

Fuzzy Accuracy Compensates for Label Subjectivity in Classification of Skin Tone Using Wearable Photoplethysmography Signals

Padmini Krishnadas
Data Science and AI Department
National Physical Laboratory
Teddington, UK
Padmini.Krishnadas@npl.co.uk

Urs Hackstein
Faculty of Life Science Engineering
Mittelhessen University of Applied Sciences
Giessen, Germany
Urs.Hackstein@lse.thm.de

Alen Bošnjaković
Laboratory for Mass and Related Quantities
Institute of Metrology of Bosnia and Herzegovina
Sarajevo, Bosnia and Herzegovina
Alen.Bosnjakovic@met.gov.ba

Philip J. Aston
Data Science and AI Department
National Physical Laboratory
Teddington, UK
Philip.Aston@npl.co.uk

Abstract—We consider the problem of classification of skin tone using photoplethysmography (PPG) signals with labels of the ordinal six-class Fitzpatrick skin tones. A typical accuracy for this task is a poor 40-55 %. However, the labels are subjectively determined by comparing the skin with a colour chart, and hence contain widespread small-scale inaccuracies. By working with a “fuzzy accuracy”, which deems a prediction of skin tone class to be correct if its difference from the labelled class is not greater than one, much higher accuracy is obtained which provides more convincing evidence that skin tone can be accurately predicted from PPG signals. Three machine learning approaches were used, namely deep learning or tree-based approaches on raw PPG signals, deep learning on image representations of the signals generated by the Symmetric Projection Attractor Reconstruction (SPAR) method, and machine learning on features extracted from the signals. The first method also employed a fuzzy version of the cross entropy loss function, which gave the best results. Tree-based models on raw signals give accuracies up to 55 % and higher fuzzy accuracies up to 96 %, while deep learning models on the SPAR images obtained lower results of 44 % accuracy and 85 % fuzzy accuracy. The machine learning on PPG features gave similar results to the SPAR method with accuracy of 42 % and fuzzy accuracy of 87 %. We have shown that classification of skin tone using PPG signals is possible with high fuzzy accuracy which implies that our modelling approach enables accurate prediction of skin tone class within at most one class of the observer’s choice of class, from which we conclude that PPG signals are affected by skin tone in a discernible way.

Index Terms—fuzzy accuracy, fuzzy cross entropy, Fitzpatrick skin tone, photoplethysmography signals

The project (22HLT01 QUMPHY) has received funding from the European Partnership on Metrology, co-financed from the European Union’s Horizon Europe Research and Innovation Programme and by the Participating States. Funding for NPL was provided by Innovate UK under the Horizon Europe Guarantee Extension, grant number 10084125.

I. INTRODUCTION

Machine learning classification models are trained on data consisting of features and class labels, and the training process involves learning the mapping from features to class labels. This process assumes that the class labels are accurate and correct. However, in many medical disease diagnosis problems, the labels indicate the presence or absence of particular diseases and are determined by clinicians based on data from diagnostic tests (such as the electrocardiogram (ECG) for cardiovascular diseases) but clinicians do not always agree on the diagnosis as subjective judgement is required. Thus, there is the potential for a non-negligible proportion of records to be misclassified.

The problem we consider has much greater potential for label errors and consists of classification of skin tone from photoplethysmography (PPG) signals. PPG signals are generated by both pulse oximeters, which are widely used in a clinical environment, and by consumer wearable devices such as smart watches and rings. Both devices offer non-invasive continuous health monitoring, often deployed among vulnerable populations in intensive care or with pre-existing medical conditions, highlighting the need for accurate measurements. However, it was found during the Covid-19 pandemic that the blood oxygenation readings from pulse oximeters were overestimated for patients with darker skin tones which has the potential for incorrect decisions being made regarding treatment in these cases. An independent report *Equity in Medical Devices: Independent Review* [1] was commissioned in the UK after the pandemic which found “extensive evidence of poorer performance of pulse oximeters for patients with darker skin tones”. PPG signals are generated by shining light onto the skin and measuring the light that is reflected back

and so it is not surprising that the reflected light, and hence the PPG signal, is different for light and dark skin tones. We consider the classification of skin tone from PPG signals to investigate this dependence further using machine learning (ML). ML models have been shown to differentiate ethnicity from medical data that does not contain this information directly, making them well suited to the skin tone classification problem [2]–[4]. If skin tone cannot be accurately classified using PPG signals by machine learning, it suggests that the signals do not depend on skin tone. Conversely, if machine learning does give an accurate classification of skin tone using PPG signals, then the signals are clearly dependent on skin tone in some way.

Skin tone is commonly classified according to the six-class Fitzpatrick skin tone scale [5] where the lightest skin tone is class I and the darkest is class VI. The Fitzpatrick scale was originally developed to classify people with white skin in order to select the correct UV dose for treating skin disease, not for characterising skin tone [5]. However, its use as a scale for skin tone is now “nearly ubiquitous” [6] and there are a few PPG datasets that record the Fitzpatrick skin tone for subjects, along with other metadata, including the AuroraBP dataset [7] that we use. Another complication with this scale is that the classes, and their boundaries, are subjective and lack a formal, measurable definition. Moreover, for accurate classification, the skin tone should be measured using a spectrophotometer (or similar device) at the site of data collection (e.g. the wrist for a smart watch). However, this is not usual practice. Instead, skin tone is commonly assessed subjectively by comparing the skin with a colour chart showing the typical tone in each class. Using this comparison with a colour chart, there can often be uncertainty as to whether the particular skin under consideration should be in class n or $n + 1$ for some n . Indeed, in an assessment of manual grading, there were many one class differences between the classes designated by the graders for a particular facial image, and differences of two or more classes also occurred [8]. Such subjective assessment can result in a large number of misclassifications which correspond to inaccurate labels for our problem of skin tone classification.

A problem such as this with potentially widespread, small-scale inaccuracies in the labels can result in poor absolute accuracy when using a machine learning model trained on such data. As the labels could be described as “fuzzy” or uncertain, and are also ordinal, we define a “fuzzy accuracy”, whereby a class prediction of n is assumed to be correct if the assigned label is within one class of the prediction, i.e. $n - 1$, n or $n + 1$. This fuzzy accuracy is clearly greater than or equal to the accuracy by definition, so of interest will be the magnitude of the increase in the fuzzy accuracy.

In related work, the influence of skin tone on machine learning-based predictions has recently been investigated in [9], where classification of high blood pressure from PPG signals was examined. This study showed that when a model was trained exclusively on records from individuals with Fitzpatrick skin tone class I (the lightest skin tone) and subsequently evaluated on data from the other skin tone classes,

classification accuracy generally declined markedly for those classes in comparison with models which had been trained on data which included all skin tone classes. Despite this, many publicly available PPG datasets either do not record skin tone information or are heavily skewed towards the lighter skin tones, with relatively few samples from darker skin tones. This imbalance is present, for example, in the Aurora-BP dataset used in this work [7]. As a result, models trained primarily on data from lighter-skinned subjects may not generalise well to individuals with darker skin tones, raising concerns about potential performance disparities.

The objective of this study is therefore to investigate whether skin tone itself can be predicted from PPG signals using machine learning and evaluated with our concept of fuzzy accuracy, framed as a six-class classification problem based on the Fitzpatrick skin tone scale in order to provide evidence of a machine learning model’s sensitivity to skin tone in PPG signals. Machine learning methods are well suited to identifying complex relationships between signals and target labels. Consequently, strong predictive performance on this task would suggest that PPG signals exhibit systematic differences across skin tone categories in which case PPG datasets used for other machine learning tasks should include a uniform distribution of skin tones to ensure that the results are not biased. Conversely, poor classification performance would indicate that machine learning models are unable to reliably distinguish between signals associated with different skin tones, implying that skin tone may have a limited effect on the PPG signals themselves in which case skin tone would not have to be considered when constructing datasets for other machine learning tasks involving PPG signals.

The code for this work is available at https://gitlab.com/qumphy/fuzzy_accuracy_ppg_skin_tone.

A. Data

For this study, we used the AuroraBP dataset [7] which includes PPG waveforms (sampling frequency 500 Hz) collected using a wrist-worn optical device and Fitzpatrick skin tone labels for 823 participants (51.6 % female). Two distinct protocols were used, namely auscultatory and oscillometric, and we combined both of these into a single dataset giving a combined total of 19,017 records. The signals were of varying length but were generally around 15 s long.

The signals were of varying quality and a measure of “optical quality” is included in the dataset. This measure takes into account artifacts, signal to noise ratio and consistency between pulses [7]. Records with optical quality ≥ 0.65 were used in [7] so we used the same threshold. There are a total of 14,667 records satisfying this condition of which only 35 records are skin tone class VI.

The distribution of records with optical quality ≥ 0.65 by skin tone class is shown in Fig. 1, which shows that the dataset is highly imbalanced with the majority of subjects having lighter skin tones (92.8 % in classes I, II and III). Also, there are no female subjects with skin tone class VI.

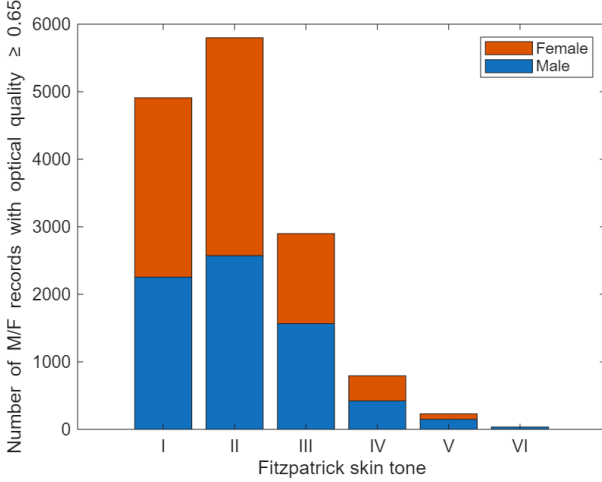


Fig. 1: THE NUMBER OF RECORDS IN EACH SKIN TONE CLASS FOR WHICH OPTICAL QUALITY ≥ 0.65 .

II. METHODS

Three different machine learning approaches will be used for this classification problem in order to consider different input representations, namely deep learning using one-dimensional convolutional neural networks (CNN) or tree-based machine learning models taking the raw PPG signals as input, transformation of the PPG signals into Symmetric Projection Attractor Reconstruction (SPAR) attractor images [10] which are used as input to a two-dimensional CNN, and machine learning using features extracted from the PPG signals. For all three methods, the normal accuracy will be reported together with our fuzzy accuracy, which we now define.

A. Fuzzy Accuracy for Ordinal Fitzpatrick Labels

Self-reported Fitzpatrick skin tone is known to exhibit subjectivity and inter-annotator variability, with frequent near-adjacent label discrepancies [11], [12]. To compensate for this uncertainty when evaluating model performance, we evaluate our models using *fuzzy accuracy*, a tolerance-aware metric tailored to ordinal classes.

Let $y_i \in \{0, \dots, 5\}$ denote the labelled class (I–VI mapped to 0–5) and $\hat{y}_i$ the predicted class for sample i . We define a tolerance radius $\tau \in \mathbb{N}$ and a weight vector $\mathbf{w} = (w_0, w_1, \dots, w_\tau)$ with $1 = w_0 \geq w_1 \geq \dots \geq w_\tau \geq 0$. The per-sample fuzzy match is

$$m_i(\hat{y}_i, y_i; \tau, \mathbf{w}) = \begin{cases} w_d, & d = |\hat{y}_i - y_i| \leq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

The *fuzzy accuracy* is the mean fuzzy match over N samples:

$$\text{Acc}_{\text{fuzzy}}(\tau, \mathbf{w}) = \frac{1}{N} \sum_{i=1}^N m_i(\hat{y}_i, y_i; \tau, \mathbf{w}).$$

In this work, we report the special case $\tau = 1$ with $(w_0, w_1) = (1, 1)$, referred to as *tolerant accuracy* (± 1), which treats

adjacent-class predictions as correct. We also provide *strict accuracy* ($\tau = 0, w_0 = 1$) for reference.

In addition to considering fuzzy accuracy, we also included a *fuzzy cross entropy* (neighbour label smoothing) computed against soft targets that allocate a small mass α to the immediate neighbours (± 1) of y_i , with edge classes having a single neighbour. For logits $\mathbf{z}_i \in \mathbb{R}^6$ and soft target $\mathbf{s}_i \in [0, 1]^6$,

$$\mathcal{L}_{\text{fuzzyCE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=0}^5 s_{i,c} \log(\text{softmax}(\mathbf{z}_i)_c),$$

where $s_{i,y_i} = 1 - \alpha_{\text{left}} - \alpha_{\text{right}}$, $s_{i,y_i-1} = \alpha_{\text{left}}$ (if $y_i > 0$), and $s_{i,y_i+1} = \alpha_{\text{right}}$ (if $y_i < 5$). We employ $\mathcal{L}_{\text{fuzzyCE}}$ as a reported metric while optimising standard cross entropy for training. This design keeps training comparable across baselines yet quantifies robustness to plausible adjacent-class ambiguity. In addition to this, we report the Earth Mover’s Distance (EMD) loss, which uses predicted probabilities of all classes and penalises missed predictions in accordance to a ground reference matrix quantifying class similarity [13], [14], making EMD well aligned with the underlying structure of skin tone gradations in comparison to a standard cross entropy loss. This formulation is consistent with fuzzy evaluation ideas in the soft classification literature [15]–[17] and recent work on ordinal-aware soft targets and calibration [18], [19].

B. Deep/Machine Learning Using Raw Signals

For the first approach, using the raw PPG signals as input to a deep learning model (1d CNN), we first truncated all signals to the first 5,877 samples (11.75 s) to achieve uniformity as signal lengths vary, and the minimum length is 5,877 samples.

Next, the signals were processed in two ways:

- i) Deep learning models were trained using the raw signals.
- ii) The signals were filtered using a fourth order Chebyshev II bandpass filter (0.5–15 Hz cutoff, 20 dB stopband attenuation) [20], and then the same models were trained.

The following ML models were evaluated to compare performance:

- Random forest [21]
- Gradient boosted decision trees [22]
- LeNet-1D [23]
- ResNet-1D50 [23]
- ResNet-1D101 [23]
- VGG16 [23]

A stratified 5-fold cross-validation by subject was employed to ensure that no signals from a subject were in both the training and test datasets. For Gradient boosted decision trees, multiple loss functions were tested, including the multinomial log-likelihood, a cross entropy loss function and two self-defined variants of multinomial log-likelihood and cross entropy respectively reflecting fuzzy accuracy. This was done in a similar way to the explanation given in Section II-A using label smoothing and an indicator matrix which treats adjacent-class predictions as correct.

This is illustrated by the following definition of fuzzy cross entropy loss:

```

1 def fuzzy_loss(
2     labels: nptyping.NDArray[np.int32],
3     predictions: nptyping.NDArray[np.float32],
4     weights: nptyping.NDArray[np.float32],
5 ) -> np.float32:
6     dimension = np.max(labels)
7     sum_exp_pred = np.sum(np.exp(predictions),
8                             axis=1)
9
10    # Indicator matrix for Fuzzy accuracy
11    indicator_matrix = np.zeros((labels.size,
12                                dimension))
13    for i, label in enumerate(labels-1):
14        if label > 0:
15            indicator_matrix[i, label - 1] =
16                0.5 # x-1
17            indicator_matrix[i, label] = 1 # x
18            if label < dimension - 1:
19                indicator_matrix[i, label + 1] =
20                    0.5 # x+1
21
22    label_exp_pred = np.exp(np.sum(predictions
23                                    * indicator_matrix, axis=1))
24    return -np.sum(np.log(label_exp_pred /
25                            sum_exp_pred)) / len(labels)

```

For Random forest and Gradient boosted decision trees, both standard accuracy and fuzzy accuracy are also computed for subsets restricted to male or female subjects, although the choice of classifier was motivated by their performance on the whole dataset.

C. Deep Learning with SPAR Attractor Images

1) *SPAR Attractor Image Generation*: The Symmetric Projection Attractor Reconstruction (SPAR) method converts an approximately periodic signal into a compact image which encapsulates the morphology and variability of the signal [10] (see Fig. 2). The SPAR attractor images can be used as input to a two dimensional CNN for classification of the Fitzpatrick skin tone.

For this approach, we used only the first 15 s of longer signals and all of the signals for shorter signals. To generate SPAR images, we first filtered the signals using a fourth order high pass Chebyshev II filter using a cut-off frequency of 0.5 Hz [20]. This removes any baseline wander from the signals which can affect the SPAR attractors. The size of the attractor is determined by the amplitude of the signals [10]. Thus, we scaled the signals in the vertical direction by the amplitude so that all signals had amplitude 1 resulting in consistently sized attractors. We then generated SPAR attractor images for each signal using a 64×64 grid and three delay coordinates, resulting in an approximate threefold rotational symmetry in the attractors.

2) *Deep Learning*: For SPAR attractor images, skin tone classification is formulated as an image classification task. We compared the performance of ResNet50, ResNet152 and VGG-16 models for classifying Fitzpatrick skin tones. We used ImageNet-pretrained ResNet-50 ResNet152 and VGG-16 backbones. ResNet models employed bottleneck residuals with final feature dimensions of 2048, while VGG-16 uses

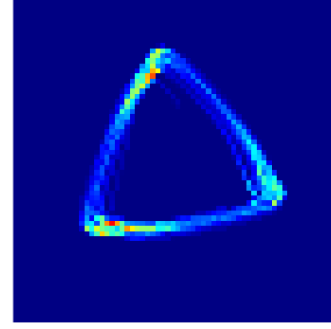
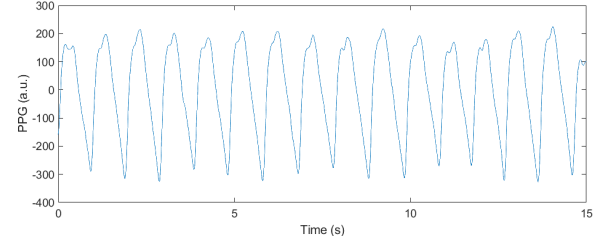


Fig. 2: A PPG SIGNAL (TOP) AND THE CORRESPONDING SPAR ATTRACTOR (BOTTOM).

sequential 3x3 convolutions followed by a 4096-unit classifier. For all architectures, we replaced the final classification layer with a task specific output head, mapping to the six skin tone classes and applied differential learning rates to the backbone and classifier, and an Adam optimiser. Prior to training, data augmentation techniques were applied to SPAR images to increase dataset diversity, improve generalisation, and mitigate overfitting. Augmentations also enhance model robustness by effectively expanding the training set [24]–[27]. Augmentations considered are: random horizontal and vertical flips, rotations and resized crop. To address uncertainty in Fitzpatrick skin tone labels, we adopt fuzzy accuracy for multi-class evaluation, implemented with the Earth Mover’s Distance (EMD) loss function.

D. Machine Learning Using PPG Features

We used the PulseAnalyse code [28] to extract PPG features from the raw signals. Further details about these PPG features are provided in [29]. This function is designed to analyse arterial pulse waves, including blood pressure and PPG signals.

We applied a fourth-order Chebyshev Type II band-pass filter (0.5–15 Hz) to the raw signals to remove low-frequency drift and high-frequency noise before passing them to the PulseAnalyse function. We extracted 66 features, including 39 fiducial point-based features, 15 amplitude-based features, and 12 cardiovascular indices, and organised them into a feature table with dimension $14,667 \times 66$. Raw signals with optical quality ≥ 0.65 were used, and all missing values were replaced by zeros. Furthermore, we were not able to process all the raw signals using the PulseAnalyse function, with approximately 0.3 % (41 signals) not being processed, and additional effort will be invested to understand this issue.

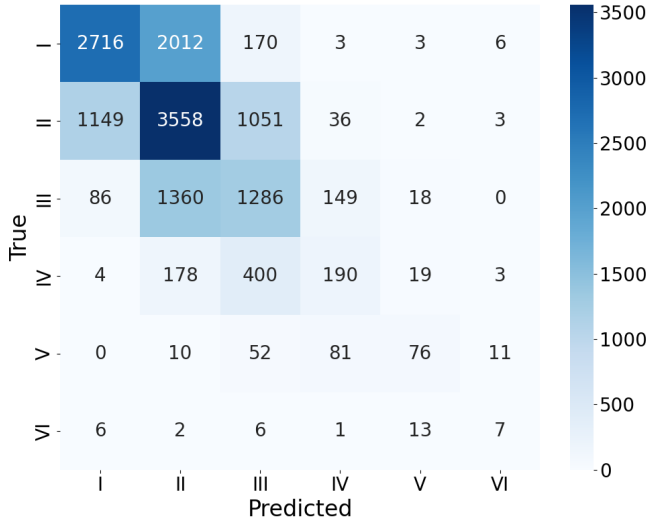


Fig. 3: CONFUSION MATRIX OVER 5-FOLD CROSS VALIDATION FOR THE GRADIENT BOOSTED DECISION TREE MODEL WITH FUZZY CROSS ENTROPY LOSS FUNCTION TRAINED ON RAW SIGNALS.

We applied three supervised machine learning methods to these features, namely neural network, classification trees, and support vector machine. The neural network involved two fully connected layers, the first having 20 nodes, followed by ReLU activations.

III. RESULTS

A. Deep/Machine Learning Using Raw Signals

The values obtained for the accuracy and fuzzy accuracy for the different classifiers and loss functions when using a tree-based model on raw signals are shown in Table I. Similarly, Table II shows analogous results for the filtered signals. The other (CNN-based) classifiers (LeNet-1D, ResNet-1D50, ResNet-1D101 and VGG-16) showed significantly lower results for both accuracy and fuzzy accuracy. The evaluation of the tree-based models (i.e. Random forest and Gradient boosted decision trees) on female or male only subsets delivered similar values for accuracy and fuzzy accuracy with a difference that was always less than 0.04, which indicates that there is no gender bias in these results. We note that the fuzzy cross entropy loss function gave the best fuzzy accuracy for the raw signals, whereas the Random forest approach gave the best fuzzy accuracy for the filtered signals. However, the best accuracy was obtained using Gradient boosted decision trees with the multinomial log likelihood loss function for the raw signals or the Random forest approach for the filtered signals. An exemplary confusion matrix combined over all folds for the case of Gradient boosted decision trees with fuzzy cross entropy loss function with 5-fold cross validation by subject is shown in Figure 3. (Note that the numbers indicate classification by record not by subject.)

B. Deep Learning using SPAR images

We assessed the ResNet50, ResNet152 and VGG-16 models using fuzzy accuracies, implemented with 5-fold cross-

TABLE I: ACCURACY AND FUZZY ACCURACY VALUES FOR DIFFERENT CLASSIFIERS AND LOSS FUNCTIONS APPLIED TO THE RAW SIGNALS. THE BEST RESULTS ARE SHOWN IN BOLD.

GBT = GRADIENT BOOSTED DECISION TREE, RF = RANDOM FOREST CLASSIFIER

Classifier	Loss Function	Accuracy	Fuzzy Accuracy
GBT	Cross entropy	0.5502	0.9561
GBT	Fuzzy cross entropy	0.5336	0.9594
GBT	Multinomial log likelihood	0.5503	0.9561
GBT	Fuzzy multinomial log likelihood	0.4585	0.9339
RF	Standard	0.5332	0.9531

TABLE II: ACCURACY AND FUZZY ACCURACY VALUES FOR DIFFERENT CLASSIFIERS AND LOSS FUNCTIONS APPLIED TO THE FILTERED SIGNALS. THE BEST RESULTS ARE SHOWN IN BOLD.

GBT = GRADIENT BOOSTED DECISION TREE, RF = RANDOM FOREST CLASSIFIER

Classifier	Loss Function	Accuracy	Fuzzy Accuracy
GBT	Cross entropy	0.5427	0.9563
GBT	Fuzzy cross entropy	0.5369	0.9554
GBT	Multinomial log likelihood	0.5427	0.9563
GBT	Fuzzy multinomial log likelihood	0.4440	0.9264
RF	Standard	0.5431	0.9564

TABLE III: OVERALL PERFORMANCE METRICS FROM RESNET50, RESNET152 AND VGG-16 FOR MULTICLASS SKIN TONE CLASSIFICATION USING CROSS ENTROPY LOSS AND FUZZY ACCURACY. THE BEST ACCURACY AND FUZZY ACCURACY ARE SHOWN IN BOLD.

Classifier	Loss function	Loss Score	Accuracy	Fuzzy Accuracy
ResNet50	Cross entropy	1.2138	0.4567	0.8435
ResNet152	Cross entropy	1.2071	0.4439	0.8450
VGG-16	Cross entropy	1.2111	0.4603	0.8459
ResNet50	Earth Mover Distance	0.0761	0.4673	0.8498
ResNet152	Earth Mover Distance	0.0783	0.4339	0.8504
VGG-16	Earth Mover Distance	0.076	0.4621	0.8588

TABLE IV: ACCURACY AND FUZZY ACCURACY VALUES FOR DIFFERENT CLASSIFIERS AND LOSS FUNCTIONS APPLIED TO THE PPG FEATURES. THE BEST RESULTS ARE SHOWN IN BOLD.

Classifier	Loss Function	Accuracy	Fuzzy Accuracy
Trees	Classification error	0.3473	0.8100
Neural Network	Cross entropy	0.4219	0.8705
Support Vector Machine	Classification error	0.4024	0.8638

validation. All models reported similar classification accuracies, with VGG-16 giving a marginally higher fuzzy accuracy of 0.8588. Table III lists the performance of all models when classifying the Fitzpatrick skin tones. Figure 4 shows the confusion matrix across all 5 folds when using the VGG-16 model. This illustrates the model's capability to discriminate between classes I, II and even class III to some extent, with many misclassifications. However, the model was unable to identify records for classes IV, V and VI with any accuracy. This is likely due to a combination of imbalance in the dataset causing the model to favour learning light skin tones over dark as well as very small sample sizes effectively rendering classes V and VI as statistical outliers.

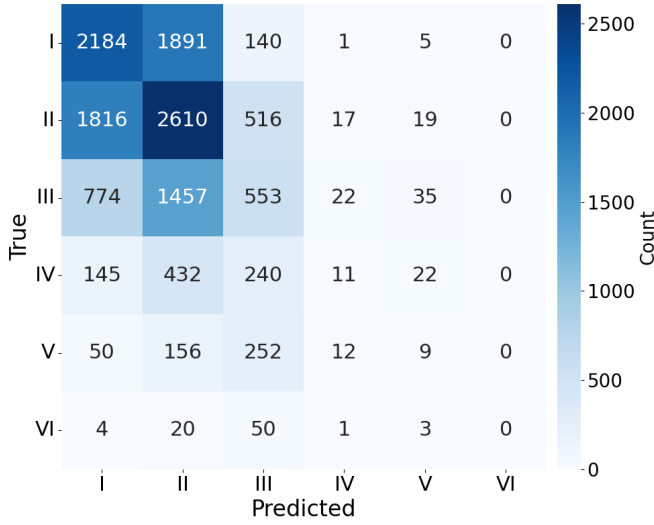


Fig. 4: CONFUSION MATRIX OVER 5-FOLD CROSS-VALIDATION USING VGG16 WITH EMD LOSS TRAINED ON SPAR IMAGES.

C. Machine Learning Using PPG Features

Neural Networks and SVM methods achieved almost the same accuracy and the same fuzzy accuracy on PPG features (see Table IV) while Trees reported lowest performance metrics. The highest fuzzy accuracy was obtained using the neural network method. The results are presented using a confusion matrix that combines all folds for the neural network method with five-fold cross-validation performed by subject, as shown in Figure 5. The numbers represent classification by record rather than by subject.

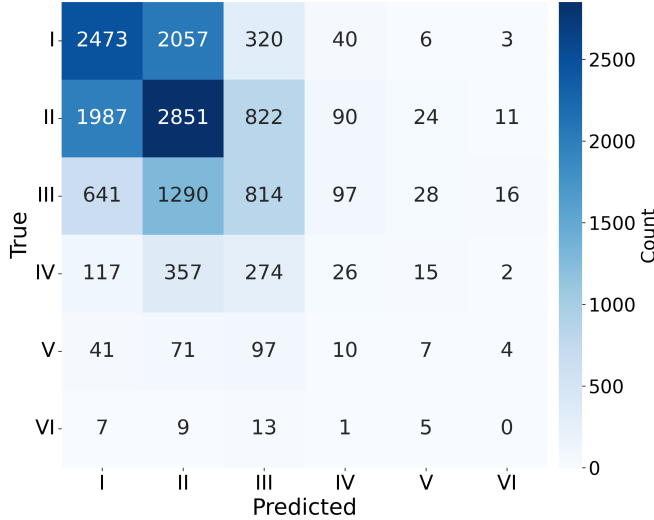


Fig. 5: CONFUSION MATRIX OVER 5-FOLD CROSS-VALIDATION FOR THE NEURAL NETWORK MODEL USING PPG FEATURES.

CONCLUSIONS

We have shown that in order to deal with widespread small-scale inaccuracies in the skin tone labels for our classification

problem, the introduction of a fuzzy accuracy provides a metric that scores much higher than the normal accuracy, thus showing that machine learning prediction of skin tone class can be performed very accurately within one class of the subjective class label. Moreover, the use of a fuzzy cross entropy also resulted in a (marginally!) higher fuzzy accuracy.

The best results for this classification problem when considering fuzzy accuracy were obtained by tree-based models using raw signals as input. In particular, the best fuzzy accuracy was obtained using a Gradient boosted decision tree model with a fuzzy cross entropy loss when working with the raw signals, or a Random forest model when working with filtered raw signals. There was very little difference between the results using raw and filtered signals and predictions of the under represented darker skin tones was reasonable. The SPAR and PPG features methods gave fuzzy accuracies approximately 10 % lower, and with poor prediction of the darker skin tones, with the dominant lighter skin tone classes being favoured. However, we note that the SPAR method also normalised the amplitudes of the signals in order to give comparably sized attractors and so any variation in amplitude with skin tone is lost in this case. Since the PPG signals are generated by light passing through the skin, it is very likely that the amplitude of the signals will be skin tone dependent.

The classification results that we have presented support our concept of fuzzy accuracy. The values for the accuracy are significantly higher than the uniform value of 16.7% which would arise from prediction of the classes if there was no dependence of the PPG signals on skin tone, but they are not high enough to give accurate prediction of the labelled Fitzpatrick skin tone of the subject. However, the fuzzy accuracies are reaching up to almost 96 %, strongly supporting the hypothesis that skin tone classes can be determined from PPG signals up to a deviation of ± 1 . This also suggests that if the skin tone was measured more precisely, resulting in more accurate labels, then the prediction accuracy would be much better also.

The Aurora BP dataset that we used for this study is highly imbalanced with regard to skin tone, with very few records with the darker skin tones. Future work could include repeating this study with a more balanced dataset.

A consequence of this work is that skin tone should be taken into consideration when constructing PPG datasets for machine learning, and in the calibration and use of wearable devices and pulse oximeters. Failure to do so could result in biased results and, potentially, consequent harm to the wearer.

REFERENCES

- [1] "Equity in Medical Devices: Independent Review," <https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report>, 2024, accessed: 22/07/2025.
- [2] P. A. Noseworthy, Z. I. Attia, L. C. Brewer, S. N. Hayes, X. Yao, S. Kapa, P. A. Friedman, and F. Lopez-Jimenez, "Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis," *Circulation: Arrhythmia and Electrophysiology*, vol. 13, no. 3, p. e007988, 2020.

- [3] J.-S. Kim, X. Gao, and A. Rzhetsky, "Riddle: Race and ethnicity imputation from disease history with deep learning," *PLoS computational biology*, vol. 14, no. 4, p. e1006106, 2018.
- [4] J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang *et al.*, "Ai recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health*, vol. 4, no. 6, pp. e406–e414, 2022.
- [5] T. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.*, vol. 124, pp. 869–871, 1988.
- [6] E. Monk, "The Monk skin tone scale (MST)," *SocArXiv*, vol. <https://doi.org/10.31235/osf.io/pdf4c>, 2023.
- [7] B. Mieloszyk, H. Twede, J. Lester, J. Wander, S. Basu *et al.*, "A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting," *IEEE J. Biomed. Health Inform.*, vol. 26, pp. 2864–2875, 2022.
- [8] K. Krishnapriya, G. Pangelinan, M. King, and K. Bowyer, "Analysis of manual and automated skin tone assignments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2022, pp. 429–438.
- [9] P. Aston, "Does skin tone affect machine learning classification accuracy applied to photoplethysmography signals?" *Comp. Cardiol.*, vol. 51, p. 038, 2024.
- [10] P. Aston, M. Christie, Y. Huang, and M. Nandi, "Beyond HRV: Attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction," *Phys. Meas.*, vol. 39, p. 024001, 2018.
- [11] S. Tian, "Shades of skin: Limitations of the Fitzpatrick scale with CIELAB," 2024.
- [12] V. R. Weir, Y. Li, M. C. Gillis, N. R. Kurtansky, T. Salvador *et al.*, "Evaluating skin tone scales for dermatologic dataset labeling: A prospective-comparative study," *npj Dig. Med.*, vol. 8, p. 787, 2025.
- [13] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.
- [14] Z. Zhang, Y. Zhang, X. Zhao, and Y. Gao, "Emd metric learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [15] D. Gómez and J. Montero, "Determining the accuracy in supervised fuzzy classification problems," in *Computational Intelligence in Decision and Control*. World Scientific, 2008, pp. 411–416.
- [16] E. Binaghi, P. A. Brivio, P. Ghezzi, and A. Rampini, "A fuzzy set-based accuracy assessment of soft classification," *Patt. Rec. Letts*, vol. 20, no. 9, pp. 935–948, 1999.
- [17] J. L. Silván-Cárdenas and L. Wang, "Sub-pixel confusion-uncertainty matrix for assessing soft classifications," *Rem. Sens. Envir.*, vol. 112, no. 3, pp. 1081–1095, 2008.
- [18] D. Kim, H. Chung, and I. Jang, "Calibration of ordinal regression networks," *arXiv preprint arXiv:2410.15658*, 2024.
- [19] C. Kang, X. Yao, and D. Novak, "Fuzzy windows with Gaussian processed labels for ordinal image scoring tasks," *Appl. Sci.*, vol. 13, no. 6, p. 4019, 2023.
- [20] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Sci. Data*, vol. 5, p. 180076, 2018.
- [21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [22] M. Guillaume-Bert, S. Bruch, R. Stotz, and J. Pfeifer, "Yggdrasil decision forests: A fast and extensible decision forests library," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 2023, pp. 4068–4077.
- [23] M. Moulaeifard, L. Coquelin, M. Rinkevicius, A. Sološenko, O. Pfeiffer, C. Bench, H. Hegemann, S. Vardanega, M. Nandi, J. Alastruey, C. Heiss, V. Marozas, A. Thompson, P. Aston, P. Charlton, and N. Strodthoff, "Machine-learning for photoplethysmography analysis: Benchmarking feature, image, and signal-based approaches," *submitted*, 2026.
- [24] P. Krishnadas, K. Chadaga, N. Sampathila, S. Rao, K. Swathi, and S. Prabh, "Classification of malaria using object detection models," *Informatics*, vol. 9, no. 4, p. 76, 2022.
- [25] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 2018, pp. 117–122.
- [26] M. Elgendi, M. Nasir, Q. Tang, D. Smith, J. Grenier *et al.*, "The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective," *Front. Med.*, vol. 8, p. 629134, 2021.
- [27] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey," 2023. [Online]. Available: <https://arxiv.org/abs/2204.08610>
- [28] P. H. Charlton, "Pulse-Analyse," <https://github.com/peterhcharlton/pulse-analyse>, 2023, gitHub repository, accessed Jan. 20, 2026.
- [29] P. H. Charlton, P. Celka, B. Farukh, P. Chowienczyk, and J. Alastruey, "Assessing mental stress from the photoplethysmogram: A numerical study," *Phys. Meas.*, vol. 39, no. 5, p. 054001, 2018.