Does Skin Tone Affect Machine Learning Classification Accuracy Applied to Photoplethysmography Signals?

Philip Aston[1,2]

[1] National Physical Laboratory, Teddington, UK
[2] University of Surrey, Guildford, UK

Abstract

A recent UK report highlighted the fact that measurements of blood oxygenation using pulse oximeters may be inaccurate for patients with darker skin tones. The photoplethysmography signals collected by pulse oximeters and a multitude of wearable devices are often combined with machine learning to derive physiological parameters that may be useful for health monitoring. So it is natural to question whether machine learning accuracy when using photoplethysmography signals as input can also lead to inaccurate results for darker skin tones, particularly when a model may be trained on data from subjects with predominantly lighter skin tones.

To test this, we chose a binary classification problem of classifying systolic blood pressure as high (≥ 140 mm Hg) or not high (< 140 mm Hg). We used the Aurora BP dataset which includes photoplethysmography signals, blood pressure readings and Fitzpatrick skin tone classification. We trained a model using only skin tone class 1 data (lightest skin tone) and then tested all the other skin tone data using this model. We found that the accuracy dropped for all the other skin tone classes. We then trained a model using data from all skin tones and the accuracy for the darker skin tones improved except class 5.

1. Introduction

During the Covid-19 pandemic, it became clear that the blood oxygenation reading from a pulse oximeter may be inaccurate for patients with darker skin tones as the true oxygen levels are overestimated, which can result in suboptimal treatment for some ethnic groups. In the UK, an independent report *Equity in Medical Devices: Independent Review* [1] was commissioned which found "extensive evidence of poorer performance of pulse oximeters for patients with darker skin tones". Due to the optical nature of the photoplethysmography (PPG) signals collected by a pulse oximeter, this is not altogether surprising.

The PPG signals generated by pulse oximeters and a multitude of wearable devices are of much current interest as their collection is non-invasive and so can be used for continuous health monitoring. However, the large quantity of data generated could not be manually reviewed by a clinician and so machine learning algorithms are a popular way of extracting useful information from the signals automatically. Applications include monitoring of blood pressure and sleep as well as detection of health conditions such as atrial fibrillation and coronary heart disease [2].

The question then naturally arises as to whether skin tone affects machine learning classification accuracy applied to PPG signals. This bias is all the more likely as many PPG datasets are dominated by signals collected from people with lighter skin tones and so often the training data may have only a small minority of signals from people with darker skin tones. This problem is compounded by the fact that skin tone is often not recorded in available datasets, so the distribution of skin tones in the data is not known. Even when the skin tone is recorded, it is generally categorised by the six-point Fitzpatrick skin tone scale [3] or the more recently developed ten-point Monk skin tone scale [4]. What is really required to fully assess this problem is a measurement of skin tone at the site of the PPG data collection (e.g. the wrist for a smartwatch) using a spectrophotometer or similar device. However, such data is not currently available.

We assess the affect of skin tone on machine learning classification accuracy by training a model on signals from subjects with Fitzpatrick skin tone class 1 (the lightest skin tone) and then testing the model on the signals from subjects with all the other skin tone classes in order to see if classification accuracy is poorer for the darker skin tones. These results are compared with results obtained using signals associated with all skin tones in the training data which it is anticipated may give better performance across the skin tone range.

2. The problem

The problem we have chosen to consider is classification of high blood pressure (hypertension) from PPG sig-

Table 1. Data for the different Fitzpatrick skin tone classes in the Aurora BP dataset [7].

Skin tone class	No. of subjects	No. of records available	No. of records used
1	311	5289	4989
2	294	7072	6064
3	138	3972	3318
4	44	1212	976
5	28	679	458
6	8	210	106
Total	823	18434	15911

nals. Blood pressure is considered to be high if it exceeds 140/90 mm Hg [5, 6]. We consider only systolic blood pressure (SBP) in this work and so we have the two classes of SPB $\geq$ 140 mm Hg (class H) and SPB $<$ 140 mm Hg (class L) giving a simple binary classification problem.

3. The data

We use the Aurora BP dataset [7] which includes PPG signals, blood pressure readings and Fitzpatrick skin tone class for a large number of subjects. There were 18,434 records from 823 subjects with recorded skin tone class and PPG signals as shown in Table 1. We note however that there is not an equal distribution of subjects and signals for all skin tones, with only a few subjects with darker skin tones. Each PPG record was around 15 s in length and was sampled at 500 Hz.

4. The features

We extracted features from each PPG record using the Symmetric Projection Attractor Reconstruction (SPAR) method [8]. The SPAR method is good for detecting changes in morphology and beat-to-beat variability in an approximately periodic signal such as the PPG.

All the signals were scaled in the vertical direction to have range 0 to 1 and then filtered using a 4^{th} order Chebyshev II filter. The SPAR method using three delay coordinates to construct a two-dimensional attractor from each PPG signal which has approximate three-fold rotational symmetry. All the points in the attractor were expressed in polar coordinates (r_i, θ_i) and then a probability density function was generated from all the r points and all the θ points [9]. We used 50 boxes for both the r density and the θ density. We used these r density values combined with the θ density values giving 100 features in total.

The SPAR method requires the computation of the average cycle length, which was found using autocorrelation after filtering the signals to remove baseline wander. The height of the first peak of the autocorrelation function is a simplistic measure of the quality of the signal. This peak has the value 1 if the signal is exactly periodic and decreases as the signal becomes less periodic. There is generally little useful information contained in a very noisy PPG signal and so we only used signals with peak value $\geq$ 0.7. This left us with 15,911 records as shown in the last column of Table 1. We note that the Aurora BP dataset includes a measure of optical quality. However, the correlation between the optical quality and our peak quality measure was poor.

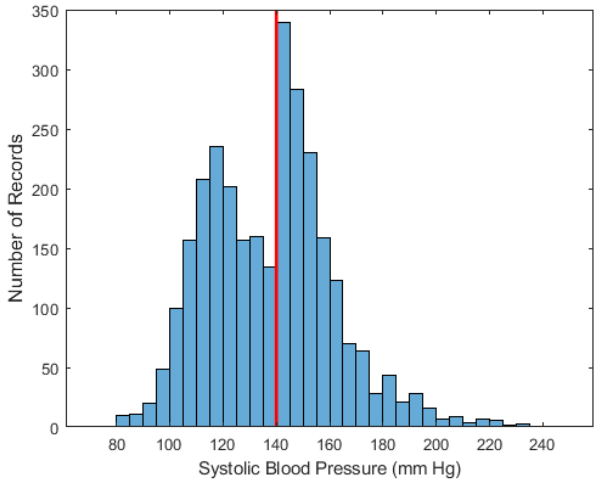

Figure 1. The distribution of systolic blood pressure values in the training data with skin tone class 1. The red line shows the division between the two classes.

5. Machine learning

We trained a machine learning model on records with Fitzpatrick skin tone class 1 and then tested the model on all the other records. There were 3,546 skin tone class 1 records in class L and 1,443 records in class H. In order to give a balanced dataset, we randomly selected 1,443 records in class L to use for training. The distribution of SBP values in this dataset is shown in Fig. 1. We used 5-fold cross validation *by subject* to ensure that all records for each subject were in the same fold.

For machine learning, we used a classification ensemble model [10] where the ensemble aggregation was performed by either bagging or boosting. The hyperparameters of the model were optimised by using Bayesian optimisation (100 iterations) of the cross validation accuracy. Using the optimal hyperparameter values, we then trained a machine learning model using all of the (balanced) data and used this model to test the records from skin tone classes 2-6.

We then created a dataset using all of the skin tones classes, again randomly selecting a subset of the class L records so that the number of records in the two classes was

the same for each skin tone class. Ideally, we would have liked data that consisted of an equal number of records of each skin tone class but our dataset is unbalanced across the skin tone classes and so this was not possible. Using this dataset, we performed 5-fold cross validation with folds that were stratified using skin tone class so that the distribution of skin tone class was similar in each fold.

6. Results

The accuracy of the cross validation performed on the skin tone class 1 data was 68.4 %. The accuracy when tested on the records associated with all the other skin tone classes is shown in Fig. 2. From this, it can be seen that there is a significant decline in accuracy for the other skin tone classes relative to class 1. Even for class 2, there is a 7.5 % drop in accuracy while for class 6, the reduction is 23.1 % resulting in an overall accuracy of just 45.3 %.

Taking class H as the positive class, the cross validation sensitivity for skin tone class 1 was 68.9 %. The sensitivity was similar for classes 2 and 3 while it exceeded 80 % for classes 4 and 6 but dropped to 43.9 % for class 5. On the other hand, the specificity for the training data was 67.9 % which decreased to below 60 % for classes 2 and 5 and below 50 % for classes 3, 4 and 6, which accounts for the corresponding decrease in accuracy.

The number of correctly classified records (green bars) and incorrectly classified records (stacked red bars) are shown in Fig. 3 for all values of SBP and for each skin tone class. We note from these plots that the number of incorrect classifications generally increases near to the 140 mm Hg threshold, which is not too surprising. We can also see why for skin tone classes 2-6 the sensitivity is high (except for class 5) and the specificity is low as the proportion of correctly classified records for class H is much higher than for class L.

For the classification using all the skin tone classes, the overall cross validation accuracy was 65.5 %. The accuracy by skin tone class is shown in Fig. 4. The accuracy for all the skin tone classes increased by between 4.3 % and 13.8 % compared to the results in Fig. 2 except for class 5 which had a small decrease. We note that class 6 increased in accuracy by 13.8 %.

7. Conclusions

Our results show that a machine learning model that is trained on skin tone class 1 records only and then tested on the remaining skin tone classes gives poorer accuracy for all the other skin tone classes, with the poorest performance for class 6 (darkest skin tone). When using data from all the skin tones in the training data, our results show an improvement in accuracy for all skin tones except for class 5, with the biggest improvement for class 6. The sensitivity for class 5 was very poor in both cases. The reason for this is not clear.

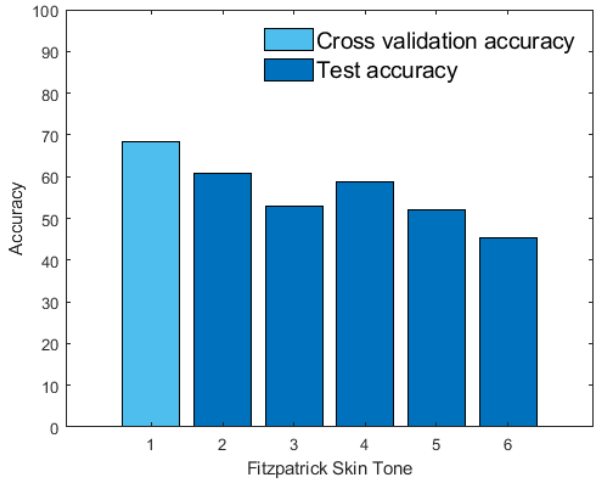

Figure 2. The accuracy when training on skin tone class 1 and testing on the other skin tone classes. For class 1, the cross validation accuracy on the training data is shown.

However, there are some issues with the dataset we have been using. We ensured that the training data for skin tone class 1 was balanced over the two blood pressure classes, but then tested all of the records for the other skin tone classes which all had many more records in class L than in class H. Since the accuracy for class H is generally higher than for class L, had the data been more balanced between the two blood pressure classes, the accuracy would probably have improved.

When training using data from all skin tone classes, we stratified the folds by skin tone class but it would have been far better to have had an equal number of records for each skin tone class. An alternative approach would be to train separate machine learning models for each skin tone class. The data was however balanced across the two blood pressure classes in this case.

Despite these caveats, our results still suggest that machine learning models should be trained using an equal number of records from all skin tone classes, and that training using records associated with mainly lighter skin tones would likely mean lower accuracy for darker skin tones.

Acknowledgments

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.

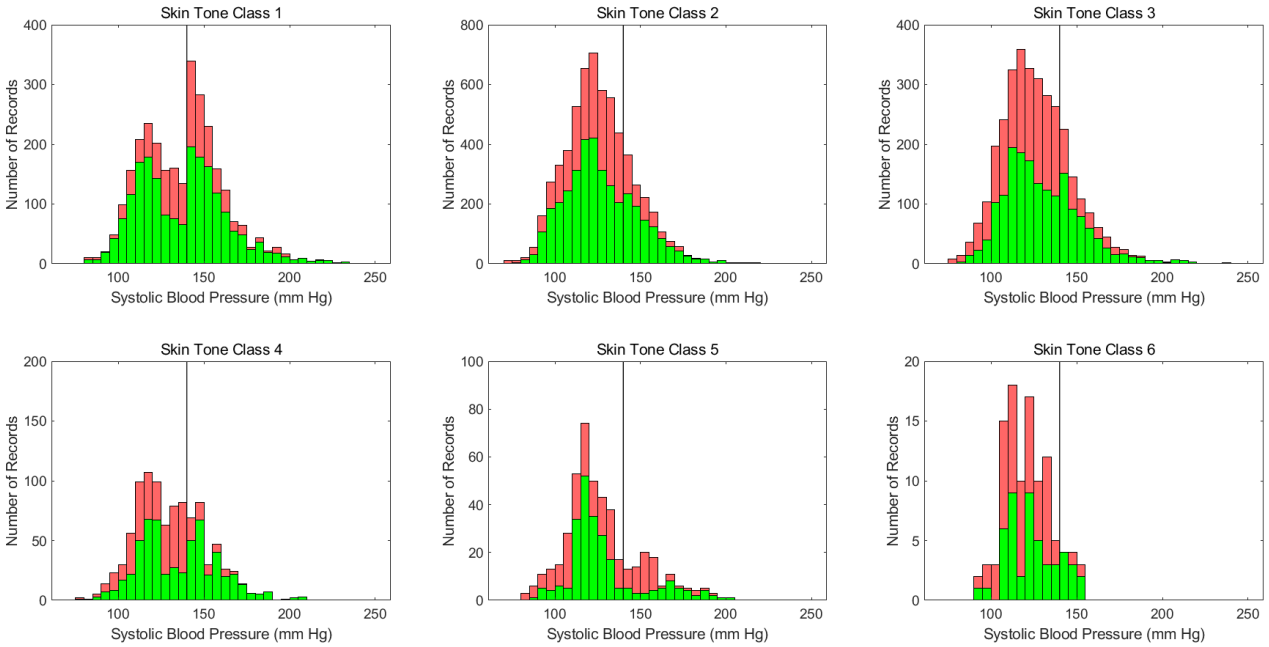

Figure 3. For each skin tone class, the green bars show the number of correct classifications in each range of SBP while the red bars on top show the number of incorrect classifications.

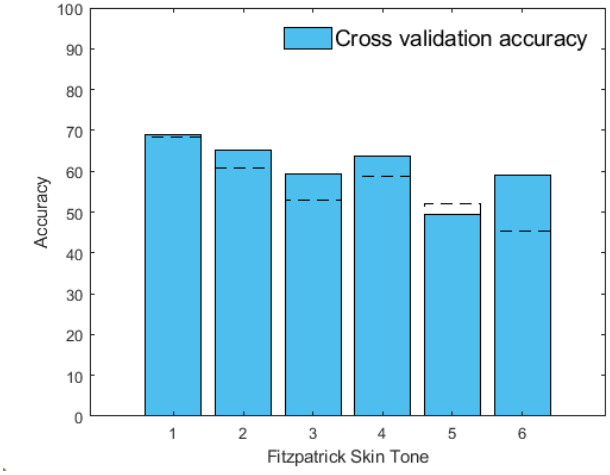

Figure 4. The accuracy for each skin tone class using cross validation on a dataset containing all skin tone classes. The accuracies from Fig. 2 are shown as dashed lines.

References

[1] Equity in medical devices: independent review, 2024. URL `https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report`.

[2] Mejía-Mejía E, Allen J, Budidha K, El-Hajj C, Kyriacou P, Charlton P. Photoplethysmography signal processing and synthesis. In *Photoplethysmography: Technology, Signal Analysis and Applications*, eds. P.A. Kyriacou, J.Allen. Academic Press, 2022; 69–145.

[3] Fitzpatrick T. The validity and practicality of sun-reactive skin types I through VI. Arch Dermatol 1988;124:869–871.

[4] Monk E. The Monk skin tone scale, 4 May 2023. URL `https://doi.org/10.31235/osf.io/pdf4c`.

[5] High blood pressure, 19 July 2024. URL `https://www.nhs.uk/conditions/high-blood-pressure/`.

[6] High blood pressure – understanding the silent killer, 11 May 2024. URL `https://www.fda.gov/drugs/special-features/high-blood-pressure-understanding-silent-killer`.

[7] Mieloszyk B, Twede H, Lester J, et al. A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. IEEE J Biomed Health Inf 2022; 26:2864–2875.

[8] Aston P, Christie M, Huang Y, Nandi M. Beyond HRV: Attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction. Phys Meas 2018; 39:024001.

[9] Lyle J, Nandi M, Aston P. Symmetric Projection Attractor Reconstruction: Sex differences in the ECG. Front Cardiovasc Med 2021;8:709457.

[10] Breiman L. Random forests. Mach Learn 2001;4:5–32.

Address for correspondence:

Prof Philip J. Aston

Data Science Department
National Physical Laboratory
Teddington TW11 0LW, UK
P.Aston@surrey.ac.uk

Department of Mathematics
University of Surrey
Guildford GU2 7XH, UK
Philip.Aston@npl.co.uk