# Paper AR03

# Harnessing AI for Health Equity and Inclusivity in Risk-Based Monitoring

Finn Janson, Roche, Welwyn Garden City, United Kingdom

**Abstract**
This research addresses bias in pulse oximetry readings, particularly in patients with darker skin pigmentation by utilizing Artificial Intelligence (AI) using the MIMIC-4 hospital dataset. Pulse oximetry (SpO2) is essential in monitoring arterial oxygen saturation (SaO2) in various clinical trials and healthcare settings. This paper seeks to refine the precision of SpO2 to inform Quality Tolerance Limits (QTLs) of SaO2 across diverse patient populations. By using boundaries derived from errors in pulse oximetry readings, by using AI trained on SaO2 levels, we can ensure data quality and effective monitoring.

**Introduction**
Pulse oximetry is a critical tool in modern healthcare, providing non-invasive measures of arterial oxygen saturation (SaO2). They are used in clinical trials to monitor oxygen saturation levels in study participants across various research areas (Cabanas et al., 2022), in order to gauge safety and efficacy in randomized controlled trials for new drugs with respiratory or cardiovascular effects, medical device trials, or experimental treatments, such as those targeting low oxygen levels in illnesses like acute respiratory distress syndrome or traumatic brain injury (Yin et al., 2022). Machine learning has previously been used to predict or calibrate measurement in trials (Ren et al., 2022) and stratify patients based on risk (Jentzer et al., 2021).

A systematic review found that pulse oximetry (SpO2) overestimates oxygenation more commonly in patients of racial and ethnic minority groups, leading to variations in treatment (Cabanas et al., 2022). . When comparing oxygen saturation measures across races, one particular study found significant overestimation from Black patients compared to White patients (Sjoding et al., 2020). These findings are consistent with other studies addressing pulse oximeters in vulnerable populations (Gottlieb et al., 2022).

In a trial setting, placing Quality Tolerance Limits (QTLs) on SpO2 levels and designating oxygen saturation estimation as a Key Risk Indicator (KRI) enables enhanced patient safety and data integrity (Bhagat et al., 2021; Wolfs et al., 2023). This approach works effectively with continuous access to oxygen saturation levels, crucial for responsively modifying protocols and facilitating timely medical interventions. The challenge, however, lies in setting appropriate QTLs given the unreliable data quality from SpO2 readings.

**Objectives**
An unbiased metric for SpO2 readings, as well as an understanding of performance across demographics, enhances standards of patient monitoring and care when considering the risks associated with inaccurate oxygen saturation estimations (Matos et al., 2023a). To achieve this, a predictive model will be trained to predict arterial oxygen saturation (SaO2) measurements based on SpO2 readings and other relevant features. The model's predictions will be used to inform Quality Tolerance Limits (QTLs). This approach is intended to improve the precision of SpO2 readings, effectively reduce error rates

across diverse demographics, and contribute to safer and more effective patient monitoring. Successfully achieving these objectives will be vital for accurately detecting conditions such as hidden hypoxemia, improving patient care and safety in populations most affected by current pulse oximetry inaccuracies.

## Methodology

### Data Sources

This study utilized data from the Medical Information Mart for Intensive Care IV (MIMIC-IV v2.2), a comprehensive and publicly available dataset of de-identified electronic health records. This was chosen due to its large size, variation in demographics, and its features that relate to those used or collected in clinical trials (Matos et al., 2023b).

### Sample size

There were 50,000 unique patients who were admitted to the Beth Israel Deaconess Medical Center in Boston, MA, during the years 2008 to 2019. The dataset covers 425,000 ED stays from 2011 to 2019. This includes critical information such as patient demographics, specific treatment information, SOFA scores, physiological measurements, vital signs, triage details, medication records, and discharge diagnoses.

### Data access

To maintain the privacy and security of patient information, all data in the MIMIC-IV dataset have been de-identified in compliance with the HIPAA Safe Harbor provision. To utilize this rich dataset, one must become a credentialed user by fulfilling ethical and regulatory requirements and signing a data use agreement.

### Ethical considerations

Despite the best intentions, bias still rears its head into any potential decision we make, even while trying to mitigate it. Limitations on the data available lead to a heavy class imbalance and also requires a manual labeling of skin pigmentation based on demographic priors (Gottlieb et al., 2022). When gathering patient data in a trial, it will require a more scientific process for identifying the skin pigmentation.

## Preprocessing

SaO2 and SpO2 values outside the standard physiological range (70-100%) were excluded, based on values provided by the Critical Datathon curator (Matos et al., 2023a). Records were eliminated where the time delta between SpO2 and SaO2 readings exceeded 5 minutes, ensuring temporal relevance between measurements for validation. Subjects with over 50% missing data were excluded from the sample. Subjects with unknown or ambiguous race were excluded, leaving only those that can be assessed for demographic bias. Races which appear an order of magnitude smaller than the modal race were excluded to avoid small numbers fallacy. SpO2 and SaO2 had no missingness; and missing values for the remaining subjects were imputed using mean imputation for continuous variables and mode imputation for categorical variables.

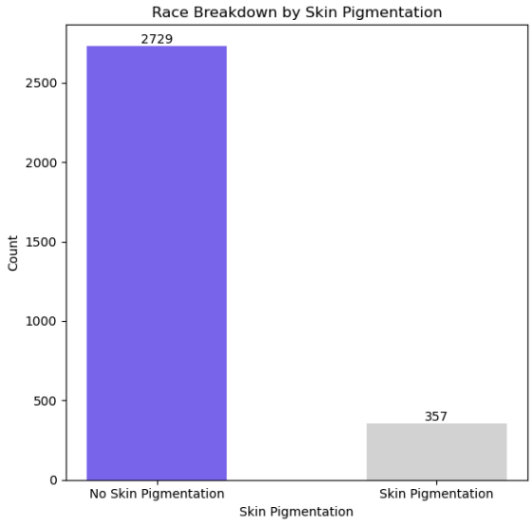
### Labeling

Since skin pigmentation was not a field included, they were broken down into two classes, based on priors that show risk associated with race. Those labeled “darker skinned” were mostly African or African American or Hispanic (Gottlieb et al., 2022), while those labeled “not” were primarily White and European.

### Features

The total number of potential features was constrained to 16 based on model performance considerations. Categorical features were one-hot encoded. From this candidate pool, features were selected, across categorical (encoded) features and numerical features using ANOVA F-value, based on their association with SaO2 target (from train set only to avoid data leakage)

when tested for association with the outcomes on the training data. Redundant features were further trimmed by removing variables that demonstrated high collinearity (correlation above 40%) with other retained features. Categorical features were one-hot encoded and numerical features underwent standardization to ensure comparable scales when modeling.




Following processing and a train-test split (70% used in training), the training set comprised 12,202 samples, while the test set contained 3,086 samples, divided into 5 k-folds for hyperparameter optimization.

Figure 1: This figure shows the sample size is 2729 of patients in the training data set without dark skin pigmentation and 357 with dark skin pigmentation, with remaining 6030 samples in the training set not classified in either subgroup due to ambiguity.

**Model choices**

CatBoostRegressor uses an efficient encoding scheme for categorical features that improves performance, especially with many high-cardinality categories. CatBoost is designed to be fast even with large datasets and has potential to be deployed in real world systems. It is designed to balance model complexity and overfitting through regularization techniques like ordered boosting, which grows trees level-wise to prevent overfitting by limiting model peeking at labels. This more balanced boosting approach can reduce bias compared to other gradient boosting methods (Prokhorenkova et al., 2019).

**Fairness objective function**

A custom objective function was used to ensure that the model performs well for these groups without overfitting to any specific race (which would harm generalization) or underfitting across all races (which would lead to a model that is too simplistic and not useful). Initially, the CatBoost model uses MSE in order to fit the dataset. This function is used during hyperparameter tuning, by computing Mean Squared Error (MSE) from patients with darker skin pigmentations, then squaring the number of samples when calculating the weighted MSE to amplify the influence of small sample sizes within the darker skin pigmentation patient group. By using the square of the sample size as a weighting factor, the algorithm assigns a disproportionately higher weight to the errors from smaller patient groups. This helps to ensure that the model pays sufficient attention to these groups, which might otherwise be underrepresented in the error metric due to their smaller sample sizes.

**Hyperparameter Tuning**

Hyperparameter optimization was conducted to fine-tune our models using the fairness objective function, ensuring the best possible predictive performance. Bayesian optimization was chosen for its time-cost efficiency. This was used to find optimal values for learning rate, max depth, subsampling and number of estimators.

**Results**

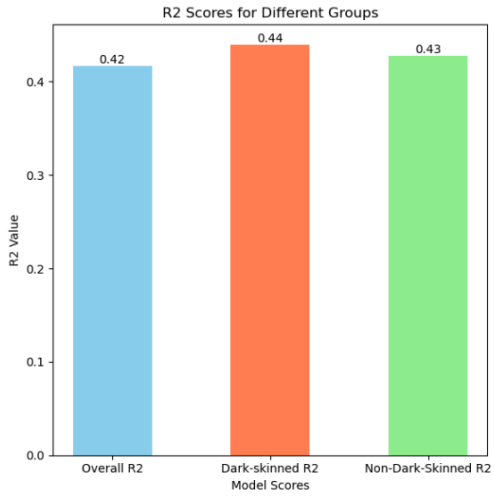



Figure 2: This figure shows R2 across different subgroups. Overall R2 for the entire dataset, without any subgroup segmentation, shows that 42% of the variability in the dependent variable can be explained by the model, with similar results across the two different classes.

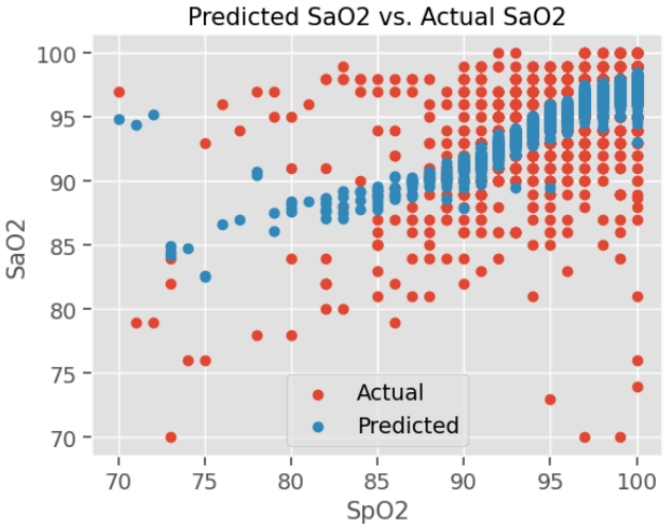



Figure 3: This figure illustrates the relationship between measured peripheral oxygen saturation (SpO2) and arterial oxygen saturation (SaO2). Two sets of data points are shown, one representing model predictions for SaO2 values and the other representing SpO2 readings used conventionally to estimate SaO2.

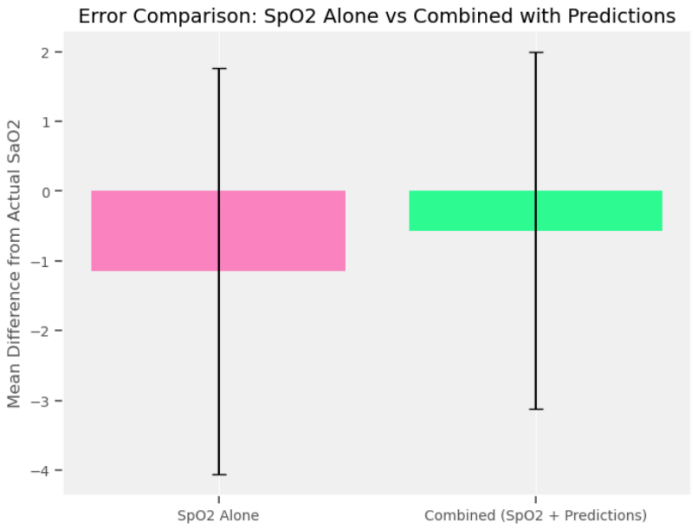



Figure 4: This figure shows the mean differences from actual SaO2 and predicted SaO2, illustrating how model predictions reduce the mean error rate, showing how predictions are less likely to overestimate oxygen saturation levels.

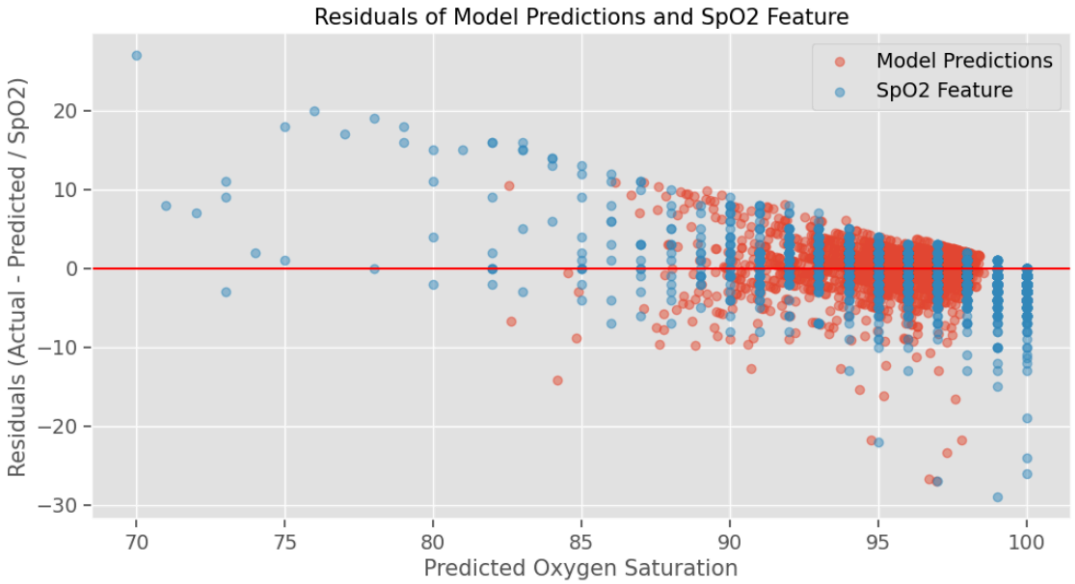



Figure 5: This figure shows residual plots comparing actual SaO2 values against residuals from model predictions and the SpO2 feature.

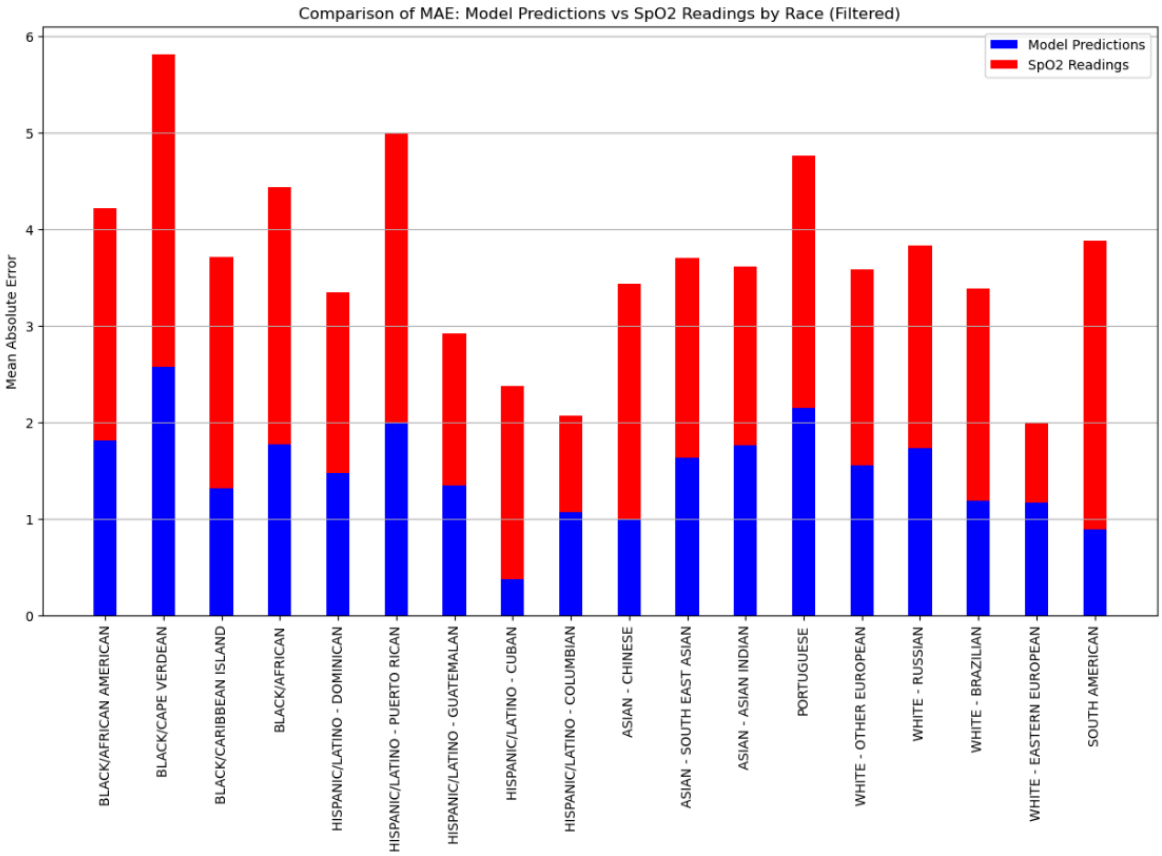



Figure 6: This figure displays MAE for each racial subgroup as a stacked bar chart, showing MAE for: This figure displays MAE for each racial subgroup as a stacked bar chart, showing MAE for each, from model and from SpO2 feature, Candidates for this chart were either classified as likely to have darker-skinned based on race feature. Across races flagged up as both likely to be darker-skinned and those not, the model is able to successfully reduce the MAE from the SpO2 reading across demographics from these two polarities.

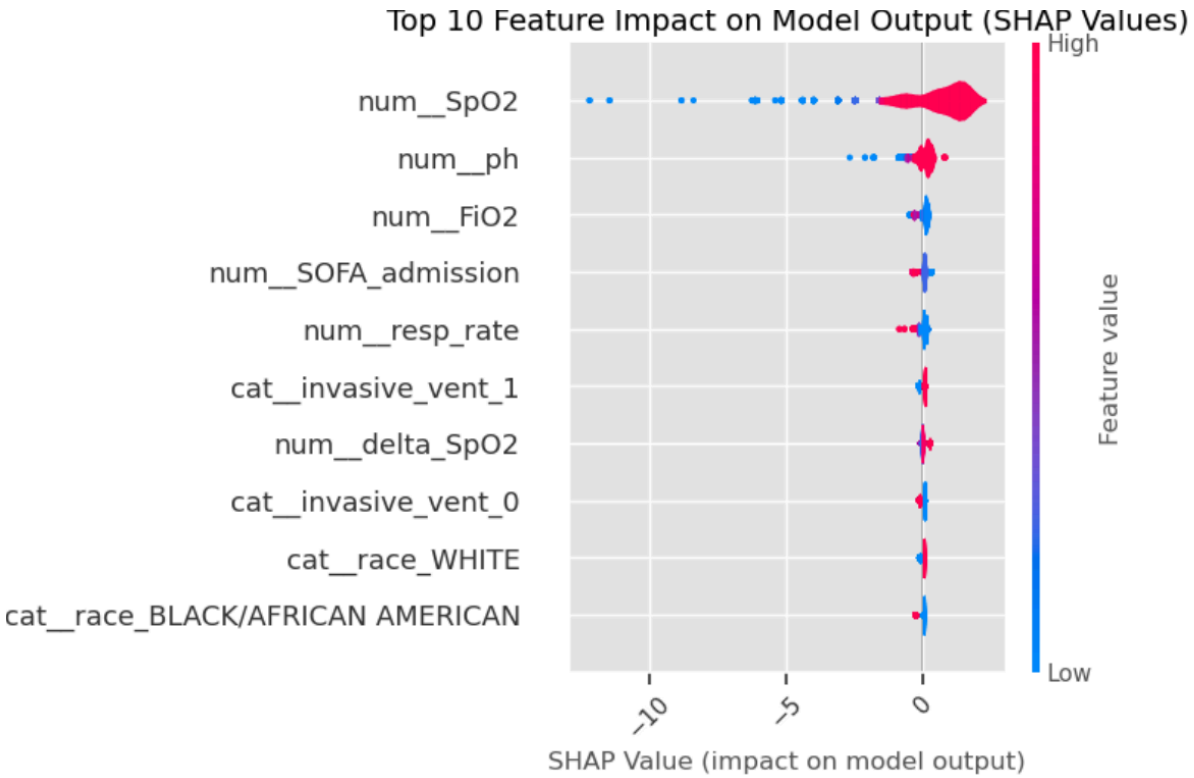



Figure 7: Top features indicating the importance of each feature and how the distribution of each feature contributes positively or negatively to model output.

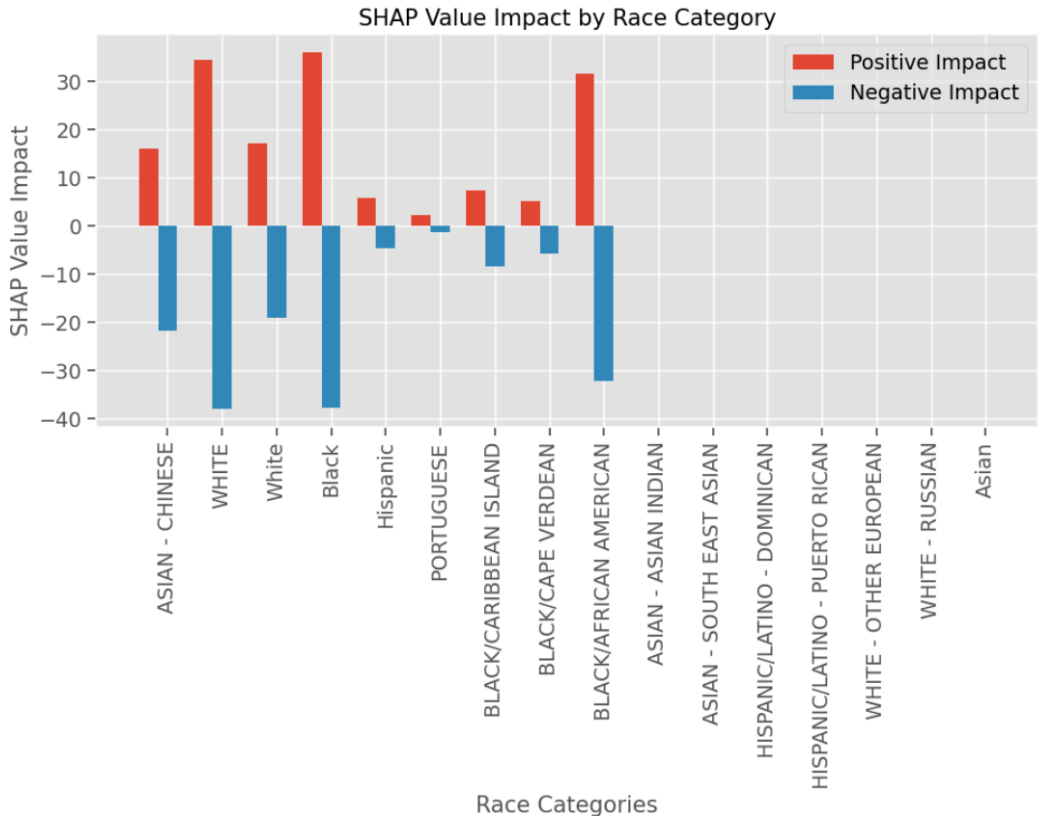



Figure 8: Top features indicating the importance of each race and how the distribution of each race contributes positively or negatively to model output.

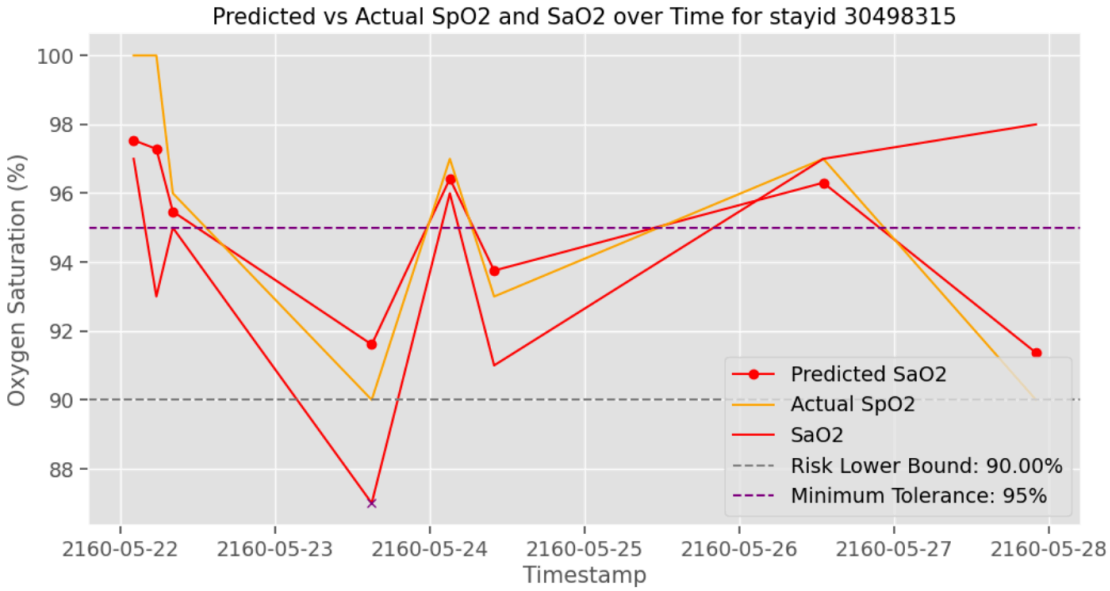



Figure 9: Along with an example QTL in purple indicating 90% as SaO2 minimum, the gray line shows a more accurate minimum estimation, derived from predicted SaO2. The cross indicates a hidden hypoxemia event. The timestamps are based on anonymized date-times but with preserved ranges from a patient's stay.

## Discussion

The model's performance, cross-validated in a test set, has a Squared Error (MSE) of 5.28, a Mean Absolute Error (MAE) of 1.51, and an R-squared (R2) value of 0.42. The model's estimation of SaO2 improves estimations from SpO2 readings alone as shown in Figure 3. They show that the model provides SaO2 estimations closer to its true values compared to SpO2 on its own. SpO2 also shows higher variance than the model's prediction, especially at lower levels of oxygen saturation. While some predictions are very close to the line of perfect correlation, others show more significant deviations. As shown in Figure 4, the mean difference between the SaO2 and predicted SaO2 vs SpO2 is reduced by a half, which could be significant for patients who might be at risk of conditions like hidden hypoxemia. The residuals in Figure 5 show that the predictions centered closer to zero and had less bias than SpO2 readings.

In Figure 2, the known issue of bias with pulse oximeters is addressed by observing R2 for the dark-skinned patients, indicating that the model explains 44% of the variability for this subgroup, showing that the model's predictions appear not to reproduce the bias. Although those with dark skin pigmentation are in the minority (as seen in Figure 1), the model still appears to predict well for them, due to the CatBoost Regressor being able to learn the race feature distribution without overfitting. This is also shown in Figure 6 where MAE was not particularly high for darker-skinned patients compared to those without darker skin.

Figure 9 demonstrates the model's ability to accurately estimate SaO2 fluctuations in a test set, despite the model being trained on patient stay's with varied time windows or intervals. The depicted QTL, highlighted in purple, represents a threshold of 95% SaO2, which is considered a minimum acceptable level before indicating potential hypoxemia. The model's predicted SaO2 values provide a refined minimum estimation, surpassing the accuracy of traditional SpO2 readings. The cross symbol on the graph pinpoints an incident of hidden hypoxemia, which is a critical event where the patient's oxygen saturation levels drop below a safe threshold but may not be immediately apparent through standard monitoring, for example, if QTL is set at 95%. The prediction reveals a calculated lower bound of the SaO2 levels. This bound is set initially by the differences between predicted SaO2 and SpO2. This could even be further calibrated by each time step depending on how well it over or underestimated SaO2, meaning the model can give more precise bounds over time.

The model was able to learn from features with well known clinical significance. In Figure 7, the top ten features are shown, demonstrating that higher (or increasing) values of SpO2 influences the model towards predicting higher SaO2 levels. Ph values, also having a linear positive relationship with oxygen saturation, came in as the second most impactful feature. Next, the fraction of inspired oxygen (FiO2), often collected along with SpO2. The SOFA score indicates organ function and failure which may affect certain patients oxygen saturation estimates. The presence of invasive ventilation features suggests the model differentiates between patients who are ventilated and those who are not. Respiratory rate and changes in SpO2 are also significant, highlighting the model's sensitivity to dynamic respiratory parameters. Finally, the inclusion of race categories implies that demographic factors may influence the model's predictions. Since the model is attempting to address measurement bias from SpO2, it makes sense it is leveraging demographic information to calibrate its prediction, with specific races having an impact on the model's predictions as shown in Figure 8.

## Conclusion

ML models present a refined approach for evaluating whether QTLs have been met or exceeded, thus improving the data quality of oxygen saturation measures. Utilizing the CatBoostRegressor on a robust dataset demonstrates superior SaO2 prediction, including for darker-skinned individuals, over traditional SpO2 measurements. This advancement promises improved oximetry accuracy, informing QTLs that more effectively counteract bias. Clinicians stand to gain a tool for more precise interventions, elevating patient care and mitigating risks associated with flawed oximetry, particularly vital in managing ARDS, TBI, and preventing critical hypoxic episodes.

## Future Work

Beyond just oxygen saturation, other kinds of measurement that risk capturing unwanted biases across vital monitoring, image diagnostics and skin-based treatments benefit from a data-driven approach. In oncology, breast cancer mortality is around 40% more for Black women compared to white (Yedjou et al., 2019), based on a sample from a US population. Melanoma also has high mortality rates despite low representation in medical datasets for Black populations, especially women (Wen et al., 2022). Such conditions may benefit from an approach comparable to the solution for biased SpO2 readings outlined in this dataset, when it comes to measurements that could impact their treatment or standard of care.

This could involve using QTLs that are derived from or compared against calibrated measurements from ML models trained on historical, benchmark, and even real world data. Further research should focus on validating the predictive model against gold-standard measures of skin pigmentation to ensure its accuracy across diverse patient groups, rather than a race label. Stratifying patients by specific treatments and interventions will allow for a more nuanced understanding of the model's performance in various clinical contexts. Collecting a more diverse dataset would enable a better model, including predictive features that encompass genetic, environmental, and behavioral factors. Finally, by incorporating time-series analysis, the model can move towards real-time monitoring, adapting to the continuous nature of physiological data.

## Reference

Cabanas, A.M., Fuentes-Guajardo, M., Latorre, K., León, D. and Martín-Escudero, P., 2022. Skin Pigmentation Influence on Pulse Oximetry Accuracy: A Systematic Review and Bibliometric Analysis. Sensors (Basel), 22(9), p.3402. Available at: [URL] (Accessed: [01/06/10])

Gottlieb, E. R., Ziegler, J., Morley, K., Rush, B., & Celi, L. A. (2022). Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit. JAMA Internal Medicine, 182(8), 849-858. doi:10.1001/jamainternmed.2022.2587 (Accessed: [01/01/2024])

Jacob C Jentzer, Anthony H Kashou, Francisco Lopez-Jimenez, Zachi I Attia, Suraj Kapa, Paul A Friedman, Peter A Noseworthy, Mortality risk stratification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients, European Heart Journal. Acute Cardiovascular Care, Volume 10, Issue 5, May 2021, Pages 532–541, https://doi.org/10.1093/ehjacc/zuaa021 (Accessed: [01/01/10])

Matos, J., Struja, T., Gallifant, J., Nakayama, L., Charpignon, M.-L., Liu, X., Economou-Zavlanos, N., Cardoso, J.S., Johnson, K.S., Bhavsar, N., Gichoya, J.W., Celi, L.A., & Wong, A.I. (2023) 'BOLD: Blood-gas and Oximetry Linked Dataset – Open Source Research', medRxiv, preprint: https://doi.org/10.1101/2023.10.03.23296485 (Accessed: [01/01/10])

Matos, J., Struja, T., Restrepo, D.S., Nakayama, L.F., Gallifant, J., Weishaupt, L., Mullangi, N., Loureiro, M., Shapiro, S., Carrel, A., & Celi, L.A. (2023) 'MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models', PhysioNet. Available at: https://doi.org/10.13026/jfpc-pz79 (Accessed: [01/01/2024)

Ren, S., Zupetic, J.A., Tabary, M., DeSensi, R., Nouraie, M., Lu, X., Boyce, R.D. & Lee, J.S. (2022). 'Machine learning based algorithms to impute PaO2 from SpO2 values and development of an online calculator'. Scientific Reports, [12], Article number: 8235. Available at: https://doi.org/10.1038/s41598-022-12419-7 (Accessed: [01/01/2024])

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. The New England journal of medicine, 383(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240 (Accessed: [01/01/2024])

Wen, D., Khan, S.M., Xu, A.J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X. and Matin, R.N., 2022. Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet Digital Health, [e-journal] 4(1), pp.e64-e74. Available at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext (Accessed: [01/01/2024])

Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A., Payton, M., &

Tchounwou, P. B. (2019). Health and Racial Disparity in Breast Cancer. Advances in experimental medicine and biology, 1152, 31–49. https://doi.org/10.1007/978-3-030-20301-6_3 (Accessed: [01/01/2024])

Yin, H., Yang, R., Xin, Y., Jiang, T. and Zhong, D. (2022) 'In-hospital mortality and SpO2 in critical care patients with cerebral injury: data from the MIMIC-IV Database', BMC Anesthesiology, 22(1), p. 386. Available at: https://doi.org/10.1186/s12871-022-01933-w (Accessed: [01/01/2024])

**Contact Information**

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Finn Janson
Company: Roche
Address: Hexagon Place, Shire Park, Falcon Way, Welwyn Garden City AL7 1TW
Work Phone: +44 (0)7513045612
Email: finn.janson@roche.com
Website: https://finnjanson.substack.com/