Addressing Racial Disparities in Pulse Oximetry: A Machine Learning Perspective

Hakan B. Karli and Bige D. Unluturk
Department of Electrical and Computer Engineering, Michigan State University,
East Lansing, MI 48824, USA

2024 58th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-5405-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF60004.2024.10943105

***Abstract*— Pulse oximetry is a widely used non-invasive method for estimating arterial oxygen saturation (SaO_2). However, it has been shown to exhibit racial biases in its blood oxygen saturation (SpO_2) measurements, primarily due to differences in skin color, which can lead to misdiagnosis and improper treatment. This study utilizes the publicly available BOLD dataset on PhysioNet to investigate these biases. It employs machine learning models, including Random Forest, Gradient Boosting, and XGBoost regressors, to predict SaO_2 from SpO_2 readings, incorporating interpersonal parameters such as race, with an initial focus on the most distinguishable groups by skin color: Black and White. The developed models significantly improve prediction accuracy, reducing the Mean Squared Error from 30.76 (baseline) to 4.72. SHAP analysis underscores the critical role of race as a surrogate for skin tone in enhancing predictions. These findings emphasize the need for equitable signal processing techniques for medical devices to decrease the bias and improve health outcomes across diverse populations.**

I. Introduction

Pulse oximetry is a cornerstone in non-invasive monitoring of peripheral oxygen saturation (SpO_2) and is widely used in clinical settings to assess patients' oxygenation status. SpO_2 readings are commonly used as an estimate of arterial oxygen saturation (SaO_2), providing clinicians with a convenient method to infer blood oxygen levels without the need for invasive procedures. However, recent studies have highlighted significant limitations in the accuracy of pulse oximeters for individuals with darker skin tones. In these cases, SpO_2 readings tend to overestimate the true arterial oxygen saturation (SaO_2), leading to what is termed "occult hypoxemia." This condition occurs when SpO_2 readings fail to detect dangerously low oxygen levels, posing a particular concern in urgent and high-stakes clinical contexts where precise monitoring is critical [1].

The root of this disparity lies in the reliance of pulse oximeters on light absorption technology influenced by melanin levels in the skin [2]. Melanin, present in higher concentrations in individuals with darker skin tones, absorbs a portion of the emitted light, thereby reducing the light signal reaching the detector and causing inaccuracies in SpO_2 readings as illustrated in Figure 1. Devices calibrated predominantly on lighter skin tones, can yield inaccurate SpO_2 measurements in individuals with darker skin, obscuring hypoxemic conditions that require timely intervention [3]. Multiple studies have documented this skin color dependency in pulse oximetry accuracy, affirming that increased melanin pigmentation can lead to overestimation of oxygen saturation [4]–[6]. For instance, [7] demonstrated that low perfusion levels exacerbate inaccuracies in pulse oximetry for patients with darkly pigmented skin. Furthermore, recent reviews and modeling studies emphasize that the spectral characteristics of light sources in pulse oximeters inadequately account for the optical properties of darker skin, underscoring the need for device recalibration or alternative monitoring approaches [8], [9].

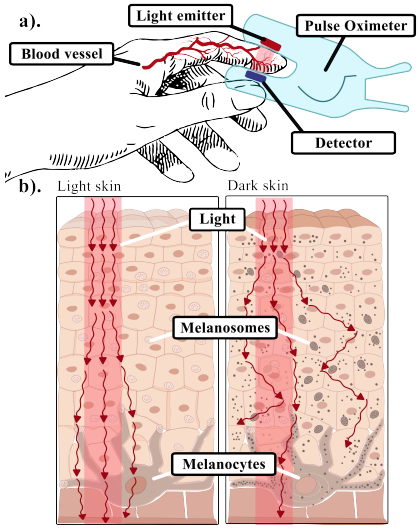

Fig. 1: a). Pulse oximeter placement. b). Melanin in melanosomes—produced by melanocytes—affects signal accuracy, especially in darker skin, which has larger and more numerous melanosomes.

An approach by [10] attempted to correct SpO_2 readings by estimating SaO_2; however, this method has several limitations. First, the model did not fully address dataset bias, as the data included underrepresented groups, potentially skewing results. Second, some features used to correct SpO_2 could only be obtained through invasive measures such as blood samples, contradicting the main advantage of pulse oximeters being non-invasive.

In this study, we address these limitations by developing a model that predicts SaO_2 levels more accurately using SpO_2

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant OAC-2203827, in part by the National Institutes of Health (NIH) under Grant 1R01HL172293-01 and in part by 1R21EB036329-01. Corresponding author: Bige Deniz Unluturk.

readings, while accounting for interpersonal parameters and the time gap between these measurements. This approach aims to enhance the robustness of oxygen saturation monitoring by accounting for skin color dependencies that affect the accuracy of traditional SpO_2 readings. By refining this model, we strive to ensure that SpO_2 measurements are more reliable for individuals of all skin tones.

The organization of this paper is as follows. Section II introduces the dataset and pre-processing steps. Section III elaborates on the methodology for the development of the model. Following that, Section IV presents our findings, discussing the results obtained from the model's output. Finally, in Section V, we provide concluding remarks.

II. Dataset and Processing

We utilized the BOLD (Blood-gas and Oximetry Linked Dataset) [11], which offers a comprehensive collection of patient demographics, physiological measurements, and laboratory results from Intensive Care Unit (ICU) stays. The dataset was created by combining three large Electronic Health Record (EHR) databases: MIMIC-III, MIMIC-IV, and eICU-CRD. It includes 49,099 paired measurements of SpO_2 and SaO_2 within a strict 5-minute window, with oxygen saturation levels ranging between 70% and 100%.

The BOLD dataset comprises 49,099 pairs representing 44,907 patients. The dataset includes detailed demographic information such as age, sex, race, weight, and height. Clinical parameters include vital signs such as temperature, blood pressure, heart rate, respiratory rate, and SpO_2 readings; laboratory values such as arterial blood gas measurements with SaO_2 levels, complete blood count, coagulation profile, basic metabolic panel, hepatic function panel, and enzyme levels; and clinical scores like the Sequential Organ Failure Assessment (SOFA) scores.

All time-varying data are aligned with each (SaO_2, SpO_2) pair using left-sided windows to include only past data relative to the SaO_2 timestamp. This temporal alignment ensures that the dataset reflects the patient's clinical state at the time of SaO_2 measurement.

A. Preprocessing and Balancing

We addressed data quality issues through preprocessing, focusing on outliers and missing data.

1) Data Cleaning: Outliers included implausible values such as a body mass index (BMI) exceeding 220,000 kg/m^2 and a pH value of 70. To correct these, we capped BMI values at a maximum of 100 kg/m^2 and corrected extreme pH values into physiologically reasonable ranges. For missing data, columns with more than 50% missing values were dropped, reducing the feature set to 122 variables. Numerical variables with missing data were replaced using the mean or median based on their distribution and skewness, while categorical variables were filled with the most common category value.

2) Labeling: To focus on clinically relevant cases, we filtered the dataset to include SaO_2 values between 85% and 99%. SaO_2 values below 85% are often associated with more severe hypoxemia [12], which may present with noticeable symptoms [13]. By concentrating on the 85% to 99% range, we aim to identify patients who may not exhibit obvious symptoms, yet are at risk of hypoxemia. SaO_2 levels were categorized into two groups: *Treatment Needed* for values between 85% and 90% and *Treatment Not Needed* for values between 90% and 99%. Similarly, SpO_2 readings were labeled to reflect treatment suggestions based on pulse oximetry. According to clinical guidelines, SpO_2 values below 93% indicate that treatment may be necessary [14], and these were labeled as *Treatment Suggested*, while values of 93% and above suggest no immediate treatment and were labeled as *No Immediate Treatment Suggested*. By labeling SaO_2 and SpO_2, we evaluated discrepancies between these measurements to highlight potential misdiagnoses from pulse oximetry inaccuracies.

3) Balancing: To ensure our model could effectively learn from the data, we performed balancing procedures. We focused on patients identified as Black or White, as these groups have distinct skin color differences relevant to our study. We equalized the number of instances in the "Treatment Needed" and "Treatment Not Needed" categories based on SaO_2 levels, as explained in the II-A.2 section. This approach mitigates potential biases due to class imbalance and enhances the model's ability to generalize across different subgroups.

III. Methodology

A. Feature Selection

Our goal was to predict SaO_2 values using interpersonal parameters and clinical measurements that are non-invasive and known to affect SpO_2 readings. We selected features based on their relevance to oxygen saturation levels and their availability in the dataset.

Race (White or Black) was used as a surrogate for skin color differences, addressing documented discrepancies in pulse oximetry accuracy due to variations in light absorption [15], [16]. Anthropometric measurements such as weight and height were included in the analysis, recognizing that obesity and body habitus can significantly influence oxygen saturation levels. Prior studies have shown that increased body mass index (BMI) correlates with lower resting oxygen saturation [17], [18].

Age was considered an important feature, as age-related changes in cardiovascular and respiratory function can influence oxygen delivery and utilization. For instance, older individuals often experience reduced oxygen transport efficiency, making age a relevant factor for SaO_2 prediction [19].

Sex was included due to physiological differences between males and females, such as hormonal variations and differing hemoglobin levels. These factors can result in variations in oxygen transport and saturation levels [19].

SpO_2 readings were used as direct inputs for predicting SaO_2 levels. Given the inherent limitations of pulse oxime-

try in accurately reflecting arterial oxygen saturation, these readings were critical in bridging the gap between non-invasive measurements and SaO_2 estimations. Lastly, the time difference between SpO_2 and SaO_2 measurements was incorporated to account for potential physiological changes that could occur between the two measurements.

By focusing on these non-invasive features, which are known to influence oxygen saturation readings, we aimed to develop a model that could more accurately predict SaO_2 levels across diverse patient populations.

B. Model Selection

We experimented with several machine learning algorithms to identify the most effective model for predicting SaO_2 levels. The models tested included linear regression, support vector regressor (SVR), decision tree regressor, k-nearest neighbors (KNN), gradient boosting regressor, XGBoost regressor, and random forest regressor.

For each model, we performed hyperparameter optimization to fine-tune performance. Parameters such as learning rate, maximum depth, and the number of estimators were adjusted to enhance predictive accuracy. The dataset was divided into training and testing sets; the training set was used to fit the models, while the testing set was reserved for evaluating predictive performance.

We evaluated the models using several metrics to provide a comprehensive assessment: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 Score (Coefficient of Determination). Models achieving lower MSE, RMSE, and MAE values, along with higher R^2 scores, were considered more effective in predicting SaO_2 levels. The final model selection was based on a combination of these performance metrics, interpretability, and practicality for clinical application.

IV. Results and Discussion

After preprocessing the data and selecting relevant features, we trained and evaluated multiple machine learning models to predict SaO_2 levels using non-invasive parameters. This section presents the performance metrics of each model and discusses the implications of our findings, particularly the improvement over measured SpO_2 and the impact of incorporating interpersonal parameters such as race.

A. Model Performance

We first established a baseline by examining the error between the measured SpO_2 and SaO_2 values without any correction. The baseline Mean Squared Error (MSE) was 30.76, indicating significant discrepancies between SpO_2 and SaO_2 readings. This high error underscores the limitations of traditional pulse oximetry, which often overestimates oxygen saturation levels, especially in patients with darker skin tones.

We then trained several machine learning models, applying hyperparameter optimization and scaling techniques to enhance their performance. Table I summarizes the results, including the best scaler used, MSE, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score.

TABLE I: Model Performance Metrics

Model	Best Scaler	MSE	RMSE	MAE	R^2 Score
Original Data (Baseline)	None	30.7553	5.5457	4.1488	-0.3323
Linear Regression	Min-Max	16.4068	4.0505	3.2759	0.2893
Support Vector Regressor	Standard	11.1839	3.3442	2.3071	0.5155
Decision Tree Regressor	Min-Max	8.8471	2.9744	1.6978	0.6167
K-Nearest Neighbors	Robust	6.4684	2.5433	1.3523	0.7198
Gradient Boosting Regressor	Standard	5.6347	2.3737	1.5109	0.7559
XGBoost Regressor	Min-Max	5.2142	2.2835	1.2350	0.7741
Random Forest Regressor	Robust	5.1346	2.2660	1.3505	0.7776

The results show a substantial reduction in error when using machine learning models compared to the baseline. Models such as Random Forest, XGBoost, and Gradient Boosting Regressor achieved the lowest MSE and RMSE values, along with the highest R^2 scores. Specifically, the XGBoost Regressor demonstrated strong performance with an MSE of 5.21 and an R^2 score of 0.77.

To further enhance the models, we focused on the top three performers: Random Forest, XGBoost, and Gradient Boosting Regressor, and applied additional hyperparameter optimization. Table II presents the optimized results.

TABLE II: Optimized Model Performance Metrics

Model	Best Scaler	MSE	R^2 Score
Original Data (Baseline)	None	30.7553	-0.3323
Random Forest Regressor	Robust	5.1346	0.7776
Gradient Boosting Regressor	Robust	4.7792	0.7930
XGBoost Regressor	**Standard**	**4.7164**	**0.7957**

After hyperparameter tuning, the XGBoost Regressor slightly outperformed the other models. For training, the cross-validation results showed a mean MSE of 5.64 ($\pm$0.47) and an R^2 mean of 0.76 ($\pm$0.02) over 5 folds, indicating robust performance across different subsets of the data. The model achieved a test MSE of 4.72 and an R^2 score of 0.80. The best parameters selected for the XGBoost Regressor were a learning rate of 0.03, maximum depth of 18, 250 estimators, a subsample ratio of 0.646, and the use of the standard scaler. These optimized parameters allowed the XGBoost model to capture complex patterns in the data more effectively, leading to improved predictive performance.

B. Comparison of SpO_2 vs. SaO_2 and Corrected SpO_2 vs. SaO_2

To visually illustrate the improvement achieved by our model, we compared the traditional SpO_2 readings with SaO_2 measurements and contrasted them with the corrected SpO_2 values predicted by our XGBoost model.

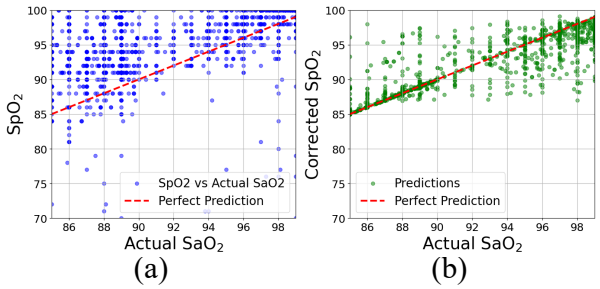

Fig. 2: Comparison of SpO_2 vs. SaO_2. (a) Measured SpO_2 vs actual SaO_2, (b) Corrected SpO_2 vs actual SaO_2, showing improved alignment.

In Figure 2-a, we observe a significant error between SpO_2 and SaO_2 values. The scatter plot shows considerable deviation from the ideal 1:1 line, highlighting the limitations of traditional pulse oximetry, which often overestimates oxygen levels, particularly in patients with darker skin tones.

In Figure 2-b, we compare our model's corrected SpO_2 with the SaO_2 values. The predictions closely align with the actual values along the 1:1 line, demonstrating that the model performs significantly better than using SpO_2 alone. By incorporating interpersonal parameters, our model is able to provide more accurate and reliable SaO_2 estimates.

C. Bland-Altman Analysis

To further assess the agreement between the measurements, Bland-Altman analysis was performed to compare SpO_2 vs. SaO_2 and corrected SpO_2 vs. SaO_2.

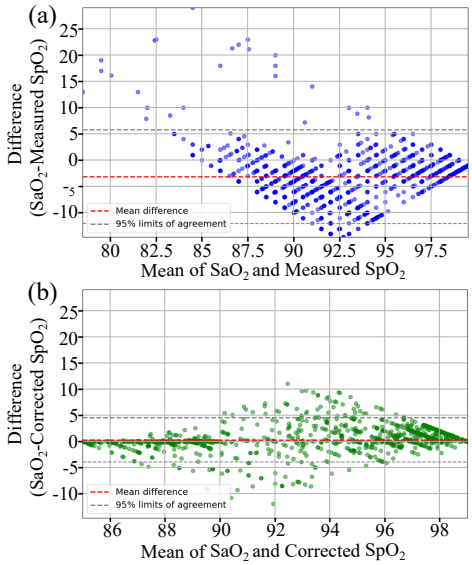

Fig. 3: Bland-Altman plots comparing (a) measured SpO_2 vs. SaO_2 and (b) corrected SpO_2 vs. SaO_2.

In Figure 3(a), we compare SpO_2 vs. SaO_2. The y-axis represents the difference between SaO_2 and SpO_2, and the x-axis represents the average of the two measurements. The plot shows a significant spread of differences, indicating large discrepancies between SpO_2 and SaO_2 values. This highlights how SpO_2 often overestimates oxygen saturation, leading to potential misclassification of hypoxemia.

In Figure 3(b), we compare the corrected SpO_2 vs. SaO_2 using our model. The differences here are much smaller, with most data points tightly clustered around the zero line. This indicates that the model provides a much closer prediction of SaO_2, improving the agreement between non-invasive measurements and arterial blood gas analysis.

D. SHAP Value Analysis

To understand each feature's influence on the model's predictions, we computed SHAP (SHapley Additive exPlanations) values. Figure 4 shows their distribution, capturing both the variability and direction of feature contributions.

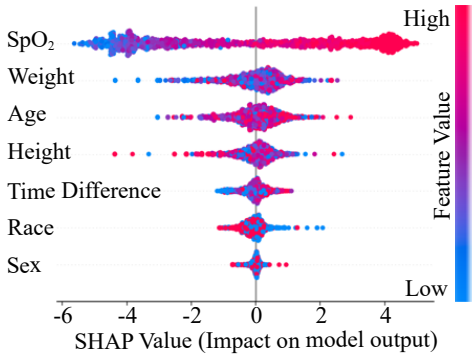

Fig. 4: Feature importance rankings using SHAP values

Figure 4 illustrates that each dot represents a SHAP value for an instance, with the color indicating the feature value, where red corresponds to high values and blue to low values. Features are listed on the y-axis by importance. The x-axis displays the SHAP value, which quantifies how much each feature contributes to increasing or decreasing the prediction.

SpO_2 has the highest mean absolute SHAP value, indicating it is the most significant feature in the model's predictions. Weight, age, and race also have substantial impacts on the predictions. Weight and age influence the predictions in both positive and negative directions, but with less impact than SpO_2. Notably, the race feature shows a significant effect, with different racial groups affecting SaO_2 predictions differently.

The SHAP analysis reveals that while clinicians traditionally consider factors like weight, age, and sex when assessing oxygen levels, our model shows that skin color, specifically race as a surrogate for skin color between Black and White individuals, has an even stronger effect than sex. This suggests that race should be given more weight in future decision-making, especially when it comes to assessing oxygen saturation in patients.

E. Confusion Matrix Analysis

To evaluate our model's practical implications, we analyzed confusion matrices comparing the classification of patients needing treatment based on SpO_2 readings versus the actual SaO_2 status, using the labeling definitions from the II-A.2 section.

Figure 5(a) represents the results of using only pulse oximetry measurements. We observe that 345 patients who actually needed treatment were misclassified as not needing it. This accounts for 26.25% of patients who required intervention but might not receive timely treatment if relying solely on SpO_2 readings. In contrast, Figure 5(b) shows the confusion matrix using our model's corrected SpO_2 values. Here, only 49 patients who needed treatment were misclassified, reducing the error rate to 3.4%. This reduction demonstrates our model minimizes under-treatment risk, ensuring patients needing intervention are reliably identified.

1) Analysis by Race: To further investigate the model's performance across different racial groups, we analyzed confusion matrices separately for Black and White patients.

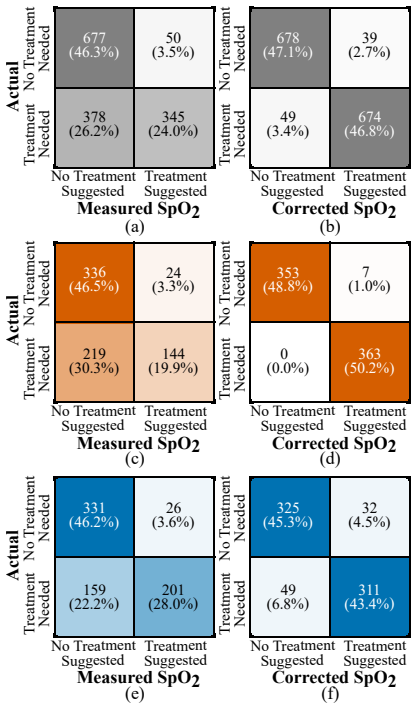

Fig. 5: Confusion Matrices for SpO_2-Suggested Treatment vs. Actual Treatment Needed and Corrected SpO_2-Suggested Treatment vs. Actual Treatment Needed. (a-b) Both Black and White patients (c-d) Black patients only. (e-f) White patients only.

In the confusion matrices for Black patients (Figures 5(c) and 5(d)), we observe that the number of false negatives (patients who needed treatment but were not identified) decreased from 219 to 0 when using the corrected SpO_2 values from our model. This indicates a significant improvement in identifying Black patients who require treatment.

Similarly, for White patients (Figures 5(e) and 5(f)), the number of false negatives decreased from 159 to 49 with the corrected SpO_2 values. Although the improvement is substantial, the model still misclassified some White patients who needed treatment.

These results suggest that our model enhances the detection of patients needing treatment across both racial groups, with a more pronounced improvement for Black patients.

V. Conclusion

This study shows that incorporating interpersonal parameters such as race, age, weight, and height into machine learning models significantly improves SaO_2 prediction accuracy over traditional SpO_2 readings. By using race as a surrogate for skin color, focusing on Black and White patients, the model addresses longstanding disparities in pulse oximetry accuracy. The improved alignment between predicted and actual SaO_2 values, as shown through Bland-Altman analysis, highlights the model's ability to correct biases inherent in SpO_2 readings. Additionally, SHAP value analysis underscores the importance of race as a critical predictor, reflecting the impact of skin pigmentation on light-based measurements.

The confusion matrix analysis further demonstrates the medical significance of the model, particularly its reduction of false negatives among Black patients. The corrected SpO_2 values provide a more reliable identification of individuals needing treatment, minimizing the risk of under-treatment. These findings emphasize the potential of machine learning to enhance diagnostic reliability, address disparities in medical devices, and contribute to more equitable healthcare outcomes across diverse populations, particularly in the detection and management of hypoxemia.

References

[1] M. W. Sjoding *et al.*, "Racial bias in pulse oximetry measurement," *New England Journal of Medicine*, vol. 383, no. 25, pp. 2477–2478, 2020.

[2] M. Bermond *et al.*, "Reducing racial bias in spo2 estimation: The effects of skin pigmentation," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023, pp. 1–5.

[3] P. E. Bickler *et al.*, "Effects of skin pigmentation on pulse oximeter accuracy at low saturation," *Anesthesiology*, vol. 102, no. 4, pp. 715–719, 2005.

[4] A. Fawzy *et al.*, "Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with covid-19," *JAMA Internal Medicine*, vol. 182, no. 7, pp. 730–738, 2022.

[5] A. L. Ries *et al.*, "Skin color and ear oximetry," *Chest*, vol. 96, no. 2, pp. 287–290, 1989.

[6] A. Jubran and M. J. Tobin, "Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients," *Chest*, vol. 97, no. 6, pp. 1420–1425, 1990.

[7] J. R. Feiner *et al.*, "Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: The effects of oximeter probe type and gender," *Anesthesia & Analgesia*, vol. 105, no. 6 Suppl, pp. S18–S23, 2007.

[8] S. Pernitez-Agan *et al.*, "Modeling skin pigmentation effects in reflectance pulse oximetry," *Sensors*, vol. 21, no. 4, p. 1116, 2021.

[9] W. P. Hannan *et al.*, "Racial bias in pulse oximetry measurement among patients about to undergo surgery," *Anesthesiology*, vol. 134, no. 6, pp. 892–898, 2021.

[10] R. I. Matos *et al.*, "Shining light on dark skin: Correction models for pulse oximetry inaccuracies in patients with darker skin tones," *Journal of Biomedical Engineering*, vol. 50, no. 2, pp. 245–259, 2023.

[11] J. a. Matos *et al.*, "Bold, a blood-gas and oximetry linked dataset (version 1.0)," https://doi.org/10.13026/phvt-3277, 2023, physioNet.

[12] I. Tyler *et al.*, "Continuous monitoring of arterial oxygen saturation with pulse oximetry during transfer to the recovery room," *Anesthesia and Analgesia*, vol. 64, no. 11, pp. 1108–1112, Nov 1985.

[13] R. P. Cafaro, "Hypoxia: Its causes and symptoms," *Journal of the American Dental Society of Anesthesiology*, vol. 7, no. 4, pp. 4–8, 1960.

[14] J. A. Sobel *et al.*, "Descriptive characteristics of continuous oximetry measurement in moderate to severe covid-19 patients," *Scientific Reports*, vol. 13, no. 1, p. 442, January 2023.

[15] M. W. Sjoding *et al.*, "Racial bias in pulse oximetry measurement," *New England Journal of Medicine*, vol. 383, no. 25, pp. 2477–2478, 2020.

[16] V. S. Valbuena *et al.*, "Racial bias in pulse oximetry measurement among patients about to undergo extracorporeal membrane oxygenation in 2019-2020: A retrospective cohort study," *Chest*, vol. 161, no. 4, pp. 971–978, 2022.

[17] Y. Xiong *et al.*, "Accuracy of oxygen saturation measurements in patients with obesity undergoing bariatric surgery," *Obesity Surgery*, vol. 32, pp. 3581–3588, 2022.

[18] V. K. Kapur *et al.*, "Obesity is associated with a lower resting oxygen saturation in the ambulatory elderly: results from the cardiovascular health study," *Respiratory Care*, vol. 58, no. 5, pp. 831–837, May 2013.

[19] A. Geirsdottir *et al.*, "Retinal vessel oxygen saturation in healthy individuals," *Investigative Ophthalmology & Visual Science*, vol. 53, no. 9, pp. 5433–5442, 2012.