenseNet121 CNN. Furthermore we discussed the sheer power of CNNs and showed that the DenseNet121 architecture learned to predict skin tones by focusing on the background of images. These results highlight the potential for accurate skin tone classification in machine learning which leads to better curated datasets.**

I. Introduction

AI fairness assessment has emerged as a prominent field of study in recent years due to perceived poor judgements, which may be caused by models being trained upon poorly curated datasets with underlying imbalances among classes. There have been multiple reports of AI's misdemeanours such as racist or sexist behaviours, or models that do not recognise people with darker skin [1]–[4]. To mitigate this problem one should start by employing a balanced and well curated dataset. For example, consider the face recognition related tasks, open datasets such as CelebA [5] and LFW [6] should be balanced in skin tone, biological gender, ethnicity and other features. Consequently, there is a demand for algorithms that classifies skin tones to be used to identify imbalances while creating such datasets. Moreover, there are potential applications in the field of social sciences, cosmetics, visagism, and makeup applications.

Skin tone classification is the process of taking a picture of an individual that contains skin and classify it according to a given scale. Examples of scales are the Fitzpatrick Skin Tone (FST) [7] used by the medical community, and the novel Monk Skin Tone (MST) [8] which is the one that we adopt in our work. As opposed to the 6-tone FST scale, the MST scale suggests 10 distinct skin tones, which makes it more representative to skin tones worldwide [9].

Traditional computer vision approaches to skin tone classification have already been proposed. Some of them extract features in the form of histograms or moments [10]–[14], fuzzy methods [15], or K-means [16]. None of the cited methods take into account important issues that affect skin colour classification such as lighting, skin redness or many other issues that arise when dealing with facial pictures. Other approaches employ Convolutional Neural Networks (CNNs) as a classifier [10], [17], [18] or a regressor [19]. Our work comprises models from both strategies aiming to predict skin tones with the MST scale.

However, most of the works cited above use labelled datasets with 6 classes or less, either the FST or handcrafted scales. Moreover, the number of images never exceeded 10,000 [10], with most datasets containing fewer than 3,000 samples [12], [14], [15], [17], [19]. Exceptions can be found in three works that employ the 17k Fitzpatrick dataset of skin diseases: Kinyanjui [11] classifies diseases only, while Groh [13] and Tadesse [18] also address skin tone classification. Additionally, some of the works use datasets acquired in controlled environments or use high tier camera sensors [14], [17], [19], [20].

Datasets that contains skin tones are difficult to obtain and may come intertwined with skin diseases [13], [21]. However, skin diseases datasets are not ideal for skin tone identification, as they prioritise diseased areas, posing an additional challenge of segmenting relevant non-diseased regions. They also contain skin from distinct anatomical body parts. While this may contribute to improve model generalisation, they are not necessarily suitable for skin tone identification. We believe actual skin tone identification should come primarily from skin tones of human faces. Face images convey ethnicity features which may help humans define skin tones as well as facilitate the task of human data annotation, even in adverse lighting conditions.

Currently, telling a person skin tone can be considered an arbitrary and subjective challenge, for which no agreement has been reached among trained annotators. Studies have shown that agreement rates among annotators can be as low as 26% for Fitzpatrick Skin Type IV [10], [13].

With that in mind, one of the main objectives of this work is to create a new balanced dataset of human facial skin tone. The dataset will be annotated from scratch using an interface from which an annotator can assign a skin tone label to all images of a single individual. By doing so, we can speed up the annotation process. After all annotation is performed, we employed two types of models: a) a classic computer vision (CCV) model containing preprocessing, detectors, feature extraction and classification and b) a CNN-based to perform classification.

Additionally, for CNNs, we employed two types of inputs: (1) segmented-skin face region and (2) full-image, to test how CNNs would perform. Our work showed that CNNs and CCVs performed similarly, achieving around 70% accuracy and 90% off-by-one accuracy. Lastly, we performed Grad-Cam visualisations [22] to identify which areas of the input image were important to our CNN model. This approach revealed that using full-image inputs led the CNNs to learn background features, which hindered their ability to generalise in skin tone classification.

II. Dataset

We created the **SkinTone in The Wild** (STW) dataset of facial pictures annotated based on the 10-tone Monk Skin Tone scale. It consists of a combination of images from the following datasets: a) CelebA; b) Labelled Faces in The Wild (LFW); c) Casia Face Africa (CFA); d) Casia Face V5 (CFV5); e) Brazilian Face Dataset (FEI); f) Faces94/95 and g) the Feret dataset. Two other datasets are expected to be added shortly: UTK Face and MORPH [23]–[29].

The aforementioned datasets, except UTK Face and MORPH, organise and group images by individuals. This setup reduces the time required for skin tone annotation, as we label the skin tone of each individual rather than each image separately.

The CFA dataset is the only to contain a reasonable amount of people from MST skin tones 7 to 10, i.e., people with darker skin tones. Hence, its inclusion contributed to reduce the imbalance generally seen in many facial skin tone datasets, normally formed by images of people with lighter skin tones. On the other hand, CelebA and LFW were chosen because they are often used in many related works and, when combined, provide more than 200k images. Since both CelebA and LFW do not have diversity of skin tones (mostly MST skin tones 2 and 3), we added the FEI and Feret datasets that have more diversity of faces from MST 1 to 7.

After collecting all the data, we performed the annotation process using our skin tone labelling interface, depicted in Figure 1. This interface has been designed so that every picture of a single individual is assigned the same skin tone label, no matter the variations of age, tanning, oiliness or lighting conditions.

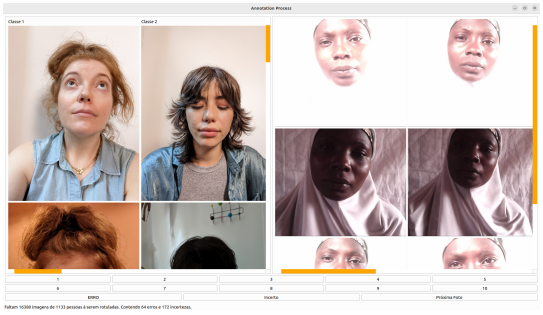


Fig. 1. Labelling interface. The gold standard images from MST (left) and images to be annotated (right)

The interface shows, on the left, 19 gold standard images representing the 10 skin tones of MST. These images have been taken in three different contexts: (1) perfect lighting and pose; (2) dark lighting; and (3) faces covered with objects. On the right, the images of the individuals to be labelled are shown. At the bottom, the annotator must choose one out of three possible choices of labels to assign: a) MST scale 1 to 10; b) not sure and c) error. Option (a) is a straightforward, when the annotator is confident about the label to choose. Option (b) should be selected when the annotator could not figure out which skin tone to select. On the other hand, option (c) or "error", is the choice when the target person could not be associated with any of the gold standard images.

A manual was also provided to annotators, explaining not only the steps of the annotation process but also the fundamentals of the MST scale. Figure 2 depicts images samples from MST for each of the 10 skin tones and their respective circular colour palette.

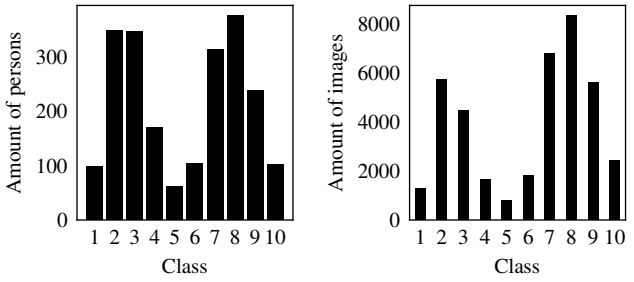
Fig. 2. MST scale gold standard pictures.

At the time of publication, the STW dataset contains 39605 annotated images of 2183 persons. Figure 4a depicts the distribution of individuals per class, while Figure 4b illustrates the number of images per class. There is a clear imbalance among classes, especially for the central classes. This is primarily due

(a) (b)

Fig. 3. Example of inputs. (a) Full-image. (b) Segmented-skin Region of Interest.

to the combination of the datasets CFA, LFW, Faces94/95 and FEI. The first (CFA) mostly contains darker-skinned people, while the second and third datasets have predominantly lighter-skinned samples. The fourth dataset (FEI) contains samples from the middle classes, but in small numbers. However, by adding more datasets with middle class samples, which include more brown and yellow skin tones, we can establish a balanced dataset.



(a) Amount of persons per class. (b) Amount of images per class

Fig. 4. Dataset Analysis

III. Proposed Classification Models

This work proposes two classification models: (1) a Classic Computer Vision (CCV) pipeline, which includes preprocessing, segmentation, feature extraction, and classification; and (2) a Convolutional Neural Network (CNN) architecture. For the CNN-based model, we fine-tuned the fourth block of the DenseNet121 network using Imagenet weights, and trained two CNNs from scratch, both derived from the work of Rachmadi [30]. As they will be later referred, we named them VehicleNet and VehicleNet revisited. VehicleNet is the original network with concatenated parallel layers, while VehicleNet Revisited is the VehicleNet with the first layer changed from a 3D convolution to a 2D convolution. We chose the VehicleNet as it was originally aimed at classifying colours of cars. On the other hand, we chose DenseNet as it connects initial layers to following layers, hence, it may transfer colour information.

For the CCV model, we have tested features such as: Colour Coherence Vectors, Histograms, Global colour Histogram, Colour Statistical Moments and Border/Interior Classification. Histograms of the colour channels RGB, Y (from YCbCr), V (from HSV), and L (from Lab) were empirically found to be the most relevant descriptors for classification which was performed with KNN, MLP, SVMs, and others. The CCV model was trained on segmented input and did not require any preprocessing. Here we did not use class weights as a mean of regularisation.

Our CNN models are straightforward applications of DenseNet121 and VehicleNet, except that we tested two different loss functions: Cross Entropy (CE) and Ordinal Cross Entropy (OCE) (Eq. 1). This is due to the fact that our data is both categorical and ordinal. Hence, the Cross Entropy Loss alone would not be appropriate to discuss results using metrics such as off-by-one accuracy and MSE. In our tests, the CE loss was applied with class weighting to improve regularisation [31].

The OCE is a simple extension of the Cross Entropy Loss (CE) by multiplying it by an error distance factor. Using k as the number of classes, y the ground truth class, $\hat{y}$ the predicted class, we have:

$$\mathrm{OCE}(y, \hat{y}) = \left(1 + \frac{\|y - \hat{y}\|}{k}\right) \mathrm{CE}(y, \hat{y}). \tag{1}$$

A. Experiments

We adopted two different strategies to evaluate our models: (a) the use of full-image and (b) segmented-skin as seen in Figure 3. To segment the face skin region, we employed Google's Mediapipe Python library [32], an AI-based solution trained on datasets that contains facial skin tones which may be imbalanced, mostly with lighter skin tones. Hence, it is bound to fail in recognising many individuals of different skin tone. Figure 5 depicts the amount of images per class that Mediapipe could not recognise even when individuals in the picture are clearly recognisable. Additionally, we must point out that a considerable amount of images have not been recognised because they contain poor lighting, face cropping or when the face was not pointing forward.

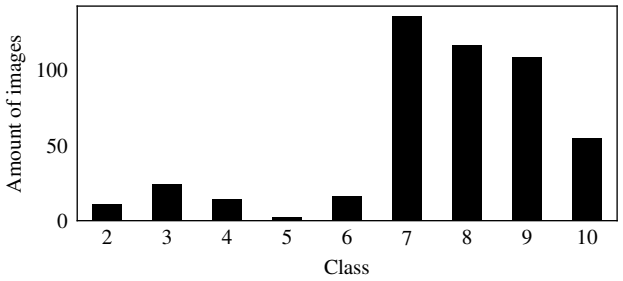


Fig. 5. Amount of images per class that were considered unreadable by Mediapipe

Quantitative evaluation has been carried out with accuracy (Acc), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and also with off-by-one accuracy (OOAcc) as an ordinal counterpart to Acc. To assess any metric advantages gained from data imbalance, we used the macro-averaged M-MSE and macro-averaged M-MAPE [33], [34]. In mathematical notation, let $\|.\|$ be a either MSE or MAPE, then we have

$$M(Y, \hat{Y}) = \frac{1}{n} \sum_{j}^{n} \frac{1}{|\{x \in Y | x = y_j\}|} \|y_j - \hat{y}_j\|, \tag{2}$$

in which Y is the set of all true labels and $\hat{Y}$ represents all predictions values, on the other hand y_j, $\hat{y}_j$ represents all true and all predictions values of a given class j, respectively, lastly, n represents the amount of classes.

Furthermore, we highlight the importance of analysing the confusion matrix when dealing with ordinal classification with imbalanced classes. The confusion matrix reveals the displacement of predictions relative to true values, and for ordinal classification, it is crucial that predictions are near (or on top of) the diagonal. Moreover, each CNN model passed through the Grad-Cam [22] algorithm where visual analysis was performed on each layer activation map with respect to the input.

For model evaluation purposes, the data has been split into: (1) a subset of 366 images, containing 5 individuals of each class, hence, 50 different individuals in total; (2) a subset containing the remaining images (39239), organised as: 15% for testing , 68% for training and 17% for validation, which were divided based on amount of images (and not individuals). The split is stratified to keep the imbalance of classes.

The purpose of the dataset with images of 5 individuals is to evaluate whether the models are overfitting by relying on contextual features rather than generalising skin tone prediction. This overfitting occurs when the models learn to recognise specific individuals based on their clothing and background, or even the overall individual, rather than the skin tone alone.

Additionally, the CNN training schemes employed data augmentation techniques from the Albumentations Python library [35]. These techniques included horizontal flip, noise, blurring, random brightness, and random contrast, and were applied to both the training and validation sets. The training set was normalised using its own average and standard deviation. We also employed learning rate schedulers, specifically, Reduce Learning Rate on Plateau with a patience of 5, and early stopping with a patience of 10. We selected the model based on the lowest validation loss.

IV. Results and Discussion

The results of our experiments are summarised in Table I. We observed that models trained on full-images consistently outperformed those trained on segmented-skin input across all evaluated metrics.

For the custom test set with 5 persons never seen before, the models showed approximately 20% accuracy and 65% off-by-one accuracy. Additionally, the poor performance of all models in the 5 person dataset sample also suggests that the model is overfitting over the imbalance of the dataset, and over the image style[1] present in each dataset as the white labelled people are mostly from LFW dataset and the darker tones are mostly from Casia Face Africa. We also suppose that by having the same individual on the test, validation and training set we may induce the model to perform face recognition rather than skin tone generalisation. However, this result is not conclusive due to the limited number of individuals (50) and images (366) in the dataset, which may not be representative of the full 10-skin tone spectrum. Increasing the number of individuals or selecting a more representative subset may improve model performance on this custom test set.

Furthermore, Table I shows that when dealing with segmented-skin images, the best performing models on the test set showed accuracy ranging from 66% to 74% and off-by-one accuracy from 86% to 92%. Furthermore, for the 5 people custom test set we obtained the same 25% accuracy and 60% off-by-one accuracy, roundly.

Continuing with segmented-skin input, all models showed a increase when using M-MSE and M-MAPE versus MSE and MAPE, which indicates that the model has overfitted around the imbalance of the dataset, however, the increase was minimal, hence, predictions were near the true label. This also explains the high OOAcc and indicates that the use of weighted OCE loss function improved skin tone generalisation. For example, DenseNet121 had 3.02 MSE over 3.67 as M-MSE which shows that the network was focusing on the imbalance of the dataset, the same analysis can be made over all models.

Figure 7 shows the confusion matrix for three classifiers trained on segmented-skin input: (a) DenseNet121, (b) VehicleNet, and (c) Random Forest. It is expected that the use of OCE loss function would bring the predicted labels near the diagonal, this is shown by both CNNs (Figures 7a and 7b) which produced values near the diagonal. However, they also showed a considerable amount of predictions far away from the diagonal. On the other hand, the Random Forest classifier gently reproduced the test set, with a small variance around the diagonal. This explains the higher OOAcc and the larger gap between CNN's and CCV's Mean Squared errors.

Our Grad-Cam analysis showed that CNNs leverage contextual information such as texture and shape from non-skin regions of the image, such as the background or clothing to distinguish the image among classes. This is corroborated by the Grad-Cam visualisations (Figure 6c) applied to a DenseNet121 network fully trained on full-image as input, where blue indicates regions with low neuron activation and red represents high activation. Additionally, the Grad-Cam algorithm output of Figure 6a indicates that the model trained on segmented-skin images correctly interprets the skin pixels as important, and the background as non important. However, the model in some images, may not return a comprehensible Grad-Cam output, as shown in Figure 6b which suggests that the entire image contains important features.

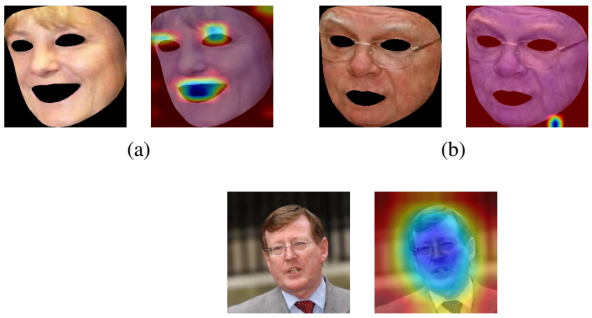
V. Conclusion

This study aimed to tackle the issue of AI Fairness in skin tone related models. To that, we developed a dataset named **SkinTone in The Wild**. We then evaluated it using various classification models. The proposed dataset consists of 39,605 images of 2,183 individuals annotated according to the Monk Skin Tone (MST) scale. The dataset was created using well-known datasets such as CelebA and LFW, which mostly contain lighter skin tones. To ensure a more representative and balanced dataset across all skin tone classes, we included the

[1]Face cropping area, face position, lighting, and blurring.

TABLE I
METRICS RESULTS OF MODELS ON THE TEST SET. THE BEST VALUES ARE IN BOLD. THE TABLE IS DIVIDED BY THE INPUT TYPE.

Models	Metrics / input	ACC	MSE	MAPE	OOAcc	M-MSE	M-MAPE	5-ACC	5-OOAcc
VehicleNet	full-image	0.5610	1.1409	0.1756	0.8613	1.5153	0.2567	0.1946	0.6703
VehicleNet Revisited	full-image	0.6155	0.9811	0.1583	0.8824	1.2893	0.2286	0.1973	0.6297
DenseNet121	full-image	**0.8642**	**0.3312**	**0.0710**	**0.9637**	**0.6706**	**0.1243**	**0.2460**	**0.6946**
Vehicle Net [30]	segmented-skin	0.3792	5.4964	0.3954	0.7196	7.6458	0.7006	0.2276	0.4808
Vehicle Net Revisited	segmented-skin	0.5152	5.2497	0.3247	0.8147	6.8380	0.5066	0.2622	0.5491
DenseNet121	segmented-skin	0.6778	3.0246	0.1907	0.8630	3.6751	0.3159	**0.3114**	0.5437
RandomForest	segmented-skin	**0.7405**	**0.9222**	**0.1094**	**0.9231**	**1.3751**	**0.1918**	0.2541	**0.6667**
KNN	segmented-skin	0.7390	1.2729	0.1480	0.8988	1.6253	0.2196	0.2077	0.6148
MLP	segmented-skin	0.6679	1.3579	0.1450	0.8914	1.7550	0.2217	0.2404	0.6148



(a) (b) (c)

Fig. 6. Grad-Cam analysis retrieved from the 5-person test dataset. (a) Grad-Cam from DenseNet121 aligned with non important pixels. (b) Grad-Cam from DenseNet121 not aligned. (c) Grad-Cam from DenseNet121 trained on full-image input.

Casia Face Africa dataset with individuals of darker skin tones. However, there are still imbalances in the brown spectrum of classes, primarily due to the composition of the parent datasets.

Our experiments demonstrates that models trained on full-image input were prone to overfitting, mainly caused by the fact that the CNN models learned to recognise specific identities and contextual features rather than generalising skin tone classification.

On the other hand, although models using segmented-skin input showed a lower performance, they presented better Grad-Cam visualisations which indicates the generalisation of skin tone classification.

Another key finding was that classic computer vision pipelines with Histograms as input and a Random Forest as classifier showed similar performance to CNNs. Both performed well, yielding around 70% accuracy and 90% off-by-one accuracy. The best classifier was RandomForest, trained on segmented-skin input with histograms of channels RGB, Y (from YCbCr), V (from HSV), and L (from Lab) as descriptors, which reached 74% accuracy and 92% off-by-one accuracy.

Furthermore, one of our experiments used a custom test set containing 5 persons of each skin tone, unknown to the classifiers. For this set, all classifiers performed poorly and somewhat equally, reaching around 20% accuracy and 65% off-by-one accuracy. However, as the CNNs and CCVs on both full-image and segmented-skin datasets achieved similar results, one should consider change the number of persons or the individuals.

Our study emphasises the critical need to address overfitting and dataset biases to achieve fair and accurate AI systems that are related to skin tone. This will require efforts to refine both datasets and model architectures.

As future work, we plan to address the imbalance of skin tone classes within the dataset by adding more images and using data augmentation techniques. We also plan to publish the dataset online. Furthermore we will test additional models aimed at tackling the imbalanced classes and the generalisation of skin tone classification. Focusing on the CCV model, we will test preprocessing techniques to tackle lighting changes.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 (grant 88887.842584/2023-00).

Portions of the research in this paper use the CASIA-Face-Africa and CASIA-FaceV5 collected by the Chinese Academy of Sciences' Institute of Automation (CASIA)

REFERENCES

[1] Jornal OGlobo, "Vídeo com saboneteira levanta debate sobre 'tecnologias racistas'," 2017. [Online]. Available: https://oglobo.globo.com/economia/video-com-saboneteira-levanta-debate-sobre-tecnologias-racistas-21720614

[2] J. Zou and L. Schiebinger, "Ai can be sexist and racist—it's time to make it fair," 2018.

[3] Repórter Unicamp, "Oxímetros podem apresentar menor precisão em pessoas negras," 2021. [Online]. Available: https://www.unicamp.br/unicamp/tv/reporter-unicamp/2021/09/30/oximetros-podem-apresentar-menor-precisao-em-pessoas-negras

[4] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.

[6] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.

[7] T. B. Fitzpatrick, "Soleil et peau," *J. Med. Esthet.*, vol. 2, pp. 33–34, 1975.

[8] E. Monk, "Monk skin tone scale," 2019. [Online]. Available: https://skintone.google

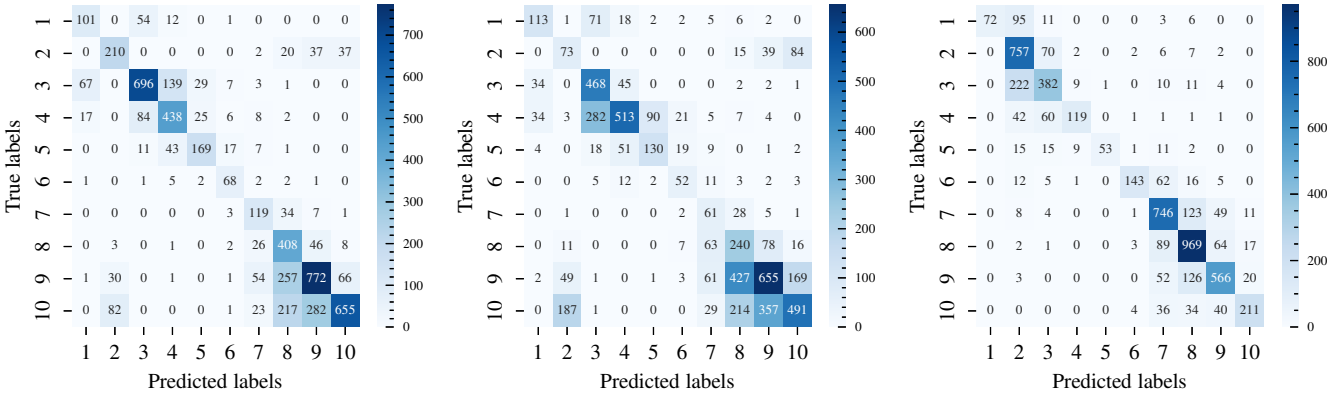


Fig. 7. Confusion matrices of classifiers applied on the test set of segmented-skin images. (a) DenseNet121 classifier fine-tuning. (b) VehicleNet revisited. (c) Random Forest.

[9] C. M. Heldreth, E. P. Monk, A. T. Clark, C. Schumann, X. Eyee, and S. Ricco, "Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence," *ACM J. Responsib. Comput.*, vol. 1, no. 1, mar 2024. [Online]. Available: https://doi.org/10.1145/3632120

[10] D. Borza, A. S. Darabant, and R. Danescu, "Automatic skin tone extraction for visagism applications." in *VISIGRAPP (4: VISAPP)*, 2018, pp. 466–473.

[11] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Estimating skin tone and effects on classification performance in dermatology datasets," *arXiv preprint arXiv:1910.13268*, 2019.

[12] M. Z. Osman, M. A. Maarof, M. F. Rohani, N. N. A. Sjarif, and N. S. A. Zulkifli, "A multi-color based features from facial images for automatic ethnicity identification model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1383–1390, 2020.

[13] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1820–1828.

[14] S. Kye and O. Lee, "Skin color classification of koreans using clustering," *Skin Research and Technology*, vol. 28, no. 6, pp. 796–803, 2022.

[15] I. Boaventura, V. Volpe, I. da Silva, and A. Gonzaga, "Fuzzy classification of human skin color in color images," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 6, 2006, pp. 5071–5075.

[16] R. A. Rejón Piña and C. Ma, "Classification algorithm for skin color (casco): A new tool to measure skin color in social science research," *Social Science Quarterly*, vol. 104, no. 2, pp. 168–179, 2023.

[17] M. Sobhan, D. Leizaola, A. Godavarty, and A. M. Mondal, "Subject skin tone classification with implications in wound imaging using deep learning," in *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2022, pp. 1640–1645.

[18] G. A. Tadesse, C. Cintas, K. R. Varshney, P. Staar, C. Agunwa, S. Speakman, J. Jia, E. E. Bailey, A. Adelekun, J. B. Lipoff *et al.*, "Skin tone analysis for representation in educational materials (star-ed) using machine learning," *NPJ Digital Medicine*, vol. 6, no. 1, p. 151, 2023.

[19] K. Robin, T. Loïc, E. MALHERBE, and M. PERROT, "Beyond color correction: Skin color estimation in the wild through deep learning," *Electronic Imaging*, vol. 32, pp. 1–8, 2020.

[20] H. Choi, K. Choi, and H.-J. Suk, "Performance of the 14 skin-colored patches in accurately estimating human skin color," in *Electronic Imaging, Computational Imaging XV 2017*. Society for Imaging Sciences and Technology, 2017, pp. 62–65.

[21] A. Ward, J. Li, J. Wang, S. Lakshminarasimhan, A. Carrick, B. Campana, J. Hartford, P. K. S, T. Tiyasirichokchai, S. Virmani, R. Wong, Y. Matias, G. S. Corrado, D. R. Webster, D. Siegel, S. Lin, J. Ko, A. Karthikesalingam, C. Semturs, and P. Rao, "Crowdsourcing dermatology images with google search ads: Creating a real-world skin condition dataset," 2024.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[23] L. Spacek, "Collection of facial images: Faces94 and faces95," *Computer Vision Science and Research Projects, University of Essex, United Kingdom*, 1995. [Online]. Available: https://cmp.felk.cvut.cz/~spacelib/faces/faces94.html

[24] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[25] L. L. de Oliveira Junior and C. E. Thomaz, "Captura e alinhamento de imagens: Um banco de faces brasileiro," Department of Electrical Engineering, FEI, São Bernardo do Campo, São Paulo, Brazil, Undergraduate Technical Report, June 2006.

[26] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, pp. 341–345.

[27] Chinese Academy of Sciences, "Casia-facev5," 2009. [Online]. Available: http://biometrics.idealtest.org/

[28] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[29] J. Muhammad, Y. Wang, C. Wang, K. Zhang, and Z. Sun, "Casia-face-africa: A large-scale african face image database," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3634–3646, 2021.

[30] R. F. Rachmadi and I. Purnama, "Vehicle color recognition using convolutional neural network," *arXiv preprint arXiv:1510.07391*, 2015.

[31] G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[32] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[33] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *2009 Ninth international conference on intelligent systems design and applications*. IEEE, 2009, pp. 283–287.

[34] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*. Springer, 2009, pp. 207–210.

[35] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.