Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contactless Human Oxygen Saturation Detection Using RGB Camera in an Induced Hypoxemia Study with Varied Skin Types

Li Zhu

zhulivictor@gmail.com

Samsung Research America

Migyeong Gwak

Samsung Research America

Korosh Vatanparvar

Samsung Research America

Qijia Shao

Samsung Research America

Sharanya Desai

Samsung Research America

Jungmok Bae

Samsung Research America

Jilong Kuang

Samsung Research America

Alex Gao

Samsung Research America

Article

Keywords:

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9053537/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The authors declare no competing non-financial interests. Authors affiliated with Samsung Research America received salary. Data collection and

research were funded by Samsung Research America (Mountain View, CA 94043, USA).

Contactless Human Oxygen Saturation Detection Using RGB Camera in an Induced Hypoxemia Study with Varied Skin Types

Li Zhu[1*], Migyeong Gwak[1*], Korosh Vatanparvar[1], Qijia Shao[2], Sharanya Desai[1], Jungmok Bae[1], Jilong Kuang[1], Alex Gao[1]

[1*]Digital Health Team, Samsung Research America, 665 Clyde Ave, Mountain View, 94043, CA, USA.

[2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Work performed during Q. Shao's internship at Samsung Research America.

*Corresponding author(s). E-mail(s): li.zhu1@samsung.com; m.gwak@samsung.com;
Contributing authors: korosh.v@samsung.com; qijiashao@ust.hk; s.desai1@samsung.com; jungmok.bae@samsung.com; jilong.kuang@samsung.com; alex.gao@samsung.com;

Abstract

Contactless monitoring of blood oxygen saturation (SpO_2) using consumer-grade cameras could enable continuous monitoring in clinical settings where contact sensors are impractical, yet existing methods require optical filters, skin contact, or are restricted to narrow saturation ranges.

We developed a physics-based framework that extracts ratio-of-ratios (RoR) features from facial video and predicts SpO_2 via a RF regression model, evaluated through leave-one-subject-out cross-validation (LOSOCV) on 19 participants (Fitzpatrick types II–VI) during controlled induced hypoxemia (SpO_2: 70–100%) with an unmodified commercial RGB camera at approximately 0.6-meter (2-foot) distance under ambient lighting.

LOSOCV yielded A_{RMS} of 6.62% (95% CI: 5.60–7.93%; PCC = 0.595), reducible to 3.87% through per-subject calibration. Slope-constrained one-point calibration using a single concurrent pulse oximeter reading achieved A_{RMS} of 4.20%. Within clinically defined SpO_2 bins, A_{RMS} ranged from 2.58% to 2.73%, with

error concentrating at range boundaries rather than within clinically actionable bands. For hypoxemia screening at the 90% threshold, the area under the receiver operating characteristic curve (AUC-ROC) was 0.797 (95% CI: 0.669–0.870) with sensitivity 0.636 and specificity 0.846. Performance for Fitzpatrick type VI was substantially degraded ($A_{RMS} = 9.71\%$), consistent with melanin-induced signal attenuation in the visible spectrum.
These results demonstrate the feasibility of contactless SpO_2 estimation from unmodified RGB cameras and identify skin pigmentation as the primary barrier to equitable performance.

Keywords: remote photoplethysmography, pulse oximetry, SpO2, contactless monitoring, skin type equity, machine learning

Introduction

Blood oxygen saturation is crucial in the assessment and treatment of cardiopulmonary diseases, including asthma [1], chronic obstructive pulmonary disease [2], coronavirus disease (COVID-19) [3], and sleep apnea [4]. Inadequate oxygenation is a strong indicator of respiratory or cardiorespiratory compromise and can cause severe adverse effects in the brain, heart, and kidneys [5]. Peripheral oxygen saturation (SpO_2), measured with a U.S. Food and Drug Administration (FDA)-cleared pulse oximeter, is the standard clinical proxy for arterial oxygen saturation (SaO_2). However, contact-based pulse oximeters are impractical in settings where patients have fragile skin (e.g., neonates), highly infectious diseases, or when unobtrusiveness is essential for daily-life monitoring [6–9]. These limitations motivate the development of contactless methods for SpO_2 estimation.

Remote photoplethysmography (rPPG) uses video cameras to measure subtle changes in skin color caused by blood volume pulses (BVP). Verkruysse et al. [10] first demonstrated that cardiac pulse signals could be recovered from facial video under ambient light, and subsequent work by Poh et al. [11, 12] established blind source separation as a practical framework for contactless vital sign monitoring. Physics-based algorithms—including CHROM [6] and the Plane-Orthogonal-to-Skin (POS) method [13]—further improved robustness to motion and illumination variation. This body of work has enabled extraction of heart rate, respiration rate, and oxygen saturation from facial video [14–24]. Several studies have demonstrated camera-based SpO_2 estimation during breath-holding protocols, achieving desaturation to approximately 85% [21, 25, 26]. However, the FDA recommends that pulse oximeter performance be validated across the 70–100% SpO_2 range [27]. Among studies that have achieved this wider validation range using controlled inspired oxygen protocols, Hoffman et al. [28] demonstrated smartphone camera oximetry using a contact-based approach in which the finger was placed directly on the camera lens, and Van Gastel and Verkruysse [29] demonstrated contactless RGB camera SpO_2 estimation using optical bandpass filters and polarizers to enhance signal purity. Since these studies, the field has advanced rapidly: deep learning approaches have achieved low error rates under controlled conditions [30], neonatal applications have demonstrated clinical feasibility in intensive care

settings [31], and comprehensive reviews have highlighted both progress and remaining challenges in translating camera-based SpO_2 to clinical practice [16, 32]. Nevertheless, existing wide-range validation studies either require hardware modifications (optical filters, polarizers) or physical contact with the sensing device, and no study has assessed truly contactless camera-based SpO_2 estimation across diverse Fitzpatrick skin types using only unmodified commercial hardware and ambient light [33].

The camera-based technique operates on a similar principle to conventional pulse oximetry, with significant additional challenges. Traditional pulse oximeters emit narrowband light at two wavelengths—typically red and near-infrared—through tissue, exploiting the differential absorption of oxygenated and deoxygenated hemoglobin. Based on the Beer–Lambert law [34], a ratio of ratios (RoR) can be derived that is highly correlated with SpO_2 and can be used for prediction via a calibrated model [35]. Camera-based approaches face additional challenges: ambient illumination has a broad spectrum rather than controlled narrowband wavelengths, light reflects from inhomogeneous tissue rather than transmitting through a thin measurement site, and involuntary subject movement introduces temporal instability.

The importance of equitable performance across skin pigmentations has been underscored by recent regulatory developments. In January 2025, the FDA issued draft guidance recommending that pulse oximeter manufacturers evaluate device accuracy across diverse skin tones, increasing the recommended number of study participants from 10 to 150 and introducing objective skin tone assessment methods [36]. This guidance reflects growing evidence that traditional pulse oximeters exhibit systematic bias for patients with darker skin tones, leading to delayed identification of hypoxemia [2]. Contactless camera-based approaches face analogous challenges, as higher melanin concentrations attenuate the optical signals used for SpO_2 estimation [8, 37]. These findings motivate the inclusion of diverse skin types and transparent stratified reporting in camera-based SpO_2 validation studies.

In this study, we developed a physics-based computational framework for contactless camera-based SpO_2 estimation, designed to be fully implementable on commercial electronic devices such as smartphones and camera-embedded monitors. The framework extracts ratio-of-ratios (RoR) features from facial video using an interpretable signal processing pipeline grounded in the Beer–Lambert model, and predicts SpO_2 using a RF regression model selected by lowest leave-one-subject-out cross-validation (LOSOCV) error. We evaluated this system on 19 participants spanning Fitzpatrick skin types II–VI during a controlled induced-hypoxemia protocol covering the full 70–100% SpO_2 range, using an unmodified commercial CMOS RGB camera at approximately 0.6-meter (2-foot) distance under ambient lighting. LOSOCV yielded an A_{RMS} of 6.62% (95% CI: 5.60–7.93%; MAE = 5.34%) across all participants, with A_{RMS} of 5.80% for Fitzpatrick types II–V. For hypoxemia screening at the 90% threshold, the system achieved an AUC-ROC of 0.797 with sensitivity 0.636 and specificity 0.846. Performance for Fitzpatrick type VI participants ($n = 3$) was substantially degraded ($A_{RMS} = 9.71\%$), consistent with melanin-induced optical signal attenuation—a finding we analyze in detail and frame as both a current limitation and a motivation for future melanin-robust algorithm development. To the best of our knowledge, this study provides the first feasibility assessment of contactless camera-based SpO_2 estimation

on the wide 70–100% range across Fitzpatrick skin types II–VI using only commercial hardware and ambient light, without optical filters or hardware modifications. We present this work as an exploratory feasibility demonstration establishing whether unmodified RGB cameras can extract clinically relevant SpO_2 information across the full desaturation range in a diverse cohort.

Results

SpO_2 prediction performance

The RoR features extracted by the signal processing pipeline were mapped to SpO_2 using leave-one-subject-out cross-validation (LOSOCV) across all 19 participants ($N = 2{,}413$ paired measurements). The best-performing regression model (RF, selected by lowest A_{RMS}) achieved an A_{RMS} of 6.62% (95% CI: 5.60–7.93%), MAE of 5.34% (95% CI: 4.51–6.42%), and Pearson correlation coefficient (PCC) of 0.595 (95% CI: 0.373–0.732). For Fitzpatrick types II–V ($n = 16$), evaluated using a dedicated 16-subject LOSOCV (type VI excluded), the primary outcome was $A_{RMS} = 5.80\%$ across the full 70–100% SpO_2 range. All four candidate models produced comparable results, with the RF outperforming the next-best linear model by 15.6% in A_{RMS}, indicating that the choice of regression model has a modest effect relative to the upstream RoR feature extraction (Table 1).

Agreement analysis

Bland-Altman analysis (Figure 1) revealed a mean bias of +0.15% SpO_2, indicating negligible systematic over- or under-estimation. The 95% limits of agreement (LOA) were −12.83% to +13.13%. The Breusch–Pagan test for heteroscedasticity [38] was not significant ($p = 0.600$), confirming that standard (constant-width) LOA are appropriate. Per-subject Bland-Altman plots are provided in Supplementary Figure S2.

Hypoxemia classification

To assess clinical screening utility, hypoxemia detection was evaluated as a binary classification task at SpO_2 thresholds of 90% and 88%. At the 90% threshold, the SpO_2-threshold-based classifier achieved an AUC-ROC of 0.797 (95% CI: 0.669–0.870), with sensitivity 0.636, specificity 0.846, and accuracy 71.6% at the optimal operating point determined by Youden's J statistic (Figure 2). At the 88% threshold, performance was comparable: AUC-ROC = 0.800, sensitivity = 0.694, specificity = 0.810, and accuracy = 74.8%.

Effect of skin type

Performance was stratified by Fitzpatrick skin type (Figure 1b, Figure 3). To ensure fair subgroup evaluation, types II–V performance was assessed using a dedicated 16-subject LOSOCV in which type VI participants were excluded from training folds

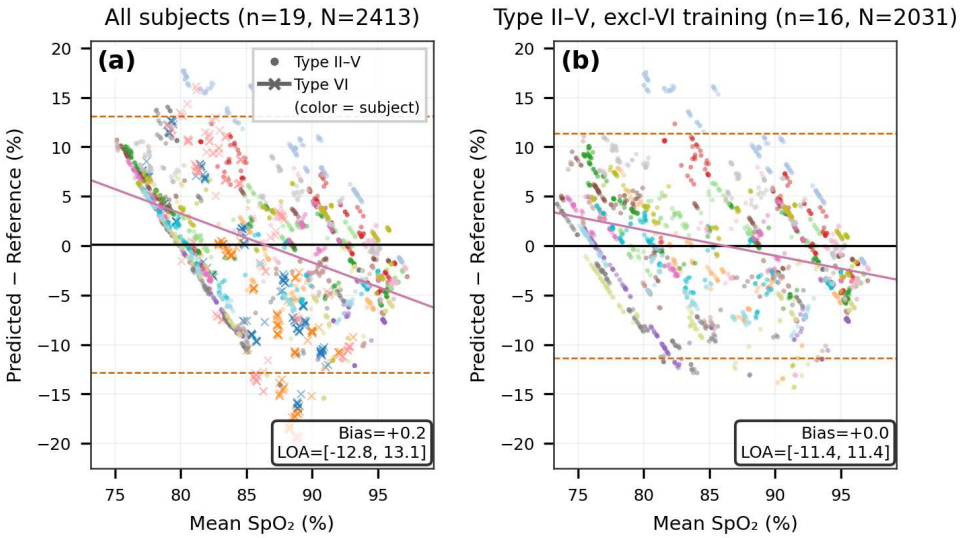


Fig. 1: Bland-Altman agreement analysis. (a) Aggregate agreement between predicted and reference SpO_2 for all 19 participants (N = 2,413) using full-cohort LOSOCV. The solid line indicates the mean bias (+0.15%), and dashed lines indicate the 95% limits of agreement (−12.83% to +13.13%). (b) Agreement for Fitzpatrick II–V participants (n = 16) using a model trained exclusively on types II–V via 16-subject LOSOCV (type VI excluded from training folds), with bias +0.01% and LOA −11.36% to +11.39%.

(type VI results are reported separately from the full 19-subject LOSOCV). For participants with skin types II–V (n = 16, N = 2,031 samples), A_{RMS} was 5.80% with LOA of −11.36% to +11.39%. For participants with Fitzpatrick skin type VI (n = 3, N = 382 samples), A_{RMS} from the full 19-subject LOSOCV increased to 9.71% with substantially wider LOA of −21.13% to +15.06%.

Per-subject analysis (Figure 3) confirmed consistent performance across skin types II–V, with the performance gap concentrated in type VI. This disparity is attributable to melanin absorption, which attenuates the optical signal available for SpO_2 estimation in the visible spectrum. Within clinically defined SpO_2 bins, however, type VI performance converges toward types II–V, suggesting that the wider full-range error reflects difficulty at range boundaries rather than a fundamental inability to track SpO_2 changes.

To test whether including type VI participants in the training set affects performance for other groups, we conducted a leave-out analysis (Figure 4). The primary II–V metric (5.80%) already uses the dedicated 16-subject LOSOCV (type VI excluded). Comparing this against the II–V slice of the full 19-subject LOSOCV (where type VI subjects were included in training) yielded similar results, confirming no cross-group contamination. However, excluding type VI from training dramatically worsened type VI prediction (A_{RMS}: 12.52% without versus 9.71% with type VI in training). This confirms that inclusive training benefits the underrepresented group at no cost to the majority, supporting the practice of training on diverse cohorts.

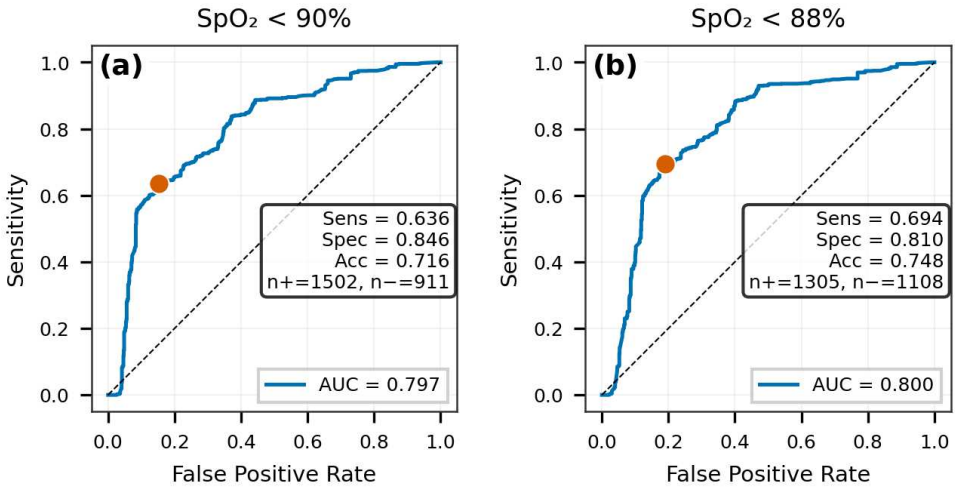


Fig. 2: ROC curves for hypoxemia classification. SpO_2-threshold-based classification at 90% threshold (AUC-ROC = 0.797, sensitivity = 0.636, specificity = 0.846, accuracy = 71.6%) and 88% threshold (AUC-ROC = 0.800, sensitivity = 0.694, specificity = 0.810, accuracy = 74.8%).

Per-subject calibration potential

Although the illumination condition and camera model were consistent for all subjects, the recorded RoR values spread differently across individuals, with significant intercept shifts in the RoR–SpO_2 correlation patterns. To quantify the potential benefit of individualized correction, we applied a per-subject ordinary least-squares (OLS) calibration to the LOSOCV predictions: for each subject k, we fitted $\hat{y}_k^{\text{cal}} = \alpha_k \hat{y}_k + \beta_k$, where α_k and β_k minimize $\sum_i (y_{k,i} - \alpha_k \hat{y}_{k,i} - \beta_k)^2$ over all n_k samples from that subject, and $y_{k,i}$ denotes the reference SpO_2. Because this fit uses each subject's own reference values for both fitting and evaluation, the resulting accuracy ($A_{\text{RMS}} = 3.87\%$) represents an *upper bound* on what individualized linear calibration could achieve in practice. Compared with the uncalibrated baseline ($A_{\text{RMS}} = 5.87\%$), per-subject calibration reduced A_{RMS} by 34% ($p = 3.05 \times 10^{-5}$, Wilcoxon signed-rank test on per-subject A_{RMS}; Figure 5), demonstrating substantial latent accuracy recoverable through individualized correction.

As a practical intermediate between no calibration and full per-subject fitting, slope-constrained one-point calibration (see Methods) reduced A_{RMS} from 5.87% to 4.20%—a 28% improvement ($p = 0.004$, Wilcoxon signed-rank test; Figure 5). By fixing the slope to the leave-one-out population mean and estimating only the per-subject intercept from a single concurrent pulse oximeter reading, this approach captures subject-specific offset information sufficient to substantially improve prediction accuracy, outperforming both uncalibrated predictions and leave-one-subject-out cross-validation (LOSOCV $A_{\text{RMS}} = 6.62\%$). Notably, the calibration point was not optimized: for each subject, the chronologically first sample in the recording was used, avoiding any selection bias that could inflate reported accuracy. Both calibration comparisons were tested against the same uncalibrated baseline; applying Bonferroni

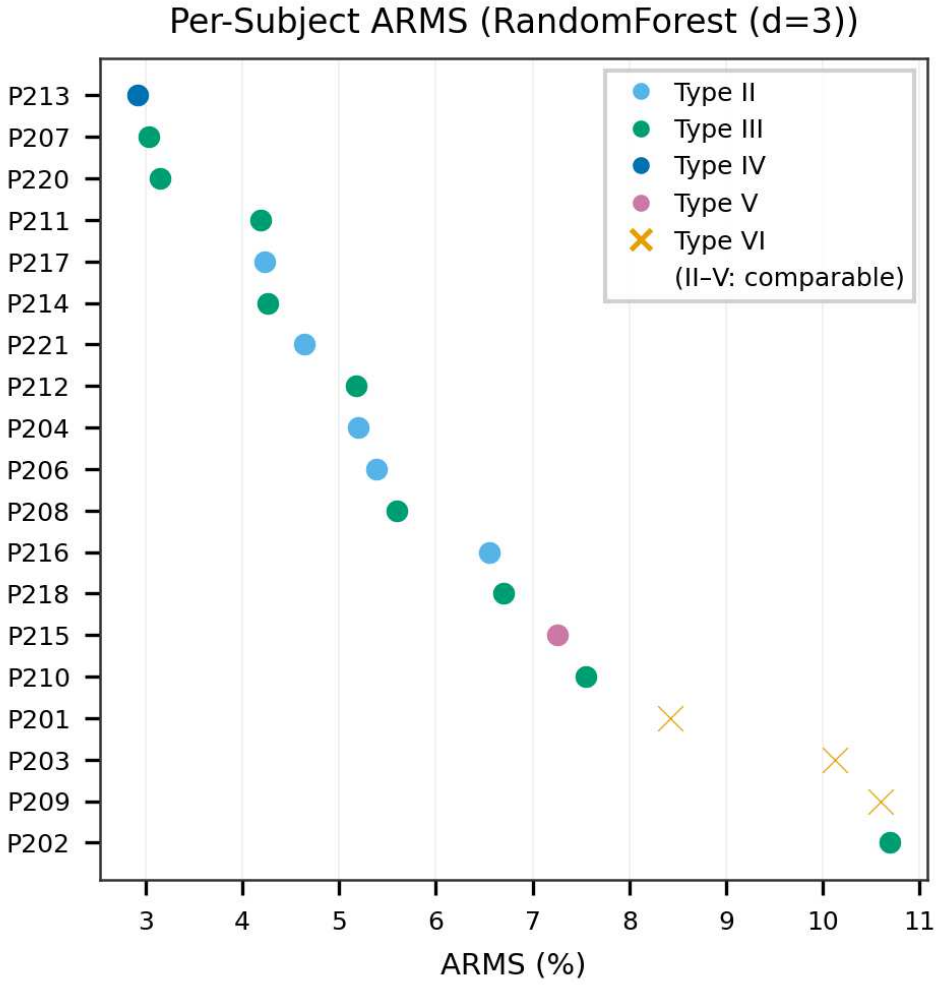


Fig. 3: Per-subject A_{RMS} stratified by Fitzpatrick skin type. Types II–V values are from the dedicated 16-subject LOSOCV (type VI excluded); type VI values are from the full 19-subject LOSOCV. Participants with types II–V showed consistent performance (A_{RMS} = 5.80%), while type VI participants exhibited substantially higher error (A_{RMS} = 9.71%), consistent with melanin-induced signal attenuation.

correction for two comparisons yields adjusted thresholds of $\alpha = 0.025$, which both p-values comfortably survive.

Some participants exhibited systematically shifted RoR intercepts, resulting in elevated uncalibrated error despite strong underlying RoR–SpO_2 correlation (e.g., P202, P215). These cases exemplify the inter-subject baseline variability that calibration corrects; a detailed outlier analysis is provided in Section 2.

Figure 6 visualizes the effect of slope-constrained one-point calibration on per-subject RoR_{RG}–SpO_2 regression curves. Before calibration, near-parallel slopes with

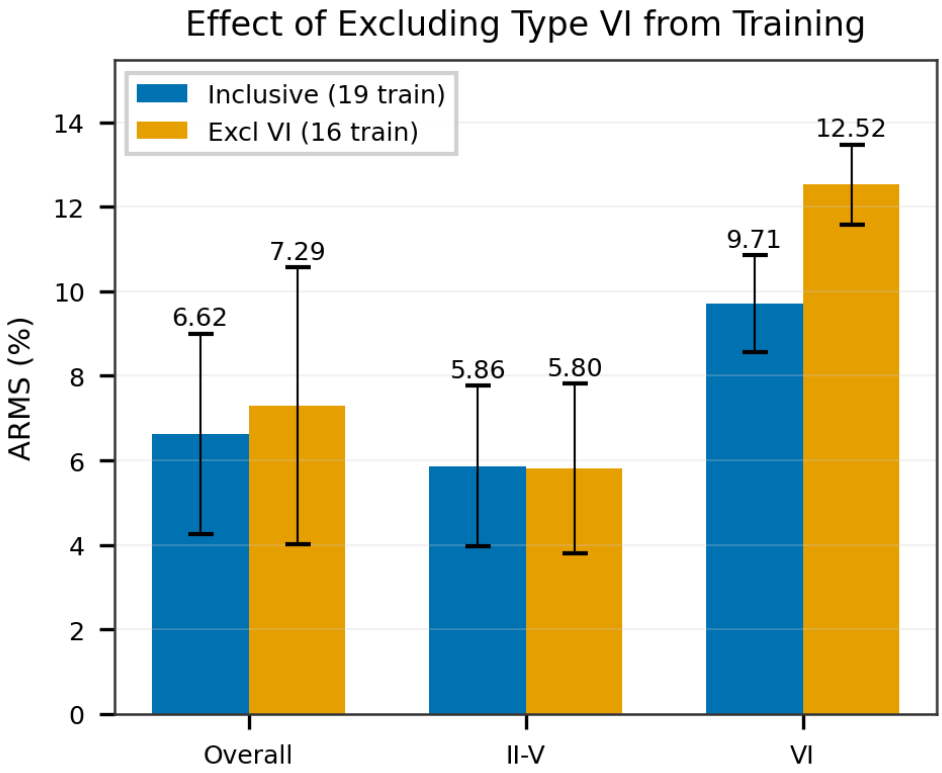


Fig. 4: Type VI exclusion analysis. Comparison of LOSOCV performance when Fitzpatrick type VI participants are included versus excluded from training folds. Inclusive training produces equal or better results, confirming that diverse training data benefits model generalization.

significant intercept shifts confirm that inter-subject variability is predominantly an offset phenomenon (panel a). After calibration, the intercept spread is substantially reduced under the population-mean slope (panel b), consistent with subject-specific optical path length and melanin content affecting the DC baseline without altering the AC modulation depth relationship.

Outlier analysis: P202 and P215

We selected two participants for detailed analysis based on distinct motivations: P202 exhibited the highest uncalibrated A_{RMS} in the cohort, warranting investigation into the source of error; P215, the sole Fitzpatrick type V participant, exhibited unexpectedly elevated error for a skin type that should fall within the system's operating range. Both cases illustrate that elevated prediction error can arise from addressable DC offset differences rather than fundamental algorithmic failure or skin type effects.

P202 (Fitzpatrick type III), the worst-performing subject, showed strong RoR–SpO_2 feature correlation (RoR_{RG} vs. reference PCC $= -0.978$, highest in cohort) but a

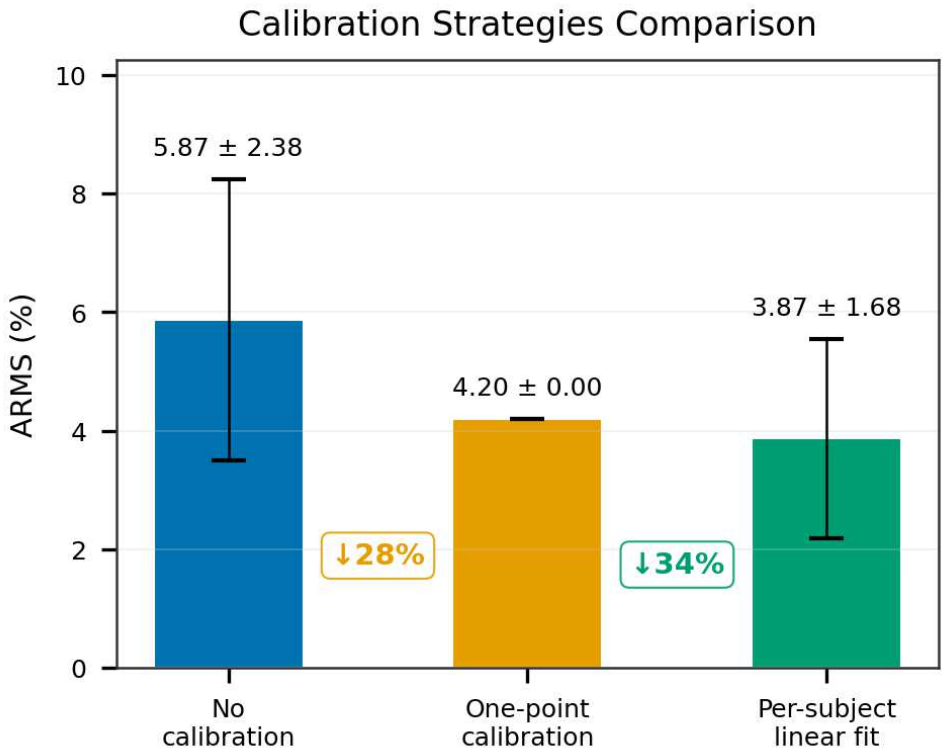


Fig. 5: **Calibration strategies comparison.** Three approaches: uncalibrated LOSOCV predictions (A_{RMS} = 5.87%), slope-constrained one-point calibration using a single concurrent measurement with population-derived slope prior (A_{RMS} = 4.20%, 28% improvement, $p = 0.004$), and per-subject linear fit upper bound (A_{RMS} = 3.87%, 34% improvement, $p = 3.05 \times 10^{-5}$). Error bars denote ±1 standard deviation of per-subject A_{RMS} values.

systematically shifted RoR intercept. Investigation revealed that P202's camera white balance configuration (Blue_BalanceRatio = 2.10) was 2.0σ below the cohort mean (mean ± SD: 3.41 ± 0.65, $n = 16$), suppressing the blue channel DC baseline and shifting the RoR intercept downward. This camera configuration parameter correlated significantly with per-subject A_{RMS} across the cohort (PCC = −0.567, $p = 0.028$).

P215 (Fitzpatrick type V) similarly exhibited elevated uncalibrated error (A_{RMS} = 5.48%) attributable to inter-subject optical path variation rather than skin pigmentation effects. Despite being the only type V participant, P215's RoR–SpO_2 feature correlation remained strong (RoR_{RG} vs. reference PCC = −0.962), and per-subject calibration reduced A_{RMS} to 4.47% (Figure 6). This confirms that elevated error in P215 reflects a DC offset shift—analogous to P202's camera-induced offset—rather than a melanin-driven signal degradation mechanism.

Both cases demonstrate that slope-constrained one-point calibration can correct subject-specific offset variability regardless of its origin (camera configuration or optical path differences).

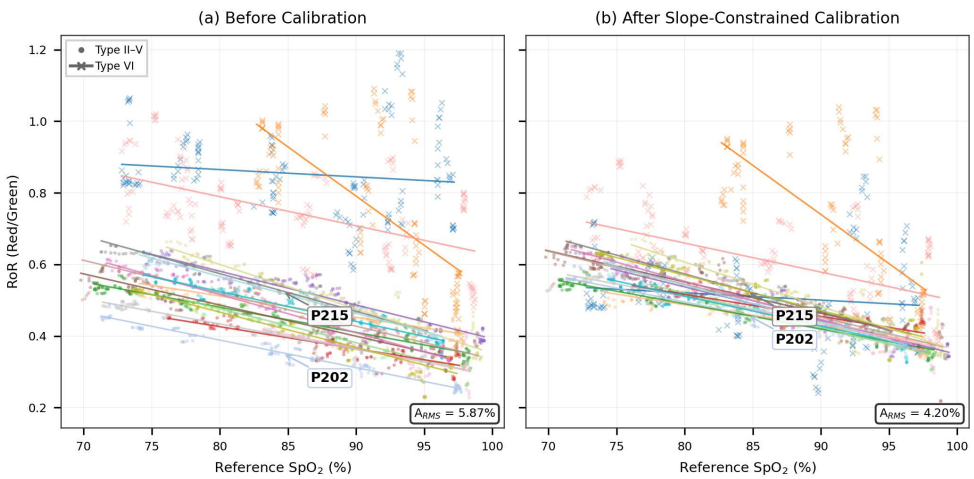


Fig. 6: Effect of slope-constrained one-point calibration on per-subject RoR–SpO_2 relationships. (a) Before calibration: individual RoR_{RG}–SpO_2 regression curves show near-parallel slopes (CV = 20.4%) with significant intercept spread across 19 participants. (b) After calibration: per-subject intercepts estimated from a single concurrent measurement converge under the population-mean slope, substantially reducing offset variability while preserving individual RoR dynamics. P202 and P215 are annotated to illustrate that subjects with high uncalibrated error exhibit strong underlying correlation, with their elevated A_{RMS} attributable to addressable DC offset differences.

Range-specific model performance

To examine how prediction accuracy varies across clinically defined SpO_2 ranges, we conducted bin-restricted LOSOCV for five ISO 80601-2-61 bins: <80%, <90%, 80–90%, ≥80%, and ≥90%. A separate model was trained and tested exclusively within each range. Bin-restricted LOSOCV yielded A_{RMS} of 2.58% (<80%, $n = 663$), 5.34% (<90%, $n = 1{,}502$), 2.77% (80–90%, $n = 839$), 4.99% (≥80%, $n = 1{,}750$), and 2.73% (≥90%, $n = 911$; Figure 7). By comparison, the full-range model evaluated on the same bins produced A_{RMS} of 7.96% (<80%), 6.46% (<90%), 4.96% (80–90%), 6.04% (≥80%), and 6.88% (≥90%), respectively. These results demonstrate that within any given clinical bin, the RoR–SpO_2 relationship is well captured, and that the full-range model's aggregate error is dominated by cross-bin prediction uncertainty.

Feature ablation

To assess the contribution of each RoR feature, we compared three feature configurations across the same RF model: RoR_{RB} only (A_{RMS} = 6.80%), RoR_{RG} only (A_{RMS} = 6.57%), and both features combined (A_{RMS} = 6.62%). The overlapping confidence intervals indicate that both channels contribute complementary information, with RoR_{RG} providing marginally better standalone performance. The combined model was selected as the primary result because the two-feature input is motivated by the physics-based Beer–Lambert derivation and provides robustness against single-channel artifacts (Supplementary Figure S3).

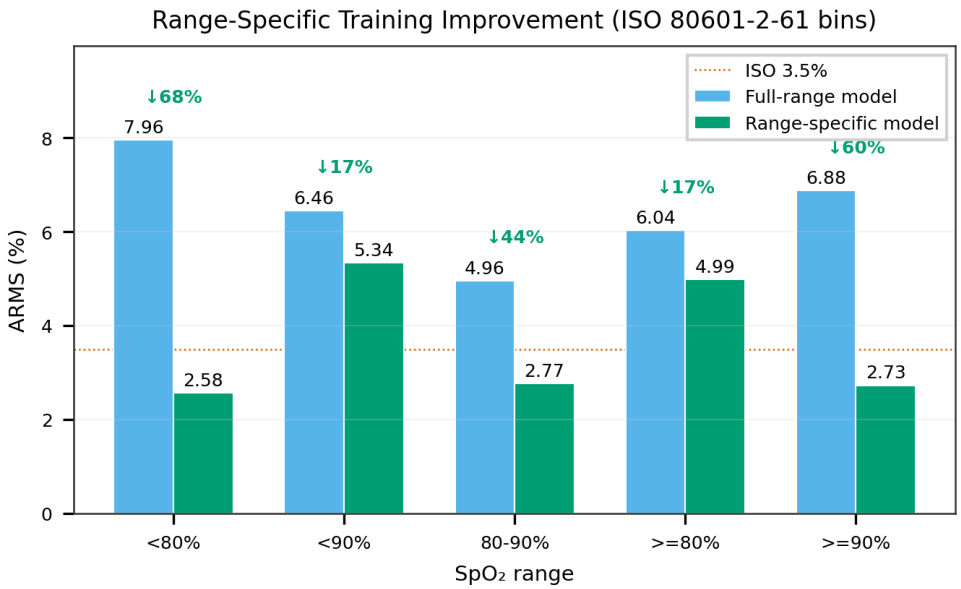


Fig. 7: Range-specific model performance across five ISO 80601-2-61 bins. Bin-restricted LOSOCV (dark bars) achieved A_{RMS} of 2.58% (<80%, n = 663), 5.34% (<90%, n = 1,502), 2.77% (80–90%, n = 839), 4.99% (≥80%, n = 1,750), and 2.73% (≥90%, n = 911), versus the full-range model evaluated per bin (light bars).

Model selection across SpO_2 ranges

The optimal model family varied with the SpO_2 range. The RF model achieved the best full-range performance (A_{RMS} = 6.62%), consistent with the nonlinear Beer–Lambert relationship between RoR and SpO_2 that spans the 70–100% range. Within narrow bins, however, linear regression models were competitive or superior, reflecting the locally linear RoR–SpO_2 relationship within clinically homogeneous ranges (Table 1). This finding has practical implications: deployment scenarios targeting specific clinical ranges (e.g., ICU monitoring within 80–100%) may benefit from simpler linear models, while general-purpose screening across the full desaturation range requires nonlinear modeling.

Table 1: Model comparison for SpO_2 prediction via LOSOCV. All models use the same two RoR features (RoR_{RB}, RoR_{RG}) with a priori hyperparameters (no tuning). Best values in bold.

Model	A_{RMS} (%)	MAE (%)	PCC
RandomForest (d=3)	**6.62%**	**5.34%**	**0.595**
Ridge ($\alpha = 1.0$)	7.95%	6.46%	0.332
ElasticNet ($\alpha = 0.1$)	7.84%	6.42%	0.340
OLS (baseline)	7.95%	6.46%	0.332

All models evaluated via leave-one-subject-out cross-validation on 19 participants (N = 2,413 samples). Model selection based on lowest A_{RMS}.

Discussion

This study demonstrates the feasibility of contactless SpO_2 estimation using an unmodified commercial RGB camera across the clinically relevant 70–100% range. LOSOCV on 19 participants (Fitzpatrick types II–VI) yielded A_{RMS} of 6.62% (Figure 1), reducible to 3.87% through per-subject calibration (Figure 5). Performance was consistent across Fitzpatrick types II–V ($A_{RMS} = 5.80\%$) but degraded for type VI ($A_{RMS} = 9.71\%$; Figure 3).

The most methodologically relevant comparator is Van Gastel and Verkruysse [29], who demonstrated contactless SpO_2 estimation using an RGB camera with controlled desaturation across diverse skin types. Their system employed optical bandpass filters and polarizers, validating on 29 participants (Fitzpatrick types I–VI) with MAE below 4% for seated scenarios. Several methodological differences preclude direct metric comparison: their hardware modifications substantially improve signal quality, their calibration-validation split differs from our LOSOCV scheme, and cohort composition differs. Our approach uses no hardware modifications, relying entirely on computational signal processing to extract SpO_2-informative features from ambient-light RGB video.

Hoffman et al. [28] demonstrated smartphone camera oximetry using a contact-based finger-on-lens approach, achieving AUC of 0.87 for hypoxemia screening in 6 participants. More recently, deep learning approaches have demonstrated low error rates under controlled conditions [30], and neonatal applications have shown clinical feasibility using notch-filtered RGB cameras [31]. These developments reflect growing interest in contactless SpO_2 monitoring across diverse clinical populations [32].

Our study contributes to this landscape through: (1) truly contactless measurement at 0.6-meter (2-foot) distance using an unmodified commercial camera under ambient lighting; (2) evaluation across the widest SpO_2 range (70–100%) among comparable contactless RGB studies; (3) a diverse cohort ($n = 19$, Fitzpatrick types II–VI) with transparent stratified reporting; and (4) a physics-based feature extraction pipeline whose intermediate outputs (AC, DC, RoR) are directly interpretable under the Beer–Lambert model.

The RF model was selected based on lowest LOSOCV A_{RMS} (6.62% versus 7.95% for Ridge regression; Table 1), outperforming the next-best linear model by 15.6%. This is consistent with the nonlinear RoR–SpO_2 relationship across the 70–100% range; within narrow clinical bins, linear models were competitive (Figure 7), suggesting that deployment scenarios targeting specific ranges may benefit from simpler models.

The physics-based RoR pipeline provides a key advantage: interpretability. The intermediate outputs—AC and DC components per channel and derived RoR values—are directly meaningful under the Beer–Lambert model, enabling diagnosis of specific failure modes. For instance, the melanin-induced attenuation in type VI participants is visible in the AC/DC ratios themselves, not hidden within learned network weights. We acknowledge that deep learning methods may achieve better accuracy with larger datasets [16], and that transformer-based [39] and GAN-based [40] architectures have shown promising results. For this exploratory study ($n = 19$), the physics-based approach was appropriate for establishing feasibility and providing mechanistic insight.

This interpretability also enables principled calibration. Although illumination and camera model were consistent across subjects, per-subject RoR–SpO_2 regression slopes were near-parallel with significant intercept shifts (Figure 6), indicating that inter-subject variability is predominantly an offset phenomenon attributable to skin characteristics and tissue composition. Slope-constrained one-point calibration using a population-derived slope prior and a single concurrent pulse oximeter reading reduced A_{RMS} from 5.87% to 4.20%—a 28% improvement ($p = 0.004$; Figure 5)—demonstrating that subject-specific DC offsets can be corrected with minimal additional instrumentation. The per-subject OLS upper bound ($A_{RMS} = 3.87\%$, 34% improvement) represents the ceiling with full reference data. Notably, participants identified as "outliers" in the uncalibrated analysis—such as P202 (camera configuration shift) and P215 (inter-subject optical path variation)—showed strong underlying RoR–SpO_2 correlation, with their elevated error attributable to addressable DC offset differences rather than fundamental algorithmic failure (Figure 6). The slope-constrained approach (CV = 20.4%) stabilizes prediction across the SpO_2 range by avoiding the noise amplification of single-subject slope estimation. This reframes calibration not as an optional post-processing step but as an integral component of a deployment-ready system. In a practical hospital ward scenario, a nurse takes a spot-check pulse oximeter reading during admission (providing the calibration point), and the camera then provides continuous contactless monitoring thereafter—reconciling the apparent paradox of requiring a reference device for a contactless system.

The sensitivity of 0.636 at the 90% SpO_2 threshold and limits of agreement of -12.83% to $+13.13\%$ require clinical contextualization. These metrics are insufficient for diagnostic confirmation, ICU-grade monitoring, or standalone clinical decision-making, where contact pulse oximeters achieving sensitivity above 0.95 and limits of agreement within $\pm 4\%$ remain the appropriate standard. However, for triage screening applications—where the clinical question is "is this patient's oxygenation deteriorating?" rather than "what is the precise SpO_2 value?"—the system's specificity of 0.846 ensures that positive screening results trigger confirmatory contact measurement rather than generating excessive false alarms. However, the corresponding sensitivity of 0.636 implies a 36% false-negative rate—meaning that over one-third of truly hypoxemic episodes would be missed—which precludes use as a standalone diagnostic screen without concurrent monitoring strategies. In continuous trending scenarios, the direction and rate of SpO_2 change carry clinical significance independent of per-reading absolute accuracy: a system that detects a downward trajectory from 95% toward 85% provides actionable early-warning information even with $\pm 13\%$ individual-reading uncertainty. In settings where the alternative is no SpO_2 monitoring at all—such as general hospital wards or post-surgical recovery areas—contactless screening at this performance level provides supplementary clinical value as part of a multi-modal monitoring approach, where positive detections are confirmed by a contact pulse oximeter.

ISO 80601-2-61:2017 was designed for clinical-grade contact pulse oximeters operating with controlled narrowband illumination and direct tissue transmission; no contactless camera-based system has achieved ISO compliance across diverse skin

tones. The overall A_{RMS} of 6.62% exceeds this standard's 3.5% threshold, but contextualizing this gap against comparable systems is instructive. Van Gastel and Verkruysse [29] reported MAE below 4% but required optical bandpass filters and polarizers to isolate hemoglobin-sensitive wavelengths, while Hoffman et al. [28] achieved RMSE of 5.55% using contact-based finger-on-lens sensing rather than truly contactless measurement. Our system operates under fundamentally more challenging optical conditions—ambient light only, no hardware modifications, 0.6-meter measurement distance—and the observed accuracy is consistent with the current state of contactless technology. Notably, bin-restricted LOSOCV yielded A_{RMS} as low as 2.58% within the <80% SpO_2 range, demonstrating that within clinically defined ranges, the underlying signal quality supports accuracy approaching ISO-level performance when the prediction range is constrained. These results position the system as a feasibility demonstration for contactless SpO_2 monitoring rather than a candidate for immediate regulatory clearance.

Performance for Fitzpatrick types II–V was consistent (A_{RMS} = 5.80%, n = 16) but substantially degraded for type VI (A_{RMS} = 9.71%, n = 3; Figure 3, Figure 1b). The 95% limits of agreement were wider for type VI (−21.13% to +15.06%) than for types II–V (−11.36% to +11.39%). This disparity is attributable to melanin absorption in the visible spectrum: melanin absorbs up to 20 times more than hemoglobin in the red band (Figure 8, Table 2), consistent with prior studies of skin-tone-dependent attenuation in both contact [2] and remote [37] photoplethysmography.

Three findings contextualize this disparity. First, within clinically defined SpO_2 bins, type VI performance converges toward types II–V, indicating that the full-range error arises from difficulty at range boundaries rather than a fundamental inability to track SpO_2 changes. Second, our Type VI exclusion analysis (Figure 4) showed that including type VI in training (19-subject LOSOCV) had negligible effect on types II–V compared with the dedicated 16-subject LOSOCV, but excluding type VI from training dramatically worsened type VI prediction (12.52% versus 9.71%), confirming that inclusive training benefits the underrepresented group at no cost to the majority. Third, per-subject calibration corrects a substantial portion of the type VI error attributable to RoR baseline shifts.

This disparity is not merely a limitation—it represents a fundamental challenge for equitable clinical deployment. The FDA's 2025 draft guidance [36] recommends evaluating pulse oximeter accuracy across diverse skin pigmentations, and our stratified reporting demonstrates why: without skin-type-specific evaluation, the aggregate A_{RMS} of 6.62% would mask degraded performance for darker skin tones, perpetuating disparities documented in contact-based oximetry [2, 8]. Addressing this will require advances at multiple levels: near-infrared channels where melanin absorption is reduced, enhanced illumination strategies (e.g., screen-based illumination for smartphone deployment), and melanin-aware signal processing. Ambient lighting conditions also affect the RoR–SpO_2 relationship, impacting both slope and intercept (Supplementary Figure S4 [41]), motivating lighting-aware calibration for real-world deployment. We emphasize that only three type VI participants were included; larger cohorts balanced across Fitzpatrick types V and VI are essential for developing and validating melanin-robust algorithms.

Table 2: **Measured melanin content by Fitzpatrick skin type.** Melanin index was measured using Mexameter MX-18 [42] from the forehead region. The concentration of melanin increases exponentially across Fitzpatrick skin types and peaks dramatically at Skin Type VI.

Fitzpatrick Skin Type	ID Subjects	Mean Melanin Content (a.u.)
I	0	N.A.
II	5	134
III	9	192
IV	1	251
V	1	354
VI	3	994

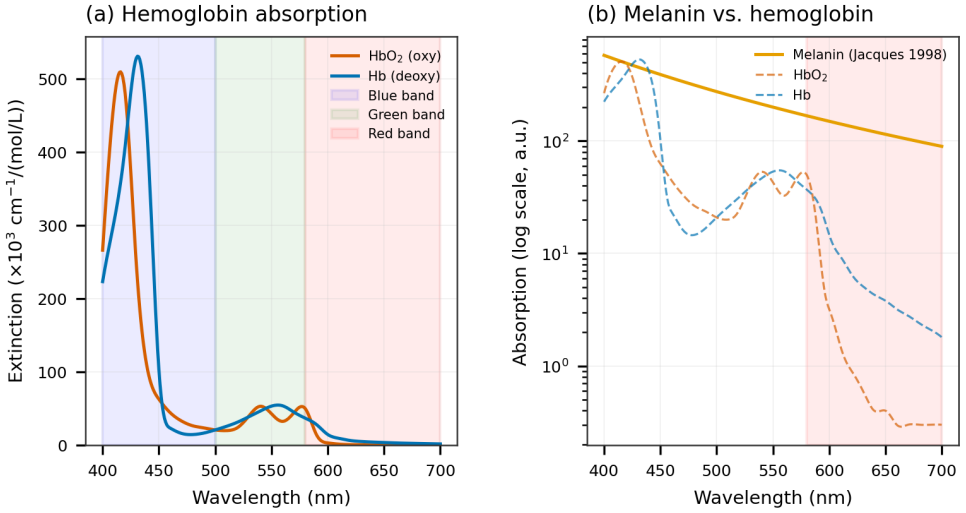


Fig. 8: **Optical absorption spectra relevant to camera-based SpO_2 estimation.** (a) Molar extinction coefficients of oxyhemoglobin (HbO_2) and deoxyhemoglobin (Hb) across the visible spectrum; RGB camera bands are indicated by shaded regions. The large HbO_2–Hb contrast in the red band (580–700 nm) enables SpO_2-dependent modulation, while the blue band (400–500 nm) provides a reference channel. (b) Melanin absorption (power-law model [43]) overlaid on hemoglobin spectra on a logarithmic scale; melanin absorbs up to 20 times more than hemoglobin in the red band, introducing substantial interference for darker skin tones. Data sources: hemoglobin values from Prahl/Oregon Medical Laser Center tabulated extinction coefficients; melanin from Jacques [43].

Limitations

Several limitations of this study should be acknowledged. First, our study enrolled 19 participants, exceeding the ISO 80601-2-61:2017 minimum of 10 subjects and more than tripling the cohort size of Hoffman et al. ($n = 6$) [28]. Nevertheless, 19 subjects remains a modest sample for establishing generalizability, particularly for the Fitzpatrick type VI subgroup ($n = 3$). To quantify uncertainty, we report 95% bootstrap confidence intervals using cluster-aware resampling at the subject level. These intervals (e.g., A_{RMS}: 5.60–7.93%, PCC: 0.373–0.732) should be interpreted as approximate indicators of uncertainty rather than precise bounds given the small cluster count.

We note that our study is exploratory in nature, providing initial evidence of feasibility and identifying skin pigmentation as a key variable for future investigation with larger, more balanced cohorts.

Second, the study was conducted under controlled laboratory conditions with fixed ambient lighting (approximately 550 lux, 3,500 K color temperature). Performance under varying real-world lighting conditions, including outdoor environments, different artificial light sources, and dynamic lighting changes, requires further investigation.

Third, the system demonstrated substantially degraded performance for participants with Fitzpatrick Skin Type VI ($A_{RMS} = 9.71\%$, $n = 3$) compared to types II–V ($A_{RMS} = 5.80\%$, $n = 16$). This finding is consistent with known challenges in optical physiological sensing for darker skin tones and is attributable to melanin absorption in the visible spectrum, which physically attenuates the pulsatile (AC) component of the reflected light signal. Figure 9 illustrates this mechanism by comparing the extracted AC amplitudes of a representative Type II–V and Type VI participant: the Type VI trace shows markedly reduced pulsatile amplitude in both the red and green channels, consistent with the theoretical absorption spectra (Figure 8). Near-infrared channels, where melanin absorption is substantially reduced, and enhanced illumination strategies represent the most promising paths to equitable performance.

Fourth, signal processing parameters (SNR weighting temperature and AC envelope exclusion percentile) were optimized via grid search on 16 participants (Fitzpatrick types II–V). While the 3 type VI participants were held out from this optimization to evaluate generalization, the optimization was not performed on an independent dataset. The RF prediction model itself used a priori hyperparameters without tuning (see Methods), but the upstream signal processing optimization may introduce indirect overfitting to the types II–V subgroup.

Fifth, the bin-restricted LOSOCV analysis demonstrated A_{RMS} as low as 2.58% within the <80% SpO_2 range, but such range-specific models are not directly deployable as a single clinical system because the patient's true SpO_2 range is unknown a priori. The full-range model ($A_{RMS} = 6.62\%$), which must operate without prior knowledge of the patient's oxygenation state, produced A_{RMS} of 7.96% in the <80% bin—an error magnitude that precludes safe absolute SpO_2 tracking during deep hypoxemia under ISO 80601-2-61 standards. The bin-restricted results should therefore be interpreted as characterizing the signal's information content within each clinical range, not as achievable operational accuracy. For the triage screening application proposed in this work, the system's value in the <80% range lies in detecting the direction and rate of desaturation rather than providing a precise absolute reading.

Sixth, participants were instructed to remain stationary during data collection. The impact of natural movements, facial expressions, and activities of daily living on measurement accuracy was not evaluated in this study.

Conclusion

In summary, this study demonstrates that an unmodified commercial RGB camera can extract SpO_2-correlated information across the full 70–100% desaturation range in a diverse cohort (Fitzpatrick types II–VI), establishing the feasibility of truly contactless pulse oximetry. The physics-based RoR approach provides interpretable intermediate

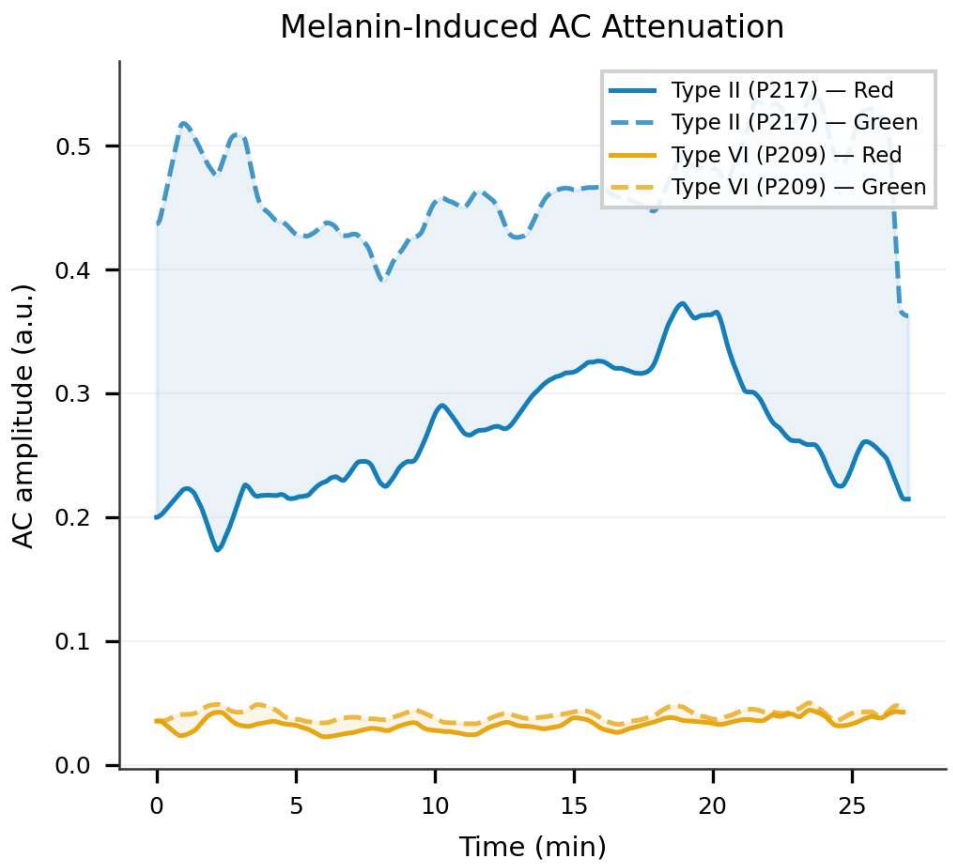


Fig. 9: **Melanin-induced attenuation of pulsatile signal amplitude.** Smoothed AC (pulsatile) components of the red (solid lines) and green (dashed lines) color channels for a representative Fitzpatrick Type II–V participant (blue) and Type VI participant (orange), overlaid on a common y-axis. The substantially reduced AC amplitude for Type VI reflects increased melanin absorption in the visible spectrum, which dampens the BVP signal that carries SpO_2 information.

outputs that enable systematic diagnosis of failure modes, as demonstrated by the outlier analyses of P202 and P215 (Section 2). Future work should address three priorities: camera-aware calibration to correct hardware-induced intercept shifts, validation on a larger multi-site cohort with balanced Fitzpatrick type representation, and real-time implementation on consumer devices to enable deployment in clinical monitoring and resource-limited settings.

Methods

Study design

The experiment was performed and administered by a clinical research organization, Clinimark, part of Element Materials Technology, which provides clinical validation services for medical devices. The experiment was approved by the Institutional Review Board (IRB) at Clinimark. A total of 20 participants were enrolled. One participant was excluded due to incomplete recordings, resulting in 19 participants included in the final analysis. All participants signed the IRB-approved informed consent form before their data collection sessions. Participants with the following conditions were excluded from the study: obesity, compromised circulation, injury, physical malformation of fingers, wrists, hands, ears, or forehead/skull or other sensor sites, tattoos, extensive moles or freckles on the face, female participants who were actively trying to become pregnant or were pregnant, respiratory conditions, and heart or cardiovascular conditions. After excluding one subject whose video was not recorded completely, data from 19 subjects with varied skin types (Fitzpatrick II–VI) were successfully collected with full controlled desaturation cycles (100–70%). A schematic illustration of the experimental setup is shown in Figure 10.

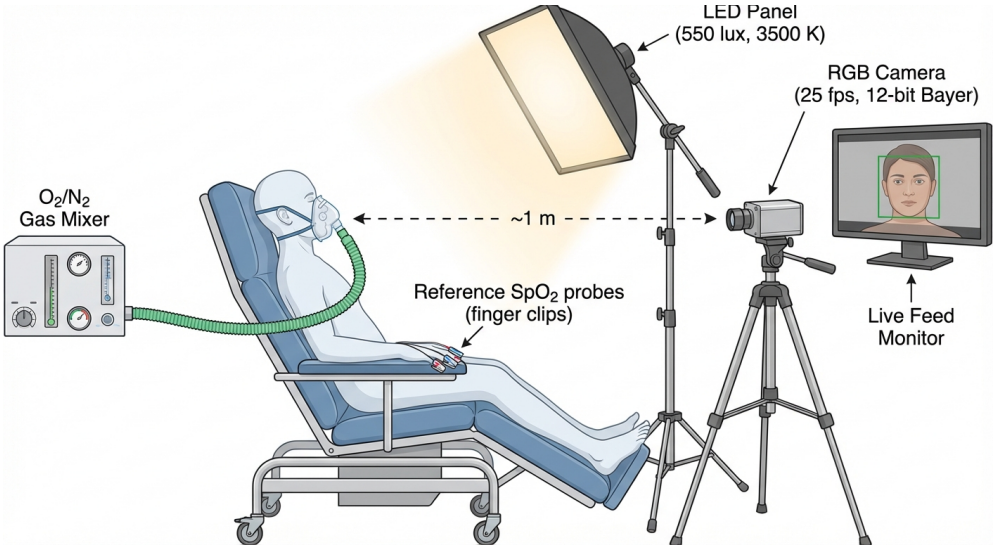


Fig. 10: Experimental setup for contactless SpO_2 measurement. A participant rests in a reclined position wearing a gas delivery mask connected to an O_2/N_2 gas mixer for controlled desaturation. An Allied Vision Alvium 1800 U-240c CMOS RGB camera, mounted on a tripod at approximately 0.6-meter (2-foot) distance, records facial video at 25 fps in raw 12-bit Bayer format. A live feed monitor displays the camera view with face detection overlay for real-time positioning verification. Two LED panels provide diffused ambient illumination at 550 lux and 3,500 K color temperature. Reference SpO_2 is measured using pulse oximeter finger clip probes (GE Datex-Ohmeda 3900). This schematic illustration was generated using Google Gemini (Gemini 3 Pro) via the OpenRouter API, with iterative prompt refinement based on laboratory photographs. The authors reviewed and verified the illustration for accuracy against the actual experimental setup.

Participant demographics are summarized in Table 3.

Table 3: **Participant demographics.** Summary of study participants including self-reported age, sex, Fitzpatrick skin type, SpO_2 range achieved during controlled desaturation, and number of paired measurements per subject.

Subject	Age	Sex	Fitzpatrick	SpO_2 Range (%)	n
P201	25	M	VI	73–97	143
P202	27	F	III	71–98	143
P203	32	M	VI	83–98	102
P204	23	F	II	74–98	111
P206	50	F	II	71–98	152
P207	41	F	III	77–98	140
P208	23	M	III	76–98	125
P209	22	F	VI	73–99	137
P210	31	F	III	75–99	121
P211	36	F	III	70–97	140
P212	35	M	III	70–99	134
P213	24	M	IV	72–97	123
P214	24	M	III	83–98	85
P215	22	F	V	71–95	121
P216	36	M	II	71–95	134
P217	41	M	II	74–97	128
P218	24	M	III	76–99	123
P220	22	M	III	74–96	124
P221	28	F	II	74–96	127

Subjects rested in a reclined position wearing a gas delivery mask for hypoxemia induction through controlled oxygen/nitrogen mixtures, operated by a trained administrator. Full desaturation from 100% to 70% took an average of 20 minutes per subject. The minimum oxygen fraction delivered was 9%, which produces blood oxygen saturation levels similar to those observed in healthy individuals at approximately 4,300 to 5,200 m elevation. The reference SpO_2 was collected from a purpose-built pulse oximeter (GE Healthcare Datex-Ohmeda 3900 with TruTrak+, which has an A_{RMS} of 2% under stationary conditions and 3% during clinical patient motion, for the range of 70–100% [44]) serving as a transfer standard that had been previously compared to gold-standard reference CO-oximetry. Measurements were obtained using a tight-tolerance finger clip, randomly rotated among the index, middle, and ring fingers of the left or right hand across subjects. The gas mixture was delivered to produce six stair-step plateaus at SpO_2 levels of 95–100%, 90–95%, 85–90%, 80–85%, 75–80%, and 67–75%, respectively, simulating blood-draw periods.

Sample size considerations

To the best of our knowledge, our cohort of 19 participants represents the largest published induced-hypoxemia evaluation of contactless camera-based SpO_2 estimation, with controlled desaturation spanning the full 70–100% range, exceeding the cohort of Hoffman et al. ($n = 6$, published in this journal) [28] by more than three-fold and meeting the ISO 80601-2-61:2017 minimum of 10 subjects [45]. Induced

hypoxemia studies require continuous physician oversight, arterial line monitoring, and IRB-approved controlled desaturation protocols, which inherently constrain practical enrollment. Despite this constraint, the cohort encompasses Fitzpatrick skin types II–VI with $N = 2{,}413$ paired SpO_2 measurements, providing sufficient statistical basis for feasibility evaluation. With $n = 3$ type VI participants, subgroup conclusions are hypothesis-generating rather than confirmatory. We report these results for transparency and to motivate targeted recruitment in future studies. To quantify uncertainty appropriately given this sample size, we report 95% bootstrap confidence intervals using bias-corrected and accelerated (BCa) resampling [46] with subject-level clustering across all primary metrics (see Statistical Analysis). A multi-site study with larger cohorts balanced across Fitzpatrick types V and VI is the essential next step for confirmatory evaluation.

Reference Standard and Subject Monitoring

The reference pulse oximeter provided real-time oxygen saturation levels for determination of plateau levels. Multiple additional pulse oximeters were used to monitor SpO_2 concurrently for safety purposes, on other fingers (Masimo Radical-7, which has a tested A_{RMS} of 2% between 70–100% and 3% between 60–80% SpO_2 [47]) and ear tip (Nellcor N-600X, which has a tested accuracy of 2.5% between 85–99% and 3% between 63–99% SpO_2 [48]), respectively.

Reference oxygen saturation was defined as a transfer-standard arterial oxygen saturation (SaO_2), derived from a clinically validated finger pulse oximeter. The transfer standard was computed as:

$$SaO_2 = \frac{SpO_2 - 7.021}{0.9222},$$

where the transformation was derived from 18 prior Clinimark hypoxemia studies involving at least 10 participants per study across a wide range of skin tones, ages, and genders.

A clinician continuously monitored the subject's vitals, including heart rhythm, pulse rate, respiration rate, inspired and expired oxygen, SpO_2, and FiO_2, and communicated with the participants regarding their comfort and sense of well-being; the study would be terminated in case of any abnormality. Blood pressure and body temperature were measured before and after data collection. The participants also wore a chest band (BioHarness) and a smartwatch (Samsung Galaxy Watch 3) to collect reference ECG and PPG signals, respectively. The Fitzpatrick scale was used to assess each participant's skin tone. Direct melanin measurements were obtained using the Mexameter [42] at the forehead.

Data annotation

Oxygen desaturation plateaus were defined as saturation ranges (e.g., 85–90%) rather than discrete values. A plateau was considered stable if the oxygen saturation variation was below 3% over a 20-second pre-window. Ground truth values were computed as

the average reference measurement within a 5–7 second simulated blood draw window, during which the min–max variation was required to be below 2%.

Several simulated "blood draw" periods (5–7 seconds each) were identified from the transfer standard pulse oximeter data at each of the SpO_2 plateaus. A valid "blood draw" period was defined by two criteria: (1) the reference SpO_2 value varied by less than 2% within the draw period, and (2) the combined minimum and maximum deviations of the reference SpO_2 were <3 during the review period of 20 seconds prior to and 10 seconds after the draw period, as defined by ISO 80601-2 [45].

Temporal synchronization across camera streams and reference devices was achieved using a combination of LED panel flashing patterns, RA-phone-triggered start/stop markers, and clock video recordings at the beginning of each session. This procedure limited inter-device temporal misalignment to within a few frames.

Camera and lighting configuration

The videos were recorded by a CMOS RGB camera (Allied Vision Alvium 1800 U-240c, 12.5mm FL f/2.5 Lens [49]). The camera was positioned at approximately 0.6 m (2 ft) from the participant's face. The camera's parameters were manually adjusted for each subject to obtain maximal brightness while preventing saturation of pixels on the face area, using minimal gain. The white balance was adjusted until the image appeared neutral in color temperature. These parameters were fixed for the respective recording sessions. The videos were recorded in a raw Bayer format with 12-bit color depth, at 25 frames per second (fps), and a $1020(H) \times 960(W)$ resolution. Two LED panels were placed in front of the participants, at a 45-degree angle with respect to the cameras to avoid specular reflection. The distance from each panel to the participant's face was approximately 216 cm. The distance between the two panels was approximately 198 cm. The light intensity and color temperature were set to 550 lux and 3,500 K.

Computational frameworks

RoI Detection

The objective of RoI detection was to identify and track regions on the human face that have visible skin and high-quality BVP signals. This process was performed in sequential steps, as described below.

1- Face Detection: The first step in identifying RoIs was to detect and track the boundary of the human face within each image frame. We used MediaPipe FaceMesh [50], a lightweight deep learning model that provides 468 facial landmarks corresponding to facial elements such as nose, mouth, eyes, and face contours. The model operated with a minimum detection confidence of 0.4 and tracking confidence of 0.3. A rectangular face region was defined based on the positions of the predicted landmarks and bounding box, with extended margins (20% width extension on each side, 40% height extension on top, and 20% on bottom) to capture the full facial area including the forehead (Figure 11a, left).

Since the subject's face may occupy only a small portion of the video frame, all subsequent steps were applied only to the extracted face region. To maintain computational efficiency and a stable frame rate, we applied adaptive image scaling. A target

face width w_{target} was defined, and the scaling ratio was determined as:

$$r = \begin{cases} 1 & \text{if } w < w_{target} \\ w/w_{target} & \text{otherwise} \end{cases} \quad (1)$$

where w is the width of the detected face. In our experiments we used $w_{target} = 140$ pixels. The face region was extracted and resized by factor r before subsequent processing, ensuring consistent resolution regardless of camera parameters and subject-camera distance.

2- Image Segmentation: The extracted face region was segmented into a 16×16 grid, yielding 256 candidate regions of interest (superpixels). This grid-based approach provides a balance between spatial resolution for capturing localized blood volume changes and computational efficiency. Prior to segmentation, a median blur filter (kernel size = 21 pixels) was applied to reduce high-frequency noise while preserving color boundaries. Each superpixel's mean RGB color was computed by averaging all pixel intensities within that region.

3- Skin Classification: Non-skin regions (e.g., eyes, hair, background) were excluded to avoid signal contamination. We employed a two-stage classification approach:

Stage 1 - Gaussian Mixture Model (GMM) Classification: A GMM with 3 components was trained on the first frame using pixels from predefined facial skin regions (forehead, cheeks). For each superpixel, a skin likelihood score was computed as the difference between the GMM score for skin versus background reference distributions. Superpixels with non-negative skin likelihood scores were classified as candidate skin regions.

Stage 2 - Homogeneity Filtering: Candidate skin regions were further filtered based on color homogeneity within each superpixel. The homogeneity score was computed as the inverse of the standard deviation across RGB channels, favoring regions with uniform skin color. Additionally, brightness thresholds were applied to exclude overexposed or underexposed regions: pixels with luminance (ITU-R BT.601: $L = 0.299R + 0.587G + 0.114B$) outside the range [50, 130] were rejected. An adaptive brightness mechanism widened this range using the 2nd and 98th percentile of the detected face luminance, ensuring that well-lit foreheads (which may exceed the nominal upper bound) were not erroneously excluded. The configured range served as a floor; the adaptive mode could only widen, never narrow, the acceptance band. A face oval filter, derived from MediaPipe FACE_OVAL landmarks, further rejected superpixels that fell outside the facial contour, preventing hair and background regions from entering the ROI set when the face padding extended into non-skin areas.

The final skin score combined both metrics, and all surviving superpixels (from the original 256) were retained as valid ROIs for signal extraction (Figure 11a, right).

4- ROI Tracking: To maintain temporal consistency of ROI signals across video frames, we implemented a coordinate-based tracking system. For each frame after the initial detection:

1. The expected position of each ROI was computed using a face-relative coordinate system that accounts for head pose changes.

2. The current superpixel at the expected location was identified as the tracked ROI.
3. If the color difference between the tracked and original ROI exceeded a threshold (30 RGB units), a local search among neighboring superpixels identified the best match based on combined position and color similarity.
4. Superpixels with insufficient pixel counts (<50 pixels) were excluded from tracking.

To balance tracking accuracy with computational overhead, face detection was repeated every 6 seconds to correct for accumulated drift. This interval was determined empirically through grid search optimization, balancing facial tracking accuracy against step artifacts introduced by detection boundary changes [51].

rPPG Signal Extraction

The recorded video used Bayer 12-bit format. Pixels were unpacked and converted to RGB format, then processed frame-by-frame to generate rPPG signals at the native video frame rate (25 fps).

Signal Averaging: For each validated skin ROI identified in the previous step, we computed the arithmetic mean of pixel intensities for each color channel (R, G, B) separately. This spatial averaging was justified by the assumption that skin tissue composition and light reflection characteristics are homogeneous within each small ROI region [52]. The resulting raw signals are shown in Supplementary Figure S5a, which illustrates per-ROI signal quality variation across good, median, and poor-quality ROIs.

Temporal Buffering: The averaged RGB intensities were concatenated over time using a sliding window approach (window length: 20 seconds, step size: 1 second) and stored in a first-in-first-out (FIFO) buffer. This yielded rPPG signals with shape $N \times T \times 3$, where N is the number of skin ROIs and T is the number of samples in the window.

Temporal Resampling: To compensate for potential frame timing irregularities and improve signal quality for subsequent spectral analysis, each rPPG signal was linearly interpolated and resampled to 100 Hz [53].

Signal Quality Screening: The eligibility of each ROI was assessed by examining the temporal consistency of its Red channel signal. For each ROI, the normalized deviation (standard deviation divided by the signal mean) was computed; ROIs exceeding a relative deviation threshold of 0.5% with more than 80% of their temporal samples flagged were excluded from all subsequent processing. This quality-first ordering ensured that step correction was only applied to ROIs with adequate signal fidelity.

Chrominance-Based Pulse Extraction (CHROM): After quality screening and step artifact correction, a chrominance-based method (CHROM) [6] was applied independently to each surviving ROI to extract the BVP signal. The CHROM algorithm linearly combines the normalized R, G, and B channels using chrominance projections designed to cancel specular reflection artifacts while preserving the pulsatile component. Specifically, the two chrominance signals $X = 3R - 2G$ and $Y = 1.5R + G - 1.5B$ were computed, and the pulse signal was obtained as $S = X - (\sigma_X/\sigma_Y) \cdot Y$, where σ denotes the standard deviation within the analysis

window. Each ROI's channels were mean-normalized and linearly detrended prior to CHROM projection to remove slow baseline drift.

Multi-ROI Fusion: The rPPG signals from multiple surviving ROIs were merged into a single set of RGB signals using signal-to-noise ratio (SNR) weighted averaging [41]. The SNR for each ROI was estimated from the spectral power at the detected heart rate frequency relative to out-of-band noise power. This weighting scheme ensured that ROIs with cleaner physiological signals contributed more to the final fused signal (Supplementary Figure S5b).

Jitter Removal: Although subjects were instructed to remain stationary, involuntary motions (e.g., respiration-induced movements) caused gradual face position drift. When face detection was repeated every 6 seconds to correct boundary drift, the statistical nature of the detection model introduced discontinuities in ROI boundaries, manifesting as step artifacts in the extracted signals. We implemented a jitter removal procedure that detected these discontinuities at known face re-detection timestamps and applied correction by subtracting the step offset from subsequent samples (Figure 11b).

Signal extraction walkthrough

The computational framework processed signals through three stages: ROI detection, RoR feature extraction, and SpO_2 prediction (Figure 11). The human face served as the primary skin region due to its large area and high blood perfusion, which enabled recovery of high-quality BVP signals. Smaller ROIs within the face were identified to reduce noise and improve BVP signal reconstruction (see Section 2). The RoR extractor converted the reflective lighting variation at different spectral bands into SpO_2-related features. To illustrate the information flow, we present a step-by-step visualization for a representative subject (P220).

The signals extracted from the videos were visualized in Figure 11. Face detection and ROI selection (Figure 11a) identified valid skin regions on the face: MediaPipe FaceMesh landmarks defined the face boundary, a superpixel grid partitioned the region into candidate ROIs, and skin classification filters retained only valid skin patches. Per-ROI signal quality varied substantially (Supplementary Figure S5a): high-quality ROIs on stable skin regions (e.g., forehead) produced clean pulsatile signals, while ROIs near the face boundary or occluded areas exhibited higher noise. Quality screening excluded noisy ROIs from downstream processing.

Step artifacts at 6-second face re-detection boundaries are visible in zoomed signal segments (Figure 11b, top, black). As tested with different face detection algorithms, including MediaPipe [50] and our in-house algorithm [51], the face could not be aligned perfectly from frame to frame due to the statistical nature of these models, causing jitters in the extracted rPPG signals. The cleaned signal after jitter removal (Figure 11b, bottom, vermillion) reveals clean blood pulses. After chrominance-based pulse extraction and SNR-weighted multi-ROI fusion (Supplementary Figure S5b), the fused signal was processed to extract AC and DC components.

Figure 11c shows the smoothed AC (vermillion, left axis) and DC (black, right axis) components from the Red channel. During desaturation, the AC signal increases as deoxygenated hemoglobin absorbs more red light, increasing the pulsatile modulation

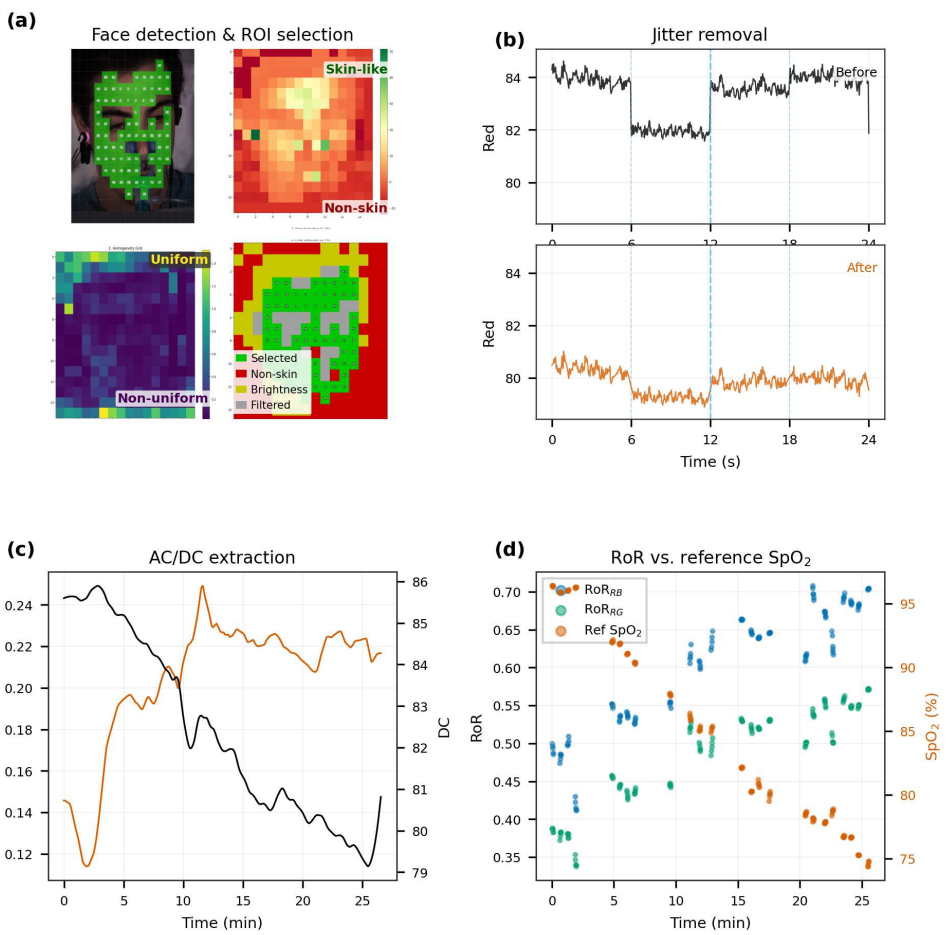


Fig. 11: **Signal extraction walkthrough for a representative subject (P220).** (a) Face detection and ROI selection: superpixel grid overlay on the detected face with selected ROIs highlighted in green (top left), GMM skin likelihood map (red-to-green scale; top right), color homogeneity map (bottom left), and final ROI classification (bottom right; green = selected skin ROI, red = non-skin rejected by GMM, yellow = brightness rejected, gray = homogeneity filtered). (b) Jitter removal at face re-detection boundaries (dashed vertical lines): raw Red channel signal with step artifacts (top, black) and corrected signal (bottom, vermillion). (c) Smoothed AC (vermillion, left axis) and DC (black, right axis) components from the Red channel; AC increases as SpO_2 decreases during the desaturation protocol. (d) RoR_{RB} (blue) and RoR_{RG} (green) versus reference SpO_2 (vermillion, right axis). Per-ROI signal quality and multi-ROI fusion details are shown in Supplementary Figure S5.

depth—consistent with the absorption spectra (Figure 8). Figure 11d presents both RoR_{RB} (blue) and RoR_{RG} (green) alongside the reference SpO_2 (vermillion, right axis), confirming that RoR features track SpO_2 throughout the desaturation–recovery protocol.

SpO_2 estimation

SpO_2 was estimated every second using a 20-second sliding window of the fused rPPG signal.

AC/DC extraction (envelope–median method): For each color channel within the analysis window, the pulsatile (AC) and steady-state (DC) components were extracted using an envelope–median approach. First, the instantaneous heart rate was estimated via spectral analysis (dominant frequency within [0.8, 3.5] Hz). A narrow bandpass filter centered at the detected heart rate with ±0.2 Hz sidebands (total bandwidth 0.4 Hz, corresponding to ±12 bpm) was applied to isolate the cardiac-frequency component. Peak-to-valley amplitudes were measured on the filtered signal, per-beat outliers beyond the 10th–90th percentile range were excluded, and the AC value was taken as the median of the surviving amplitudes. The DC component was defined as the mean of the unfiltered signal within the window. This procedure yielded one AC/DC ratio per channel per window, which was then interpolated to the full signal length using 1-second steps.

Ratio of ratios: Two RoR features were computed from the AC/DC ratios:

$$\mathrm{RoR}_{\mathrm{RB}} = \frac{\mathrm{AC}_R/\mathrm{DC}_R}{\mathrm{AC}_B/\mathrm{DC}_B}, \quad \mathrm{RoR}_{\mathrm{RG}} = \frac{\mathrm{AC}_R/\mathrm{DC}_R}{\mathrm{AC}_G/\mathrm{DC}_G} \tag{2}$$

These two features served as the input to the regression model.

In addition to linear regression, we evaluated Random Forest regression [54] with a maximum depth of 3, using the same two RoR features ($\mathrm{RoR}_{\mathrm{RG}}$ and $\mathrm{RoR}_{\mathrm{RB}}$) as input. The RF model was selected as the primary prediction model based on lowest LOSOCV $\mathrm{A}_{\mathrm{RMS}}$ (see Results). The minimal feature set of two physics-derived features was chosen to mitigate overfitting risk given the modest sample size ($n = 19$ subjects).

One-point calibration

To evaluate individualized calibration without full per-subject reference data, we implemented a slope-constrained one-point calibration strategy. For each participant, we estimated a population-level $\mathrm{RoR}_{\mathrm{RG}}$–$SpO_2$ slope by fitting ordinary least-squares linear regressions to all *other* participants' data (leave-one-out cross-validation of the slope), yielding $\alpha_{\mathrm{pop}} = -100.2$ (CV = 20.4%, range: -133.8 to -69.5). The near-parallel slopes across participants (Figure 6a) motivated fixing the slope to the population mean and estimating only the subject-specific intercept.

For each participant, the chronologically first concurrent ($\mathrm{RoR}_{\mathrm{RG}}$, SpO_2) measurement from a reference pulse oximeter served as the calibration point ($\mathrm{RoR}^{\mathrm{CAL}}$, $\mathrm{SpO}_2^{\mathrm{CAL}}$). This choice was deliberate: using the first available sample avoids any retrospective selection bias and mirrors the intended clinical workflow, in which a single pulse oximeter reading would be taken at the start of monitoring to personalize the camera-based system. The calibration point was excluded from all subsequent evaluation. The per-subject intercept was computed as $\beta_i = \mathrm{SpO}_2^{\mathrm{CAL}} - \alpha_{\mathrm{pop}} \cdot \mathrm{RoR}^{\mathrm{CAL}}$, and all subsequent predictions used $\widehat{\mathrm{SpO}_2} = \alpha_{\mathrm{pop}} \cdot \mathrm{RoR}_{\mathrm{RG}} + \beta_i$. This approach requires

only one concurrent pulse oximeter reading per new user, making it suitable for point-of-care deployment while avoiding the noise-amplification inherent in estimating both slope and intercept from a single calibration point.

Statistical analysis

All analyses were performed using Python 3.11 with SciPy [55] and scikit-learn [56].

Validation scheme. Model performance was evaluated using leave-one-subject-out cross-validation (LOSOCV): each of the 19 participants served in turn as the held-out test set, with the remaining 18 used for training. This scheme prevents data leakage across subjects and provides an unbiased estimate of generalization performance on unseen individuals.

Regression metrics. Prediction accuracy was quantified using the accuracy root mean square (A_{RMS}), defined per ISO 80601-2-61:2017 [45] as:

$$A_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

where y_i and $\hat{y}_i$ denote the reference and predicted SpO_2 values, respectively, and N is the number of samples. A_{RMS} is numerically identical to the root mean square error (RMSE). Additional metrics include mean absolute error (MAE), Pearson correlation coefficient (PCC), and prediction bias (mean signed error).

Model selection. Four candidate model families were evaluated: Random Forest regression [54] (maximum tree depth = 3), Ridge regression, ElasticNet, and ordinary least-squares (OLS) linear regression. The RF model was selected as the primary prediction model based on lowest LOSOCV A_{RMS}. A minimal feature set of two ratio-of-ratios features (RoR_{RG} and RoR_{RB}) was used to mitigate overfitting risk given the modest sample size ($n = 19$ subjects).

Agreement analysis. Bland–Altman analysis [57, 58] was used to assess agreement between predicted and reference SpO_2. The analysis reports bias (mean difference), standard deviation (SD) of the differences, and 95% limits of agreement (LOA = bias $\pm$ 1.96 $\times$ SD). Proportional bias was evaluated by linear regression of the differences on the means. Heteroscedasticity was tested using the Breusch–Pagan test; when not significant, standard (constant-width) LOA were reported. Systematic bias was assessed with the Wilcoxon signed-rank test.

Classification analysis. Hypoxemia detection was evaluated as a binary classification task at SpO_2 thresholds of 90% and 88%. Receiver operating characteristic (ROC) curves were constructed and the area under the curve (AUC) was computed. Sensitivity, specificity, and accuracy were reported at the optimal operating point determined by Youden's J statistic.

Range-specific model evaluation. To characterize prediction accuracy within clinically defined SpO_2 ranges, bin-restricted LOSOCV was performed for five ISO 80601-2-61 bins (<80%, <90%, 80–90%, ≥80%, ≥90%). For each bin, a separate model was trained and tested exclusively on samples falling within that range. Subjects with fewer than 3 samples in a given range were excluded from that bin's

analysis. Note that overlapping bins (e.g., <90% contains both <80% and 80–90%) represent distinct clinical accuracy claims under the ISO standard. This approach isolates the model's ability to discriminate SpO_2 values within clinically homogeneous ranges from the cross-bin prediction uncertainty that inflates full-range error.

Feature ablation. To assess the contribution of each RoR feature, three feature configurations were compared: RoR_{RB} only, RoR_{RG} only, and both features combined. All four model families (OLS, Ridge, ElasticNet, Random Forest) were evaluated for each configuration using LOSOCV.

Type VI exclusion analysis. To test whether including Fitzpatrick type VI participants in the training set affects performance for other groups, LOSOCV was repeated with type VI subjects excluded from training folds but retained as test subjects. Performance was compared between inclusive ($n = 19$ training) and exclusive ($n = 16$ training) conditions for both type II–V and type VI test groups.

Per-subject calibration analysis. To quantify the potential benefit of individualized calibration, each subject k's LOSOCV predictions $\hat{y}_k$ were rescaled via ordinary least-squares: $\hat{y}_k^{\text{cal}} = \alpha_k \hat{y}_k + \beta_k$, where $(\alpha_k, \beta_k) = \text{argmin} \sum_i (y_{k,i} - \alpha \hat{y}_{k,i} - \beta)^2$. Because the same reference data are used for fitting and evaluation, this represents an upper bound on achievable accuracy through individualized linear correction. The A_{RMS} improvement relative to uncalibrated predictions was assessed using the Wilcoxon signed-rank test on per-subject A_{RMS} values.

Confidence intervals. All confidence intervals were computed using the bias-corrected and accelerated (BCa) bootstrap [46] with 10,000 iterations. To account for within-subject correlation structure, subject-level cluster resampling was employed: at each iteration, subjects were resampled with replacement and all samples from each selected subject were included. This approach preserves the dependence structure inherent in repeated measurements. A percentile bootstrap fallback was used when BCa acceleration estimates produced edge cases. Given the modest sample size ($n = 19$ subjects), confidence intervals should be interpreted as approximate indicators of uncertainty rather than precise bounds.

Stratified analysis. Overall performance was evaluated using full-cohort LOSOCV ($n = 19$). For the types II–V subgroup, performance was evaluated using a dedicated 16-subject LOSOCV in which type VI participants were excluded from both training and testing, ensuring that the II–V accuracy claim reflects a model trained exclusively on types II–V. Type VI performance was evaluated within the full 19-subject LOSOCV (where type VI subjects benefit from inclusive training data). This stratification follows prior work demonstrating melanin-dependent attenuation of photoplethysmographic signals [37].

Missing data handling. Signal windows in which the quality filter rejected all ROIs (producing no valid AC/DC estimate) were excluded from the paired analysis; $N = 2{,}413$ paired measurements survived this complete-case criterion across all 19 subjects. No imputation was performed.

Model development transparency (TRIPOD+AI Item 12c)

To distinguish signal processing optimization from machine learning model evaluation, we report the following. All regression models were evaluated with a priori hyperparameters; no model-specific hyperparameter tuning was performed. Specifically: Random Forest used $n_{\text{estimators}} = 100$, maximum depth = 3, and minimum samples per leaf = 20 (with random_state = 42 for reproducibility); Ridge regression used regularization parameter $\alpha = 1.0$; ElasticNet used $\alpha = 0.1$ and ℓ_1 ratio = 0.5 with maximum 10,000 iterations; and ordinary least-squares (OLS) regression used default scikit-learn LinearRegression settings. Model selection was based exclusively on LOSOCV A_{RMS} (the lowest A_{RMS} across the four model families) without hyperparameter optimization.

Separately, two signal processing parameters were optimized via grid search on 16 participants (Fitzpatrick types II–V) prior to ML model development: the SNR weighting temperature for multi-ROI fusion (optimized value: $\tau = 4.0$, from a default of 2.0) and the AC envelope exclusion percentile for RoR feature extraction (optimized value: 10th–90th percentile, from a default of 20th–80th). The 3 Fitzpatrick type VI participants were excluded from this signal processing optimization to evaluate generalization to unseen skin type distributions. This temporal and conceptual separation ensures that the ML prediction models were evaluated with fixed, a priori hyperparameters on data whose signal processing parameters were determined independently.

All analyses were implemented in Python 3.11 using scikit-learn [56] (version 1.6.1) and SciPy [55]. A complete TRIPOD+AI compliance checklist mapping all 27 reporting items to their manuscript locations is provided in Supplementary Table S3.

Data availability

The raw video recordings and physiological data generated during this study cannot be made publicly available due to participant privacy protections under the Health Insurance Portability and Accountability Act (HIPAA) and proprietary restrictions of Samsung Research America. Processed aggregate results, statistical summaries, and all individual-level performance metrics are reported in the manuscript and Supplementary Information. Requests for data access for non-commercial research purposes may be directed to the corresponding author, subject to a formal institutional data sharing agreement.

Code availability

The computational framework developed for this study is proprietary to Samsung Research America. The core algorithms—including the physics-based ratio-of-ratios (RoR) feature extraction, CHROM-based pulse signal processing, and RF regression model specification (two RoR feature input)—are described in the Methods section with sufficient detail to enable independent replication. Requests for code access for non-commercial research purposes may be directed to the corresponding author.

Acknowledgements

The authors thank Clinimark (Element Materials Technology) for conducting the study. The authors acknowledge funding support from Samsung Research America, Inc. The schematic illustration of the experimental setup (Figure 10) was generated using Google Gemini (Gemini 3 Pro), a generative AI model accessed via the OpenRouter API. The illustration was produced through iterative prompt refinement guided by laboratory photographs, and was reviewed and verified by the authors for accuracy.

Author contributions

L.Z.: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing. M.G.: Investigation, Validation. K.V.: Methodology, Software. Q.S.: Methodology, Software. S.D.: Resources, Writing – Review & Editing. J.B.: Supervision, Project Administration. J.K.: Resources, Writing – Review & Editing. A.G.: Supervision, Funding Acquisition, Resources. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing non-financial interests, but the following competing financial interests: authors affiliated with Samsung Research America received salary. Data collection and research were funded by Samsung Research America (Mountain View, CA 94043, USA).

Ethics approval

This study protocol was approved by the Salus Institutional Review Board (Protocol Number: PR2022-504). All participants provided written informed consent prior to enrollment, including consent for the use of their physiological and video data for research purposes. No adverse events occurred during the controlled desaturation protocol. This diagnostic accuracy study was not prospectively registered in a clinical trial registry, as it evaluates a novel index test against an established reference standard under an approved IRB protocol rather than a therapeutic intervention. The full study protocol is available from the corresponding author upon reasonable request.

References

[1] Zaidan, M. F., Ameredes, B. T. & Calhoun, W. J. Management of acute asthma in adults in 2020. *JAMA* **323**, 563–564 (2020).

[2] Fawzy, A. *et al.* Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. *JAMA Internal Medicine* **182**, 730–738 (2022).

[3] Swets, M. C. *et al.* Evaluation of pragmatic oxygenation measurement as a proxy for Covid-19 severity. *Nature Communications* **14**, 7374 (2023).

[4] Levy, J., Álvarez, D., Del Campo, F. & Behar, J. A. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. *Nature Communications* **14**, 4881 (2023).

[5] Hafen, B. B. & Sharma, S. Oxygen saturation. *StatPearls [Internet]* (2023).

[6] De Haan, G. & Jeanne, V. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* **60**, 2878–2886 (2013).

[7] Fuchs, H. *et al.* Brain oxygenation monitoring during neonatal resuscitation of very low birth weight infants. *Journal of Perinatology* **32**, 356–362 (2012).

[8] Dasari, A., Prakash, S. K. A., Jeni, L. A. & Tucker, C. S. Evaluation of biases in remote photoplethysmography methods. *npj Digital Medicine* **4**, 91 (2021).

[9] Bonometti, F. *et al.* Usability of a continuous oxygen saturation device for home telemonitoring. *Digital Health* **9**, 20552076231194547 (2023).

[10] Verkruysse, W., Svaasand, L. O. & Nelson, J. S. Remote plethysmographic imaging using ambient light. *Optics Express* **16**, 21434–21445 (2008).

[11] Poh, M.-Z., McDuff, D. J. & Picard, R. W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express* **18**, 10762–10774 (2010).

[12] Poh, M.-Z., McDuff, D. J. & Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering* **58**, 7–11 (2011).

[13] Wang, W., Den Brinker, A. C., Stuijk, S. & De Haan, G. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* **64**, 1479–1491 (2017).

[14] Tarassenko, L. *et al.* Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement* **35**, 807–831 (2014).

[15] Wei, W., Vatanparvar, K., Zhu, L., Kuang, J. & Gao, A. *Remote photoplethysmography and heart rate estimation by dynamic region of interest tracking. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3243–3248 (IEEE, 2022).

[16] Huang, B. *et al.* Challenges and prospects of visual contactless physiological monitoring in clinical study. *npj Digital Medicine* **6**, 231 (2023).

[17] Van Gastel, M., Stuijk, S. & De Haan, G. New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring. *Scientific Reports* **6**, 38609 (2016).

[18] Bal, U. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical Optics Express* **6**, 86–97 (2015).

[19] Akamatsu, Y., Onishi, Y. & Imaoka, H. *Heart rate and oxygen saturation estimation from facial video with multimodal physiological data generation. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1111–1115 (IEEE, 2022).

[20] Guazzi, A. R. *et al.* Non-contact measurement of oxygen saturation with an RGB camera. *Biomedical Optics Express* **6**, 3320–3338 (2015).

[21] Kim, N. H., Yu, S.-G., Kim, S.-E. & Lee, E. C. Non-contact oxygen saturation measurement using YCgCr color space with an RGB camera. *Sensors* **21**, 6120 (2021).

[22] Casalino, G., Castellano, G. & Zaza, G. Evaluating the robustness of a contact-less mHealth solution for personal and remote monitoring of blood oxygen saturation. *Journal of Ambient Intelligence and Humanized Computing* **14**, 8871–8880 (2022).

[23] Shao, D. *et al.* Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Transactions on Biomedical Engineering* **63**, 1091–1098 (2015).

[24] Moço, A. & Verkruysse, W. Pulse oximetry based on photoplethysmography imaging with red and green light: calibratability and challenges. *Journal of Clinical Monitoring and Computing* **35**, 123–133 (2021).

[25] Akamatsu, Y., Onishi, Y. & Imaoka, H. *Blood oxygen saturation estimation from facial video via DC and AC components of spatio-temporal map. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2023).

[26] Wu, B.-J., Wu, B.-F., Dong, Y.-C., Lin, H.-C. & Li, P.-H. Peripheral oxygen saturation measurement using an RGB camera. *IEEE Sensors Journal* **23**, 26551–26563 (2023).

[27] United States Food and Drug Administration. Review of pulse oximeters and factors that can impact their accuracy. https://www.fda.gov/media/162709/download (2022). Accessed: March 5, 2026.

[28] Hoffman, J. S. *et al.* Smartphone camera oximetry in an induced hypoxemia study. *npj Digital Medicine* **5**, 146 (2022).

[29] Van Gastel, M. & Verkruysse, W. Contactless SpO_2 with an RGB camera: experimental proof of calibrated SpO_2. *Biomedical Optics Express* **13**, 6791–6802 (2022).

[30] Cheng, K.-L. & So, R. H. Y. Contactless blood oxygen saturation estimation from facial videos using deep learning. *Bioengineering* **11**, 251 (2024).

[31] Ye, Y. *et al.* Notch RGB-camera based SpO_2 estimation: a clinical trial in a neonatal intensive care unit. *Biomedical Optics Express* **15**, 428–445 (2024).

[32] Hajr, A., Tarvirdizadeh, B., Alipour, K. & Ghamari, M. Contactless health monitoring: an overview of video-based techniques utilising machine/deep learning. *IET Wireless Sensor Systems* **15**, e70009 (2025).

[33] Gupta, A., Ravelo-Garcia, A. G. & Dias, F. M. Availability and performance of face based non-contact methods for heart rate and oxygen saturation estimations: a systematic review. *Computer Methods and Programs in Biomedicine* **219**, 106771 (2022).

[34] Webster, J. G. *Design of Pulse Oximeters* (CRC Press, Boca Raton, FL, 1997).

[35] Shao, Q. *et al. Normalization is all you need: robust full-range contactless SpO_2 estimation with an RGB camera across users. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2024).

[36] U.S. Food and Drug Administration. Performance evaluation of pulse oximeters: guidance for industry and Food and Drug Administration staff (draft). https://www.fda.gov/media/175828/download (2025). Draft guidance issued January 2025. Accessed: March 5, 2026.

[37] Nowara, E. M., McDuff, D. & Veeraraghavan, A. *A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 284–285 (IEEE, 2020).

[38] Breusch, T. S. & Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–1294 (1979).

[39] Wang, R.-X., Sun, H.-M., Hao, R.-R., Pan, A. & Jia, R.-S. TransPhys: Transformer-based unsupervised contrastive learning for remote heart rate measurement. *Biomedical Signal Processing and Control* **86**, 105058 (2023).

[40] Lu, H., Han, H. & Zhou, S. K. *Dual-GAN: Joint BVP and noise modeling for remote physiological measurement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12404–12413 (IEEE, 2021).

[41] Chan, M. *et al. Estimating SpO_2 with deep oxygen desaturations from facial video under various lighting conditions: a feasibility study. 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–5 (IEEE, 2023).

[42] Courage + Khazaka Electronic GmbH. Mexameter® MX 18: Skin Pigmentation Measurement Instrument. https://www.courage-khazaka.com/en/scientific-products/occupational-health/occupational-health/169-mexameter-e (2026). Accessed: March 5, 2026.

[43] Jacques, S. L. Skin optics. *Oregon Medical Laser Center News* (1998). Available at https://omlc.org/news/jan98/skinoptics.html. Accessed: March 5, 2026.

[44] GE Healthcare. 3800/3900/3900P pulse oximeter technical reference manual (2024).

[45] International Organization for Standardization. Medical electrical equipment — Part 2-61: Particular requirements for basic safety and essential performance of pulse oximeter equipment. https://www.iso.org/standard/67963.html (2017). Accessed: March 5, 2026.

[46] Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**, 171–185 (1987).

[47] Masimo. Radical-7 pulse CO-Oximeter operator's manual. https://techdocs.masimo.com/globalassets/techdocs/pdf/lab-5475l_master.pdf (2024). Accessed: March 5, 2026.

[48] Nellcor. OxiMax N-600x pulse oximeter service manual (2024).

[49] Allied Vision. Alvium 1800 U-240. https://www.alliedvision.com/en/products/alvium-configurator/alvium-1800-u/240/ (2024). Accessed: March 5, 2026.

[50] Lugaresi, C. *et al.* MediaPipe: a framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[51] Li, J. *et al. Advancements in face alignment evaluation for contact-less vital sign detection. 2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 1–4 (IEEE, 2023).

[52] Vatanparvar, K. *et al. Enhanced contactless heart rate monitoring using camera with motion artifact removal during physical activities. 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–5 (IEEE, 2023).

[53] Chan, M. *et al. Improving heart rate and heart rate variability estimation from video through a HR-RR-tuned filter. ICASSP 2023-2023 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5 (IEEE, 2023).

[54] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

[55] Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

[56] Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[57] Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **327**, 307–310 (1986).

[58] Bland, J. M. & Altman, D. G. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**, 135–160 (1999).

[59] Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).

Supplementary Materials

Contactless Human Oxygen Saturation Detection Using RGB Camera in an Induced Hypoxemia Study with Varied Skin Types

Authors: Li Zhu[1*], Migyeong Gwak[1*], Korosh Vatanparvar[1], Qijia Shao[2], Sharanya Desai[1], Jungmok Bae[1], Jilong Kuang[1], Alex Gao[1]

Affiliations:
[1] Digital Health Team, Samsung Research America, 665 Clyde Ave, Mountain View, 94043, CA, USA
[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Work performed during Q. Shao's internship at Samsung Research America.
[*] To whom correspondence should be addressed; E-mail: li.zhu1@samsung.com; m.gwak@samsung.com.

The following supplementary materials are available:

- **Supplementary Figure S1:** STARD 2015 participant flow diagram showing enrollment, exclusions, and analysis subgroups.
- **Supplementary Figure S2:** Per-subject Bland-Altman plots from the Type VI exclusion experiment. (a) Type II–V subjects (n=16) with leave-one-subject-out cross-validation excluding Type VI from training. (b) Type VI subjects (n=3) tested against the Type II–V-trained model. All panels show one-point calibrated predictions.
- **Supplementary Figure S3:** Feature ablation analysis comparing $\mathrm{RoR}_{\mathrm{RB}}$-only, $\mathrm{RoR}_{\mathrm{RG}}$-only, and combined feature configurations for the best-performing RF model.
- **Supplementary Figure S4:** Lighting variation effects on RoR–SpO_2 correlation across ambient lighting conditions.
- **Supplementary Figure S5:** Signal extraction detail showing per-ROI signal quality variation and multi-ROI fusion for a representative subject.
- **Supplementary Table S1:** Extended per-subject performance metrics ($\mathrm{A}_{\mathrm{RMS}}$, MAE, RoR feature correlation, bias, SD) for all 19 participants.
- **Supplementary Table S2:** Bootstrap confidence intervals for aggregate performance metrics.
- **Supplementary Table S3:** TRIPOD+AI compliance checklist [59] with item-by-item manuscript section references.

Supplementary Figure S1: STARD Flow Diagram

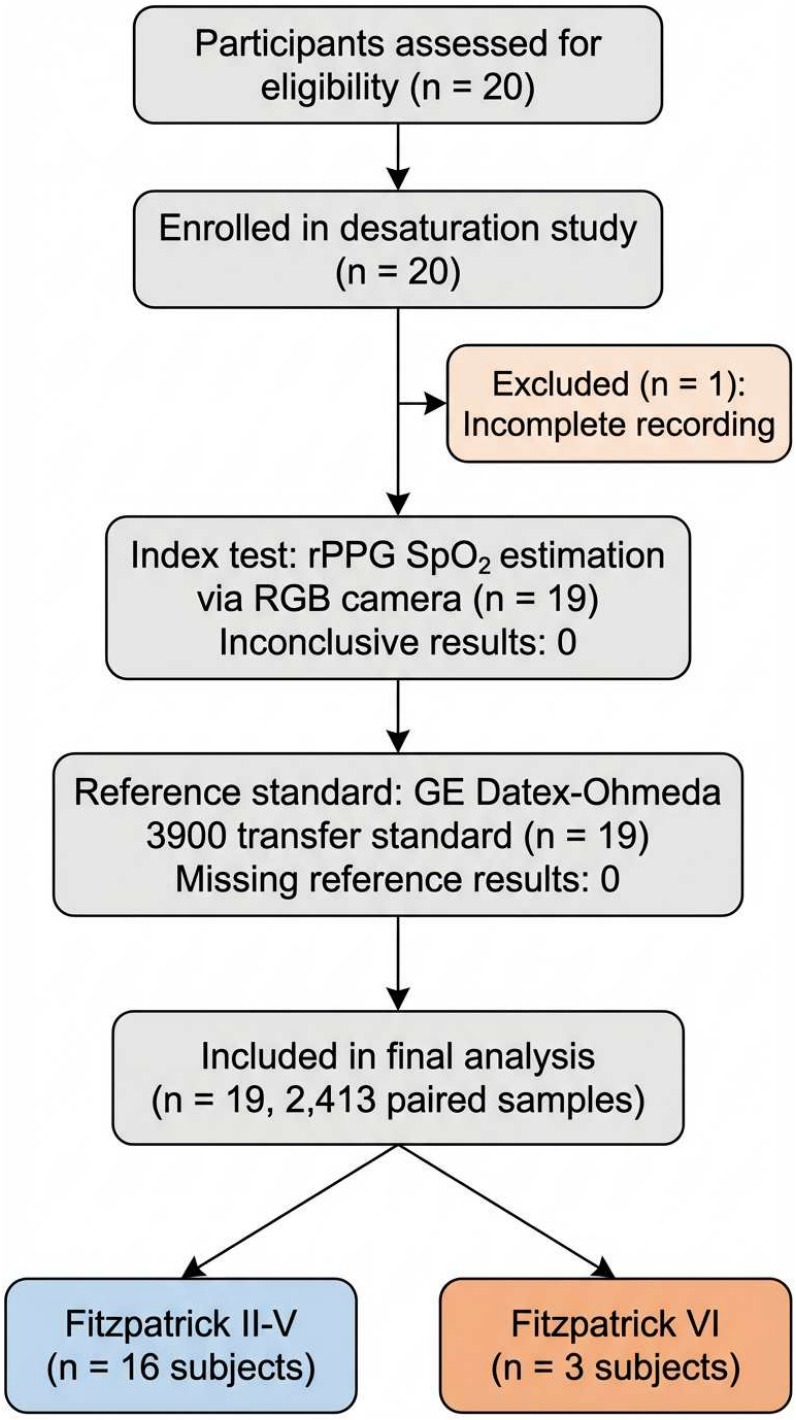


Fig. S1: **STARD 2015 participant flow diagram.** Twenty participants were assessed for eligibility, of whom 19 completed the desaturation protocol with both index test (rPPG SpO_2) and reference standard (GE Datex-Ohmeda 3900 transfer standard), yielding 2,413 paired samples. One participant was excluded due to an incomplete recording. The final cohort was stratified into Fitzpatrick Type II–V (n = 16) and Type VI (n = 3) subgroups.

Supplementary Figure S2: Per-Subject Bland-Altman Plots

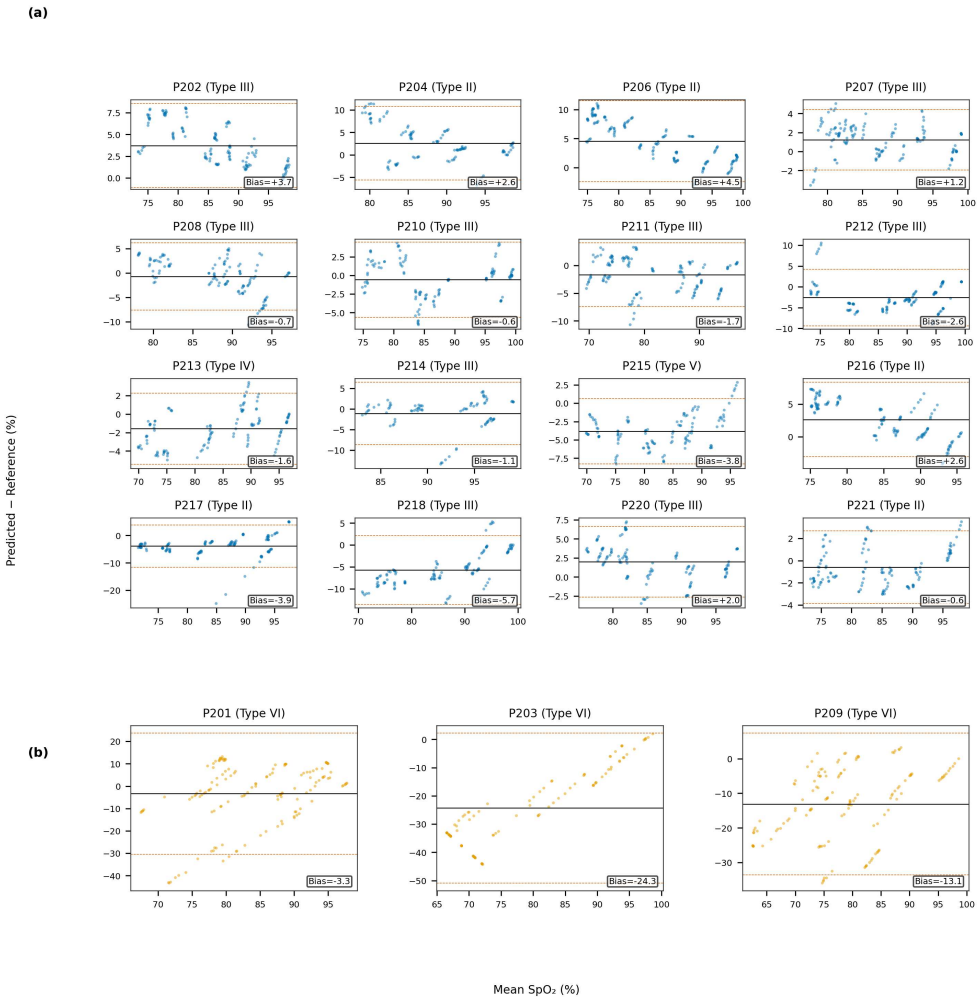


Fig. S2: **Per-subject Bland-Altman agreement plots (Type VI exclusion experiment).** (a) Individual Bland-Altman plots for 16 Type II–V participants using LOSOCV predictions from a model trained exclusively on Type II–V data (Type VI excluded from all training folds). (b) Three Type VI participants tested against the Type II–V-trained model. Solid line: mean bias; dashed lines: 95% limits of agreement. Blue: Type II–V; orange: Type VI. One-point calibrated predictions used throughout. Diagonal trends in some panels reflect proportional bias at lower SpO_2 levels.

Supplementary Figure S3: Feature Ablation

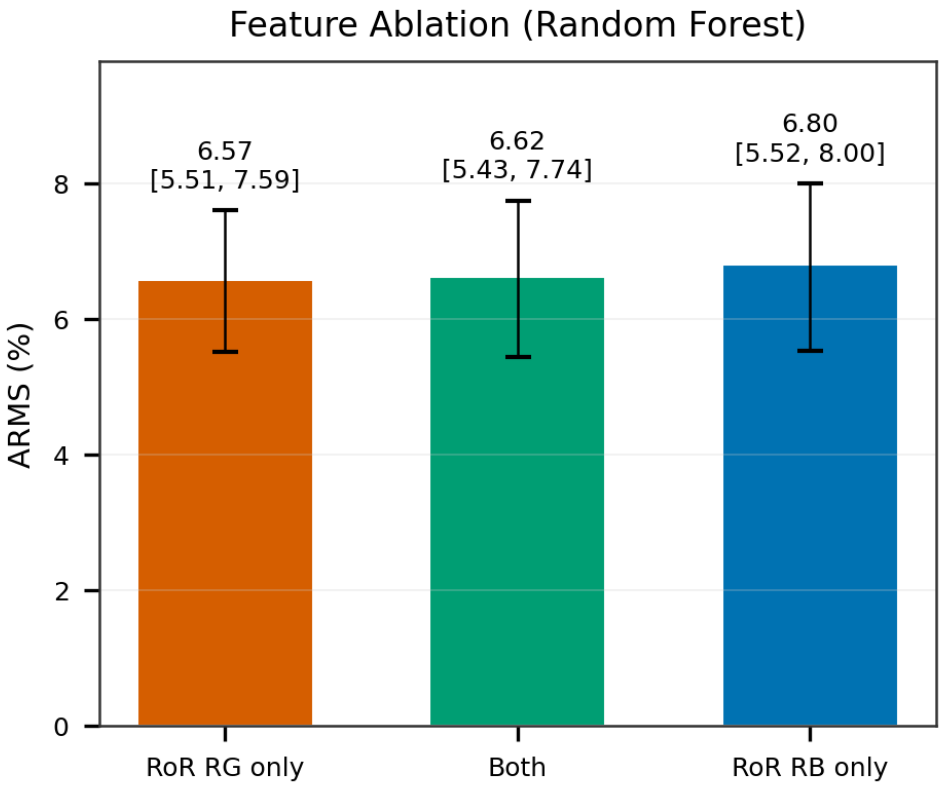


Fig. S3: **Feature ablation.** A_{RMS} for three feature configurations using the best-performing RF model: RoR_{RB} only (6.80%), RoR_{RG} only (6.57%), and both features combined (6.62%). Confidence intervals are from LOSOCV-based bootstrap within the ablation analysis (distinct from the aggregate BCa cluster bootstrap in Table S2). Overlapping intervals indicate no statistically significant difference between single-feature and combined models; the combined model was selected as the primary result because the two-feature input is motivated by the physics-based Beer–Lambert derivation.

Supplementary Figure S4: Lighting Variation

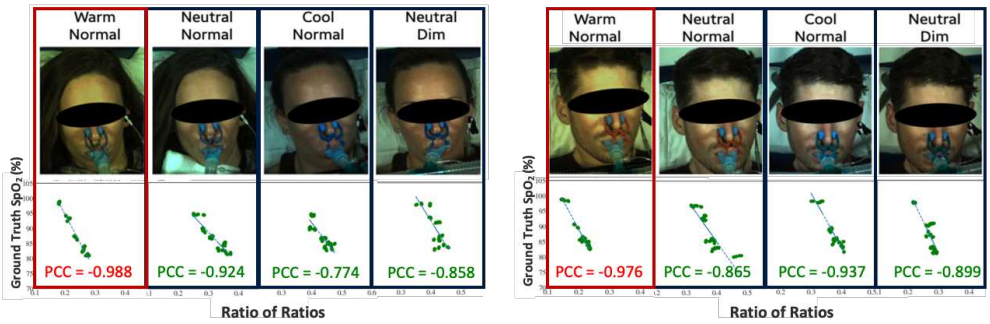


Fig. S4: **Lighting variation effects on RoR–SpO_2 correlation.** Per-subject RoR–SpO_2 scatter plots under different ambient lighting conditions from our pilot study, demonstrating that changes in illumination affect both the slope and intercept of the RoR–SpO_2 relationship. Despite these variations, the RoR features remain correlated with ground truth SpO_2 ($|\rho| = 0.77$–0.99) across all conditions, confirming the robustness of the physics-based approach. Participants provided written informed consent for the use of facial images in publication.

Supplementary Figure S5: Signal Extraction Detail

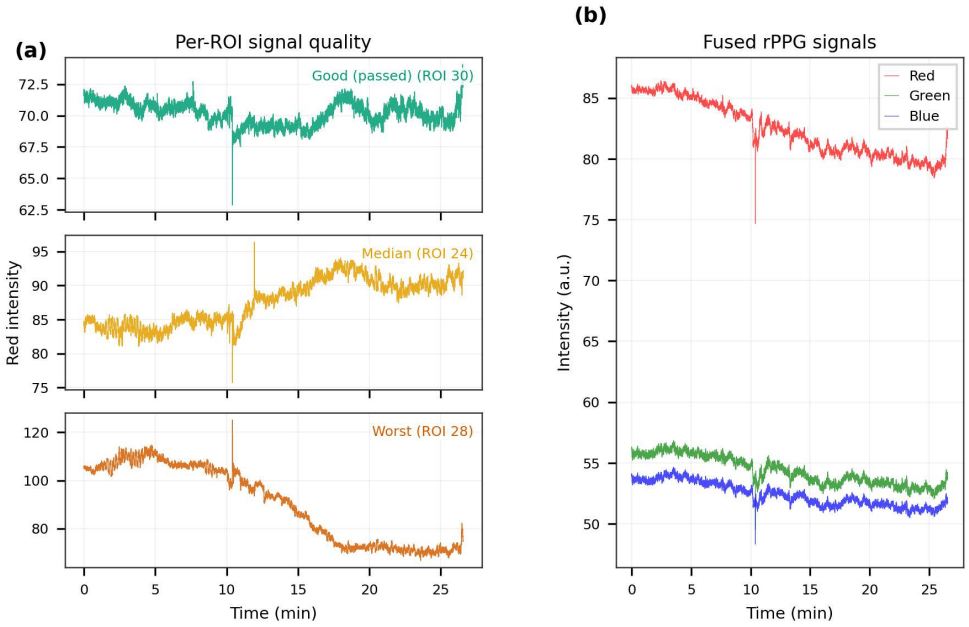


Fig. S5: **Signal extraction detail.** (a) Per-ROI signal quality: representative high-quality (top, green), median (middle, orange), and low-quality (bottom, red) ROIs from the Red channel, illustrating the signal fidelity variation across the superpixel grid. Quality screening excludes noisy ROIs from downstream processing. (b) SNR-weighted multi-ROI fusion showing fused Red, Green, and Blue channel signals after CHROM-based pulse extraction and SNR-weighted fusion.

Supplementary Table S1: Extended Per-Subject Performance Metrics

Table S1: Extended per-subject performance metrics

Subject	Skin Type	n	GT Range	ARMS (%)	MAE (%)	RoR PCC	Bias (%)	SD (%)
P201	VI	143	73–97	8.42	7.35	-0.130	-2.06	8.20
P202	III	143	71–98	10.79	9.40	-0.978	+8.93	6.07
P203	VI	102	83–98	10.13	8.08	-0.671	-8.01	6.23
P204	II	111	74–98	5.14	4.39	-0.831	-2.18	4.68
P206	II	152	71–98	5.63	4.75	-0.969	+2.13	5.23
P207	III	140	77–98	2.72	2.20	-0.966	+0.87	2.59
P208	III	125	76–98	5.57	4.62	-0.857	+3.30	4.50
P209	VI	137	73–99	10.59	9.42	-0.547	-0.34	10.63
P210	III	121	75–99	6.51	5.91	-0.953	-4.16	5.03
P211	III	140	70–97	5.01	4.15	-0.941	+2.89	4.10
P212	III	134	70–99	5.21	4.57	-0.859	-2.58	4.54
P213	IV	123	72–97	4.22	3.63	-0.980	+1.13	4.08
P214	III	85	83–98	4.00	3.08	-0.726	-0.28	4.02
P215	V	121	71–95	7.86	6.91	-0.962	-2.03	7.63
P216	II	134	71–95	6.40	5.64	-0.968	+5.35	3.53
P217	II	128	74–97	4.29	3.75	-0.898	+2.39	3.57
P218	III	123	76–99	6.37	5.25	-0.941	-4.24	4.78
P220	III	124	74–96	3.57	3.11	-0.946	-0.77	3.50
P221	II	127	74–96	4.70	4.36	-0.978	-1.56	4.45

A_{RMS}: accuracy root mean square per ISO 80601-2-61 ($= \sqrt{\text{bias}^2 + \text{SD}^2}$). RoR PCC: Pearson correlation between RoR_{RG} feature and reference SpO_2.

Supplementary Table S2: Bootstrap Confidence Intervals

Table S2: Bootstrap confidence intervals for aggregate metrics

Metric	CI Level	Estimate	CI Lower	CI Upper	SE	Method
ARMS	0.95	6.62	5.60	7.93	0.59	BCa_cluster
ARMS	0.99	6.62	5.31	8.34	0.59	BCa_cluster
MAE	0.95	5.34	4.51	6.42	0.48	BCa_cluster
MAE	0.99	5.34	4.28	6.78	0.48	BCa_cluster
PCC	0.95	0.595	0.373	0.732	0.087	BCa_cluster
PCC	0.99	0.595	0.296	0.770	0.087	BCa_cluster
AUC (90%)	0.95	0.797	0.669	0.870	0.047	BCa_cluster
AUC (90%)	0.99	0.797	0.609	0.887	0.047	BCa_cluster

All intervals computed using 10,000 BCa cluster-adjusted bootstrap resamples across 19 subjects.

Supplementary Table S3: TRIPOD+AI Compliance Checklist

This checklist documents compliance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis—Artificial Intelligence (TRIPOD+AI) reporting guideline. All 27 items and their sub-items are listed with manuscript section references. Items not applicable to this study are marked N/A with brief justification.

TRIPOD+AI Compliance Checklist

Item	Section	Checklist Item	Location
1	Title	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	Title
2	Abstract	See TRIPOD+AI for Abstracts checklist (13 sub-items covering background, objectives, setting, participants, outcome, predictors, sample size, missing data, analysis, results, conclusions, registration, funding)	Abstract
3a	Introduction	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model	Introduction, paragraphs 1–2
3b	Introduction	Describe the target population and the intended purpose of the prediction model in the context of the care pathway	Introduction, paragraphs 2–3
3c	Introduction	Describe any known health inequalities between sociodemographic groups	Introduction, paragraph 3 (skin type equity)
4	Introduction	Specify the study objectives, including whether the study describes the development or validation of a prediction model	Introduction, final paragraph (objectives)
5a	Methods	Describe the sources of data separately for the development and evaluation datasets	Methods > Study Design and Participants
5b	Methods	Specify the dates of the collected participant data, including start and end of participant accrual	Methods > Study Design and Participants

Continued on next page

Table continued from previous page

Item	Section	Checklist Item	Location
6a	Methods	Specify key elements of the study setting including the number and location of centres	Methods > Study Design and Participants
6b	Methods	Describe the eligibility criteria for study participants	Methods > Study Design and Participants
6c	Methods	Give details of any treatments received, and how they were handled during model development or evaluation	N/A — no therapeutic intervention; induced hypoxemia is the study protocol, not a treatment
7	Methods	Describe any data pre-processing and quality checking, including whether this was similar across sociodemographic groups	Methods > Signal Extraction Pipeline (quality filters, step removal, CHROM processing)
8a	Methods	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed	Methods > Reference Standard (SpO_2 from pulse oximeter)
8b	Methods	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of outcome assessors	N/A — SpO_2 reference is automated pulse oximeter reading, no subjective interpretation
8c	Methods	Report any actions to blind assessment of the outcome to be predicted	N/A — reference device operates independently of camera system
9a	Methods	Describe the choice of initial predictors and any pre-selection of predictors before model building	Methods > Signal Extraction Pipeline; Methods > Ratio-of-Ratios Feature Extraction
9b	Methods	Clearly define all predictors, including how and when they were measured	Methods > Ratio-of-Ratios Feature Extraction (RoR_{RB}, RoR_{RG} defined)
9c	Methods	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of assessors	N/A — RoR features are algorithmically computed, no subjective interpretation

Continued on next page

Table continued from previous page

Item	Section	Checklist Item	Location
10	Methods	Explain how the study size was arrived at and justify that it was sufficient to answer the research question	Methods > Sample Size
11	Methods	Describe how missing data were handled. Provide reasons for omitting any data	Methods > Study Design and Participants (exclusion criteria); Results > Participant Flow
12a	Methods	Describe how the data were used in the analysis, including whether the data were partitioned	Methods > Machine Learning Models (LOSOCV description)
12b	Methods	Depending on the type of model, describe how predictors were handled in the analyses	Methods > Machine Learning Models (two RoR features as direct inputs)
12c	Methods	Specify the type of model, rationale, all model building steps, and method for internal validation	Methods > Machine Learning Models; Methods > Model Development Transparency
12d	Methods	Describe if and how any heterogeneity in estimates across clusters was handled and quantified	Results > Skin Type Performance (Fitzpatrick stratification); Results > P202 Case Study
12e	Methods	Specify all measures and plots used to evaluate model performance	Methods > Statistical Analysis (A_{RMS}, PCC, AUC, Bland-Altman)
12f	Methods	Describe any model updating arising from the model evaluation	N/A — no model updating performed
12g	Methods	For model evaluation, describe how the model predictions were calculated	Methods > Machine Learning Models (LOSOCV prediction procedure)

Continued on next page

Table continued from previous page

Item	Section	Checklist Item	Location
13	Methods	If class imbalance methods were used, state why and how this was done	N/A — primary task is regression (continuous SpO_2); binary classification for hypoxemia uses natural prevalence with imbalance-aware evaluation metrics (AUC-ROC, sensitivity, specificity at optimal Youden threshold) rather than accuracy alone
14	Methods	Describe any approaches that were used to address model fairness and their rationale	Methods > Statistical Analysis (stratified analysis); Results > Skin Type Performance; Discussion > Equity
15	Methods	Specify the output of the prediction model and provide details and rationale for any classification	Methods > Machine Learning Models (continuous SpO_2 output); Methods > Statistical Analysis (90% threshold for hypoxemia classification)
16	Methods	Identify any differences between the development and evaluation data in healthcare setting and eligibility criteria	N/A — single dataset with leave-one-subject-out cross-validation
17	Methods	Name the institutional research board or ethics committee that approved the study	Ethics Approval
18a	Open Science	Give the source of funding and the role of the funders for the present study	Acknowledgements
18b	Open Science	Declare any conflicts of interest and financial disclosures for all authors	Competing Interests

Continued on next page

Table continued from previous page

Item	Section	Checklist Item	Location
18c	Open Science	Indicate where the study protocol can be accessed or state that a protocol was not prepared	Methods > Study Design and Participants (Clinimark protocol reference)
18d	Open Science	Provide registration information for the study	N/A — observational diagnostic accuracy study, not a registered clinical trial
18e	Open Science	Provide details of the availability of the study data	Data Availability
18f	Open Science	Provide details of the availability of the analytical code	Code Availability
19	PPI	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination	N/A — no patient or public involvement occurred
20a	Results	Describe the flow of participants through the study, including the number with and without the outcome	Results > Participant Characteristics; Supplementary Figure S1 (STARD flow diagram)
20b	Results	Report the characteristics overall and, where applicable, for each data source or setting	Results > Participant Characteristics (demographics table)
20c	Results	For model evaluation, show a comparison with the development data of the distribution of important predictors	N/A — single dataset, no separate development data
21	Results	Specify the number of participants and outcome events in each analysis	Results > Model Performance (participant counts reported)
22	Results	Provide details of the full prediction model to allow predictions in new individuals	Methods > Model Development Transparency (all hyperparameters, feature definitions, and model specifications reported)
23a	Results	Report model performance estimates with confidence intervals, including for any key subgroups	Results > Model Performance; Results > Skin Type Performance (A_{RMS} with 95% BCa CIs)

Continued on next page

Table continued from previous page

Item	Section	Checklist Item	Location
23b	Results	If examined, report results of any heterogeneity in model performance across clusters	Results > Skin Type Performance; Results > P202 Case Study
24	Results	Report the results from any model updating, including the updated model and subsequent performance	N/A — no model updating performed
25	Discussion	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives	Discussion > Summary; Discussion > Equity
26	Discussion	Discuss any limitations of the study and their effects on any biases, statistical uncertainty, and generalisability	Discussion > Limitations
27a	Discussion	Describe how poor quality or unavailable input data should be assessed and handled when implementing the model	Discussion > Limitations (signal quality, environmental constraints)
27b	Discussion	Specify whether users will be required to interact in the handling of the input data or use of the model	Discussion > Limitations (contactless monitoring context, no user interaction required for prediction)
27c	Discussion	Discuss any next steps for future research, with a specific view to applicability and generalisability	Discussion > Limitations, final paragraph (future directions); Conclusion

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf