Research Report:

Multimodal Approach by Shen.AI Using Remote Photoplethysmography (rPPG) and Ballistocardiography (rBCG) for Heart Rate and Heart Rate Variability Prediction

Abstract

Remote photoplethysmography (rPPG) and remote ballistocardiography (rBCG) are promising techniques for contactless cardiovascular monitoring. Both modalities can estimate heart rate (HR) and heart rate variability (HRV), but each has specific limitations: rPPG is sensitive to lighting and skin pigmentation, whereas rBCG is less affected by these factors but generally noisier. Shen.AI has developed a multimodal, signal-quality–driven approach that dynamically selects or combines modalities based on confidence metrics to optimize performance.

We evaluated this strategy in a large cohort of 5,311 participants spanning a broad demographic spectrum (mean age 53.8 years, 64.7% female). Data acquisition involved simultaneous recording of facial video (rPPG), micro-movement signals (rBCG), and pulse oximetry (ground truth). Each 60-second recording was segmented, and HR and HRV (SDNN) were computed for each modality. Quality scores were assigned to every window, enabling the best-of algorithm to select the modality with higher expected accuracy.

Results showed that rPPG achieved near-identity agreement with the reference for HR (MAE ≈ 0.37 bpm, R = 0.99), while rBCG exhibited higher error overall (MAE ≈ 3.6 bpm, R = 0.82) but markedly improved in high-quality windows. For SDNN, rPPG again outperformed rBCG (MAE ≈ 6 ms vs 36 ms), though variability estimates were inherently more error-prone than HR. Applying the best-of strategy yielded measurable gains: HR error was reduced by ~3% and SDNN by ~7% overall. Stratification by Fitzpatrick skin type highlighted equity-relevant patterns: rPPG error increased progressively with darker phototypes (III–VI), whereas rBCG remained stable. The selector accordingly shifted weight toward rBCG in these subgroups, mitigating disparities. At phototype VI, SDNN accuracy improved by ~18%, while HR error remained below 1.3 bpm.

These findings demonstrate that multimodal, quality-gated selection effectively leverages the complementary strengths of rPPG and rBCG, ensuring robust and equitable contactless measurement of HR and HRV across a large, diverse population. Future work should extend validation to additional HRV endpoints (e.g., RMSSD, frequency-domain measures), unconstrained real-world settings, and clinical cohorts with arrhythmias or impaired perfusion.

1. Introduction

Remote monitoring methods such as photoplethysmography (rPPG) and ballistocardiography (rBCG) are gaining popularity in telemedicine and consumer health applications. Their potential to assess heart rate (HR) and heart rate variability (HRV)—important indicators of autonomic nervous system activity and overall health—is particularly promising.

In this study, we investigate a multimodal approach developed by Shen.AI, which combines rPPG and rBCG signals to predict HR and HRV. The system dynamically evaluates the quality of each signal modality using dedicated quality metrics and selects the optimal input for final prediction. This signal-aware selection strategy aims to maximize accuracy in heterogeneous real-world conditions, such as varying lighting, motion, and diverse skin tones.

HR and HRV are clinically significant parameters used in a wide range of applications, including cardiovascular risk monitoring, mental stress and fatigue assessment, sleep quality evaluation, and early detection of arrhythmias. HRV in particular is a non-invasive biomarker of autonomic nervous system regulation and is associated with overall health, resilience to stress, and mortality risk. Continuous and reliable remote estimation of these metrics opens new opportunities for preventive care, chronic disease management, and digital therapeutics.

The objective of this study is to evaluate the effectiveness of the Shen.AI multimodal system for estimation of HR and HRV against the ground truth results obtained from pulsoxymetry measurements, across different subject demographics.

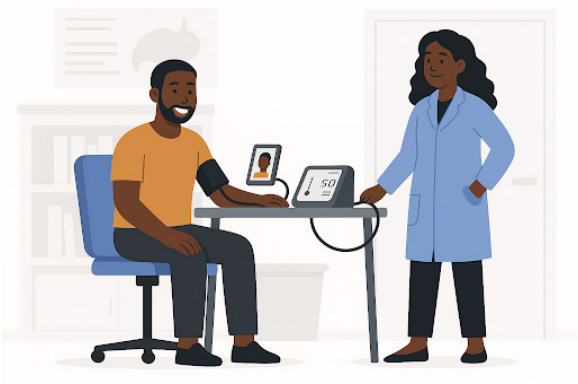
2. Data and Methods

2.1 Data set description

The data acquisition process in this study involved a custom data collection protocol, for which the team obtained approvals from the Bioethics Committee of the Medical University of Wrocław and the Lower Silesia Chamber of Physicians.

Recording of participants' facial videos and the collection of pulse oximetry-based reference measurements under controlled conditions were performed simultaneously. Recordings were conducted in three one-minute measurement sessions. A smartphone with a camera was mounted on a stable tripod equipped with an LED lamp, providing uniform and direct facial illumination while avoiding light reflections. Participants were seated with their feet flat on the floor and their forearm supported on a table or thigh, maintaining their face centrally aligned in the camera frame and positioned perpendicularly to the lens. During recording, participants remained motionless, refrained from facial expressions or speaking, and ensured that the skin of the face was fully exposed.

Throughout the measurements, signals from a Bluetooth-connected pulse oximeter (Berry BM1000C) were recorded simultaneously with video to confirm heart rate values and interbeat intervals. The intervals were calculated as a time difference between consecutive peaks in photoplethysmographic signal.



Participant metadata were collected, including age, sex, height, weight, Fitzpatrick scale of the skin tone and information regarding hypertension, diabetes, smoking, arrhythmia, or anemia.

3 types of information were extracted separately from both PPG and rBCG signals, during a 60-seconds continuous measurement:

- Inter-beat (IBI) intervals, expressed in milliseconds
- Heart Rate (or Pulse Rate), expressed as number of beats per minute
- HRV-SDNN, calculated from previously extracted IBI data.

2.2 Exclusion and inclusion criteria.

Inclusion criteria

- Open enrollment: Any individual from the general population, irrespective of age, sex, or skin phototype, was eligible to take part.
- Willingness to participate: Participants had to declare their willingness to join the study and follow basic instructions (e.g., remain seated and face the camera during brief recordings).
- Informed consent: Participants were required to read the study information and sign written informed consent prior to any procedures (for minors, parent/guardian consent with participant assent, where applicable).

Exclusion criteria

- Lack of informed consent for the measurement procedure.
- Significant facial anatomical deformity, e.g., due to neoplastic lesions, trauma, or other structural pathologies that may interfere with signal acquisition
- Inability to maintain stable head positioning during the measurement
- Respiratory dysfunction, such as dyspnea, irregular breathing patterns, or shallow respiration, which may compromise measurement quality or participant safety
- Presence of an implanted cardiac pacemaker, due to potential interference with the measurement process and associated safety concerns
- Requirement for continuous medical supervision, such as in cases of severe chronic illness requiring constant clinical oversight.
- Acute life- or health-threatening conditions, including but not limited to acute coronary syndromes, respiratory failure, shock, or major trauma
- Extensive pathological processes affecting the face, which could impair signal detection or introduce significant measurement artifacts
- Large facial dressings or bandages that obstruct facial features or interfere with optical or visual signal acquisition
- Extensive facial tattoos or permanent facial makeup, which may affect the accuracy of optical measurement methods
- Marked or persistent facial pallor, or cardiovascular/respiratory disorders such as heart failure, left ventricular systolic dysfunction, aortic valve stenosis, or other structural or functional abnormalities of the heart or respiratory system that may result in low stroke volume, low blood pressure amplitude, or the presence of pulsus paradoxus.

2.2 Subject Demographics and Characteristics of the study population

The analytic sample comprised N = 5,311 participants (64.7% women, n = 3,436; 35.3% men, n = 1,875). The mean age was 53.8 ± 18.4 years (median 57.0; interquartile range [IQR] 31.0; range 10–99). Mean body weight was 72.7 ± 14.9 kg (median 70.0; IQR 20.0; range 30–140), and mean height was 166.5 ± 9.5 cm (median 166.0; IQR 13.0; range 125–200). The resulting BMI averaged 26.2 ± 4.8 kg/m² (median 25.6; IQR 6.3; range 14–49). Values are reported as mean ± SD unless otherwise specified.

Fitzpatrick classification was available for 5,306/5,311 participants (99.9%). The distribution was highly skewed toward mid-range phototypes: Type IV was most common (3,068; 57.8%), followed by Type III (1,465; 27.6%). Darker phototypes accounted for 14.4% of the cohort (Type V: 593; 11.2% and Type VI: 169; 3.2%). Types I and II were incidental, with only 7 (0.1%) and 4 (0.1%) participants, respectively.

Within-type sex composition showed a female predominance across most categories (female share within type: I 71%, II 100%, III 75%, IV 62%, V 57%), while Type VI was the only group with a male majority (47% female / 53% male).

Age distributions by phototype (medians from the boxplots) indicated that participants with Types V and VI tended to be younger than those with Types III–IV: median age was: I 46.0 years, II 41.5 years, III 62.0 years, IV 57.0 years, V 46.0 years, VI 50.0 years. Thus, compared with the two dominant groups (III–IV), the darker-skin groups (V–VI) were under-represented but skewed toward lower median age.

In sum, the cohort is overwhelmingly Type IV and III (85.4%), with Types I–II rare and Types V–VI present but less frequent and younger on average. This distribution should be considered when interpreting modality-specific performance across skin tones and age strata.

The analytic sample comprised N = 5,311 participants (64.7% women, n = 3,436; 35.3% men, n = 1,875). The mean age was 53.8 ± 18.4 years (median 57.0; interquartile range [IQR] 31.0; range 10–99). Mean body weight was 72.7 ± 14.9 kg (median 70.0; IQR 20.0; range 30–140), and mean height was 166.5 ± 9.5 cm (median 166.0; IQR 13.0; range 125–200). The resulting BMI averaged 26.2 ± 4.8 kg/m² (median 25.6; IQR 6.3; range 14–49). Values are reported as mean ± SD unless otherwise specified.

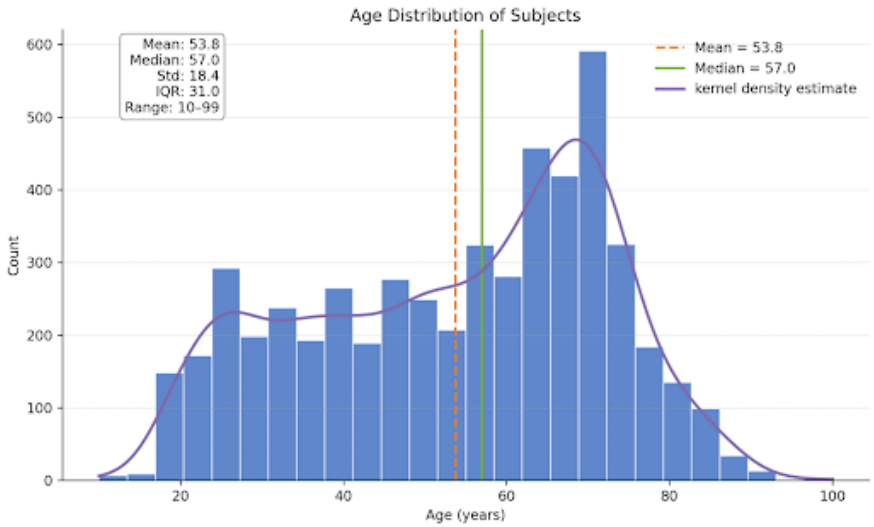
Fitzpatrick classification was available for 5,306/5,311 participants (99.9%). The distribution was highly skewed toward mid-range phototypes: Type IV was most common (3,068; 57.8%), followed by Type III (1,465; 27.6%). Darker phototypes accounted for 14.4% of the cohort (Type V: 593; 11.2% and Type VI: 169; 3.2%). Types I and II were incidental, with only 7 (0.1%) and 4 (0.1%) participants, respectively.

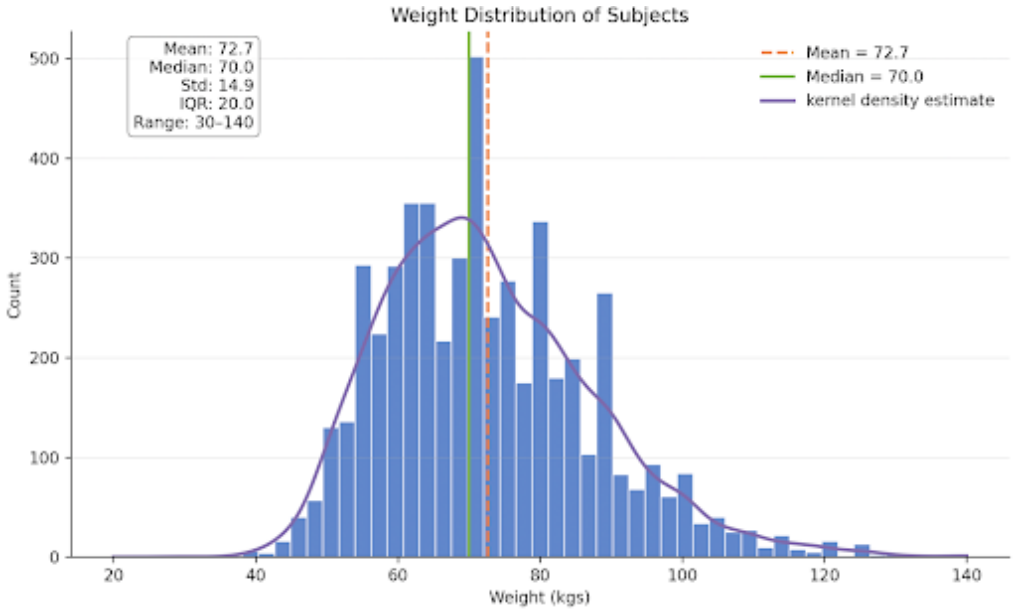
Within-type sex composition showed a female predominance across most categories (female share within type: I 71%, II 100%, III 75%, IV 62%, V 57%), while Type VI was the only group with a male majority (47% female / 53% male).

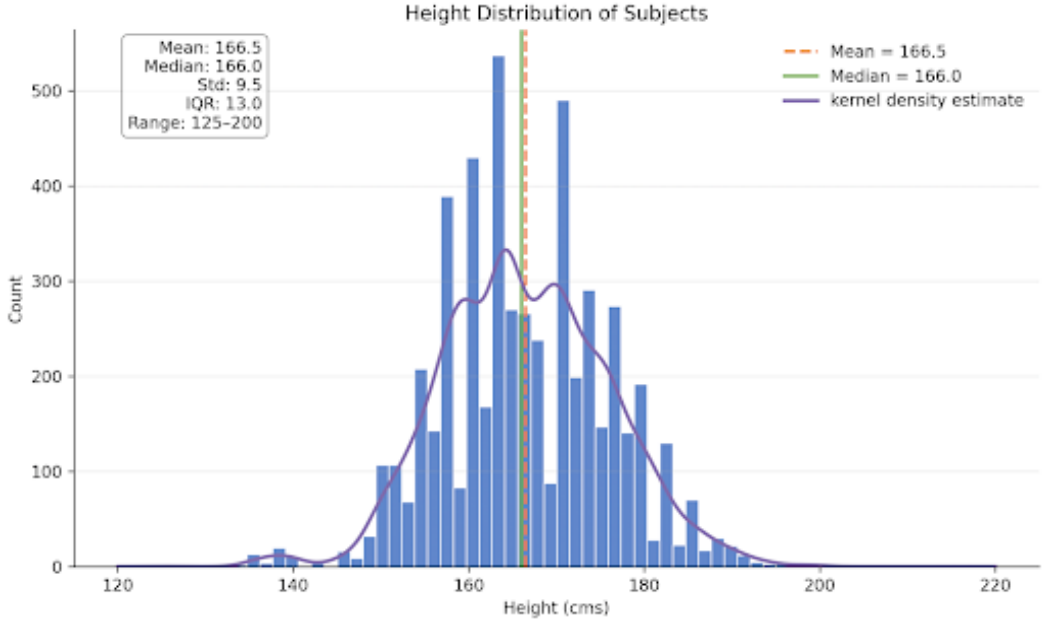
Age distributions by phototype (medians from the boxplots) indicated that participants with Types V and VI tended to be younger than those with Types III–IV: median age was: I 46.0 years, II 41.5 years, III 62.0 years, IV 57.0 years, V 46.0 years, VI 50.0 years. Thus, compared with the two dominant groups (III–IV), the darker-skin groups (V–VI) were under-represented but skewed toward lower median age.

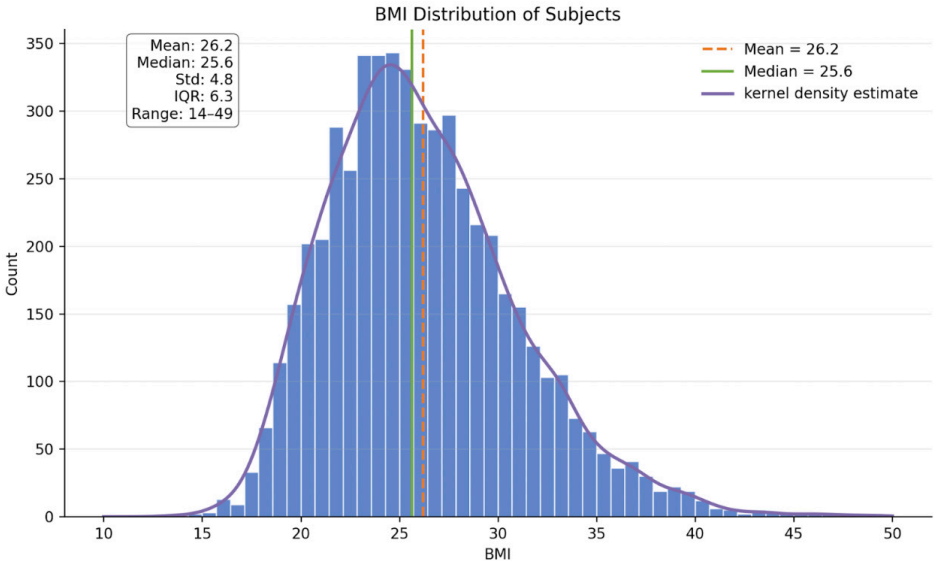
In sum, the cohort is overwhelmingly Type IV and III (85.4%), with Types I–II rare and Types V–VI present but less frequent and younger on average. This distribution should be considered when interpreting modality-specific performance across skin tones and age strata.

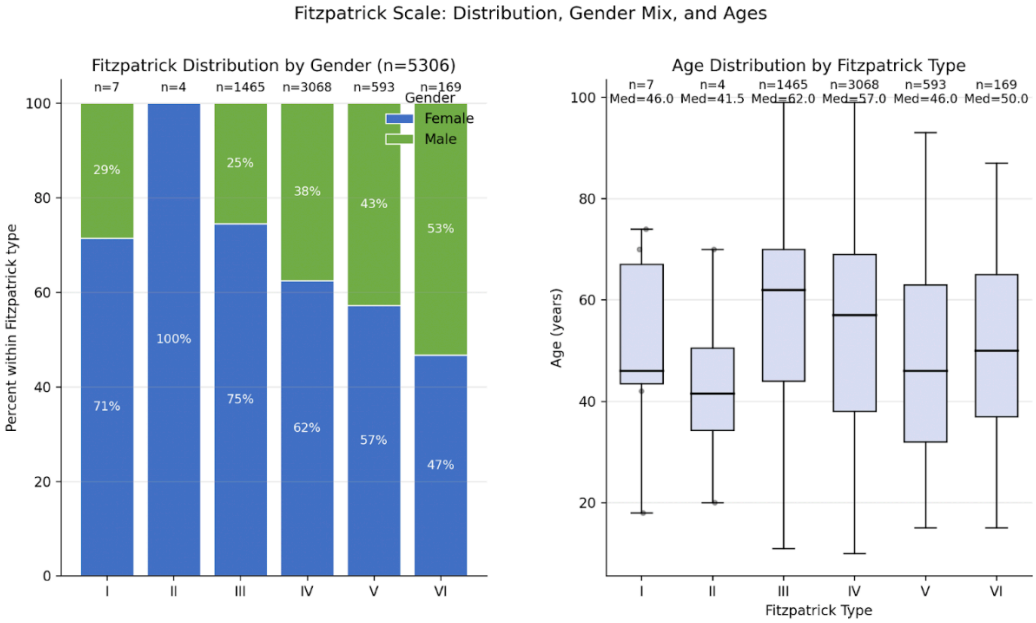
Figure 1. Age distribution of the study population
Figure 2. Gender distribution of the study population
Figure 3. Weight distribution of the study population
Figure 4. Height distribution of the study population
Figure 5. Body mass index (BMI) distribution of the study population
Figure 6. Distribution of skin tones by Fitzpatrick phototype in the study population.


Age Distribution of Subjects
Mean: 53.8
Median: 57.0
Std: 18.4
IQR: 31.0
Range: 10–99
Mean = 53.8
Median = 57.0
kernel density estimate
Count
Age (years)


Weight Distribution of Subjects
Mean: 72.7
Median: 70.0
Std: 14.9
IQR: 20.0
Range: 30–140
Mean = 72.7
Median = 70.0
kernel density estimate
Count
Weight (kgs)


Height Distribution of Subjects
Mean: 166.5
Median: 166.0
Std: 9.5
IQR: 13.0
Range: 125–200
Mean = 166.5
Median = 166.0
kernel density estimate
Count
Height (cms)





2.4 Signal Acquisition and Preprocessing

The PPG-based dataset was collected using a pulse oximeter that continuously measured blood-volume changes, paired with synchronized video recording. Using a heuristic peak-detection algorithm, we can count pulsations in real time, enabling precise estimation of metrics such as BPM (beat per minute) and SDNN. An example PPG waveform with detected peaks is shown in Figure 7.

2.5 Signal-Quality–Driven Modality Selection (Best-of) Strategy

A signal-quality–dependent modality selection ("best-of") strategy was implemented, which chooses the prediction based on quality metrics. The rPPG signal is obtained directly from facial pulsations: video frames captured by the camera are first processed for facial-texture detection, after which a dedicated neural network reconstructs the pulse--wave signal. Similarly, rBCG also reconstructs the pulse wave, but it relies on facial micro-movement information and a separate neural network.

For both modalities, quality metrics were developed to indicate whether the confidence level of the predicted signal is sufficient to return a result and which modality should be selected for the final estimate. The quality-assessment procedure is deterministic and based on well-established digital signal processing methods: pulse-wave amplitudes and their variability are evaluated to determine the confidence

of each peak, enabling assignment of a percentage quality score to every single beat in the inferred pulse wave. In this study, depending on the values of the quality metrics, the entire rPPG or rBCG signal is discarded; only the modality that meets the quality criterion is used for the final estimate (without combining/fusing results).
In summary, the rPPG/rBCG approach based on automatic modality switching enables extraction of a reliable pulse-wave signal from facial recordings under varied environmental conditions, because the system dynamically selects the highest-quality modality according to precisely defined metrics.

Figure 8A and Figure 8B present two error surfaces—one for rPPG and one for rBCG—mapped over their respective quality metrics. Both surfaces exhibit the expected monotonic trend: as a modality's quality increases, its mean SDNN or HR error decreases. The two surfaces are strongly correlated and intersect along a ridge that acts as a data-driven decision boundary: on one side, rBCG is predicted to yield the lower SDNN or HR error; on the other, rPPG is favored. This boundary operationalizes the "best-of" selection logic by directing the algorithm to choose the modality with the lower expected error for the final estimate. Operationally, the quality plane can be divided into four regimes: (i) High rPPG / low rBCG quality → rPPG dominates; (ii) High rBCG / low rPPG quality → rBCG dominates; (iii) Both high → both modalities achieve low error and the boundary passes near the line of equality, so either choice performs well; and (iv) Both low → both errors are high, suggesting that no-return or re-measurement criteria may be appropriate. In the complementary "difference-surface" view, the surface lies at approximately 45° to the quality axes, indicating that—after normalization—the two quality measures contribute comparably to error discrimination; a simple linear decision rule (e.g., a weighted difference of normalized qualities) is therefore well-justified. These plots provide construct validity for the selection strategy: (a) the consistent inverse relationship between quality and error supports the use of deterministic, signal-processing–based quality metrics; (b) the clear intersection ridge yields a stable, interpretable decision boundary rather than an ad-hoc threshold; and (c) the geometry of the difference surface explains why the "best-of" selector improves aggregate SDNN and HR accuracy

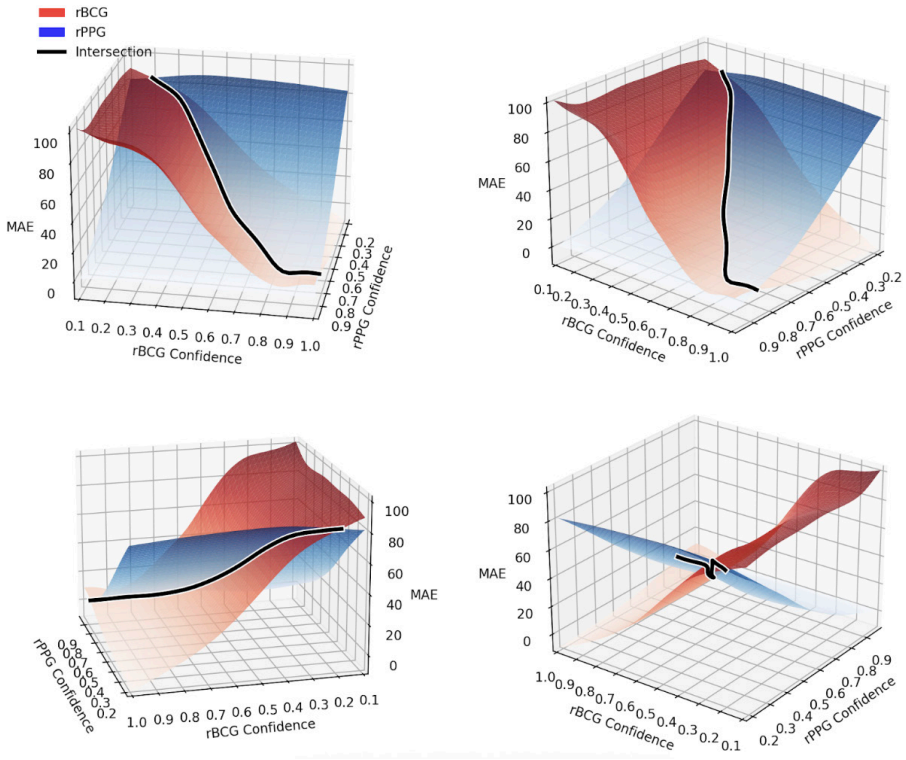


Figure 8A. shows the relationship between the rBCG and rPPG quality metrics and their mean SDNN errors. The two error surfaces are clearly correlated and intersect along a curve that serves as a decision boundary: on one side the rBCG modality is expected to yield the lower SDNN error, and on the other side rPPG; the algorithm selects the corresponding modality for the final estimate.

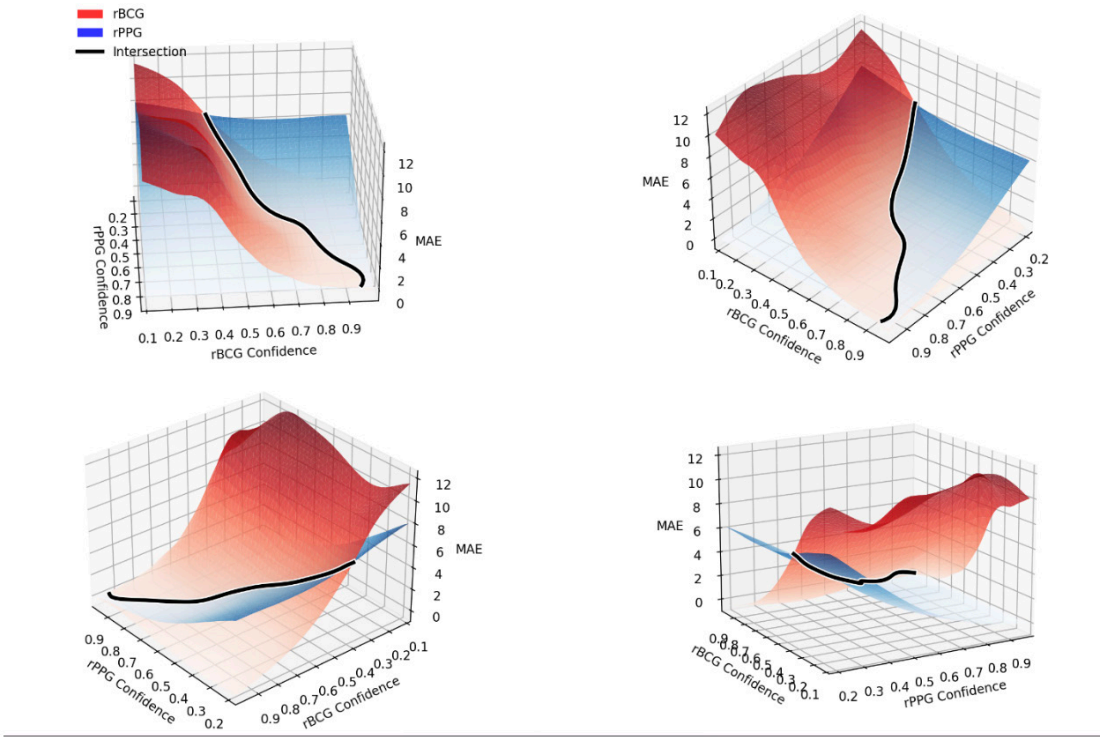


Figure 8B. shows the relationship between the rBCG and rPPG quality metrics and their mean HR errors. The two error surfaces are clearly correlated and intersect along a curve that serves as a decision boundary: on one side the rBCG modality is expected to yield the lower HR error, and on the other side rPPG; the algorithm selects the corresponding modality for the final estimate.

Figure 9A and Figure 9B show an equivalent representation in which a single surface maps the difference in errors between the two modalities onto the quality plane (Q_{rPPG},Q_{rBCG}), e.g. , E_{rPPG}–E_{rBCGE}

- The sign is immediately decision-relevant: negative values (under this definition) indicate rPPG has the lower error; positive values indicate rBCG is more accurate. The absolute magnitude $|E_{rPPG}–E_{rBCG}|$ can be read as a confidence measure: the larger it is, the clearer the advantage of one modality.
- The surface lies at roughly 45° to the quality axes. After normalizing the quality metrics, this means both quality measures contribute comparably to discriminating which modality yields the lower error. Consequently, a simple linear decision rule—such as comparing normalized qualities or using a weighted difference—is well-justified.

Algorithmically:

- The decision boundary (where the error difference ≈ 0) runs close to the line of equal quality Q_{rPPG}≈Q_{rBCGQ}. On one side the selector should choose rPPG; on the other, rBCG.
- When both qualities are high, the difference in errors is typically small (both methods perform well), so the choice is low-risk.
- When both qualities are low, both errors tend to be high and the difference may be unstable; this is a natural region for no-return / re-measurement policies.
- In practice, introducing a small margin δ around the boundary (i.e., abstain when $|E_{rPPG}–E_{rBCG}|<\delta|$ reduces misclassifications in the uncertainty zone.

Practical implications.
(1) A simple linear selector is adequate because the two quality metrics are similarly informative.
(2) A confidence score can be derived directly from $|E_{rPPG}–E_{rBCG}|$.
(3) Measurement policy (repeat/adjust conditions) should trigger when both qualities are low or the error difference is small, i.e., near the decision boundary.

Legend:

- Q_{rPPG},Q_{rBCG} - normalized quality metrics (scale 0–1 or 0–100).
- E_{rPPG},E_{rBCGE} - SDNN error [ms].
- $\Delta E=E_{rPPG}–E_{rBCG}$ error-difference surface;
  $\Delta E<0 \Rightarrow$ rPPG has lower error,
  $\Delta E>0 \Rightarrow$ rBCG has lower error.
- Decision boundary: $\Delta E=0$

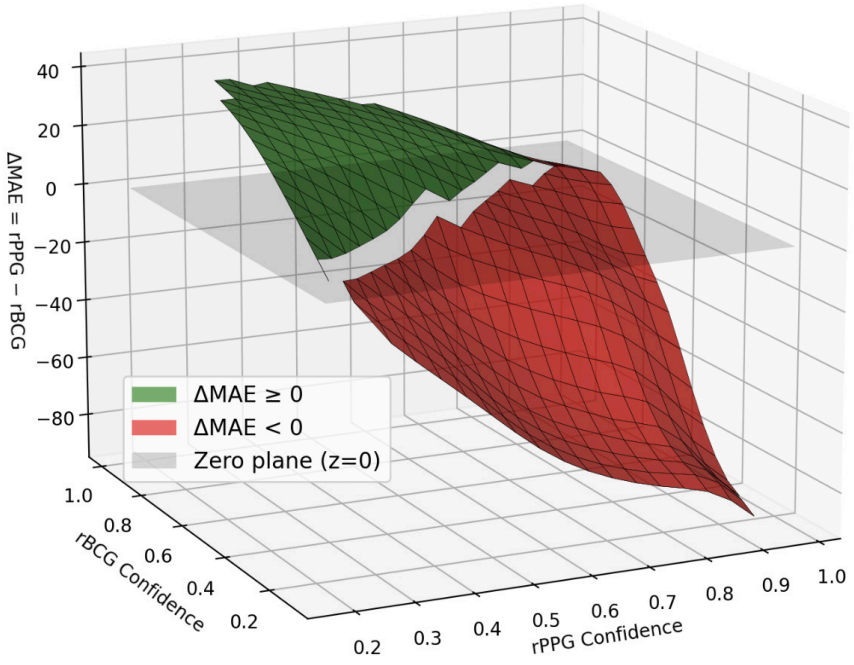


Figure 9A presents an equivalent view in which the surface represents the difference between the two errors; in this representation, the surface lies at approximately 45° to the quality plane, supporting the conclusion that the quality measures are informative indicators of the lower-error modality. This plot is shown for the SDNN measure of HRV.

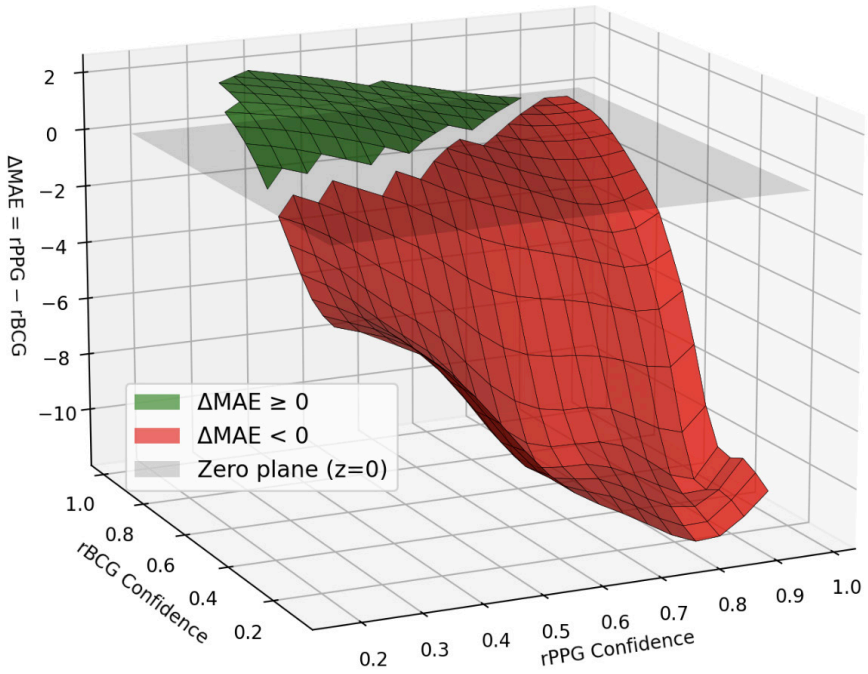


Figure 9A presents an equivalent view in which the surface represents the difference between the two errors; in this representation, the surface lies at approximately 45° to the quality plane, supporting the conclusion that the quality measures are informative indicators of the lower-error modality. This plot is shown for the HR.

2.6 Statistical analysis

3. Results

3.1 rPPG: Agreement with Ground Truth and Dependence on Quality (HR and SDNN)

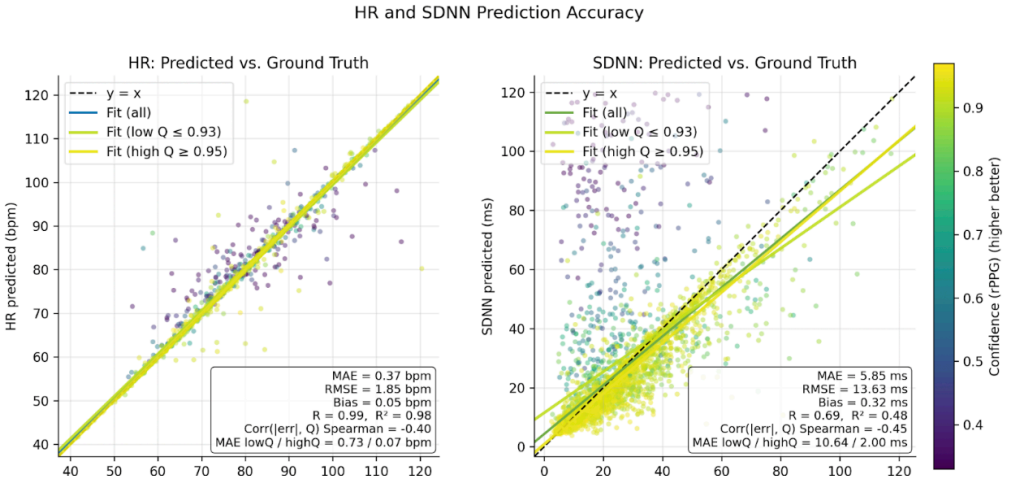


Figure 10. Heart rate and heart rate variability (SDNN) measured by rPPG: predicted vs. ground truth, colored by confidence; fits for all, low Q, and high Q.

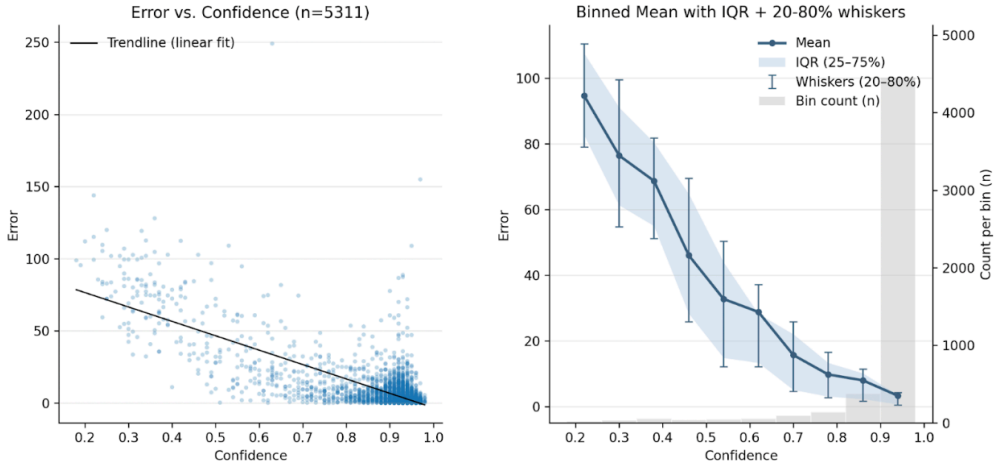


Figure 11A. Heart rate variability (SDNN) measured by rPPG: Error vs. confidence: scatter with linear trend and Error vs. confidence: binned mean with IQR and 20–80% whiskers (with bin counts).

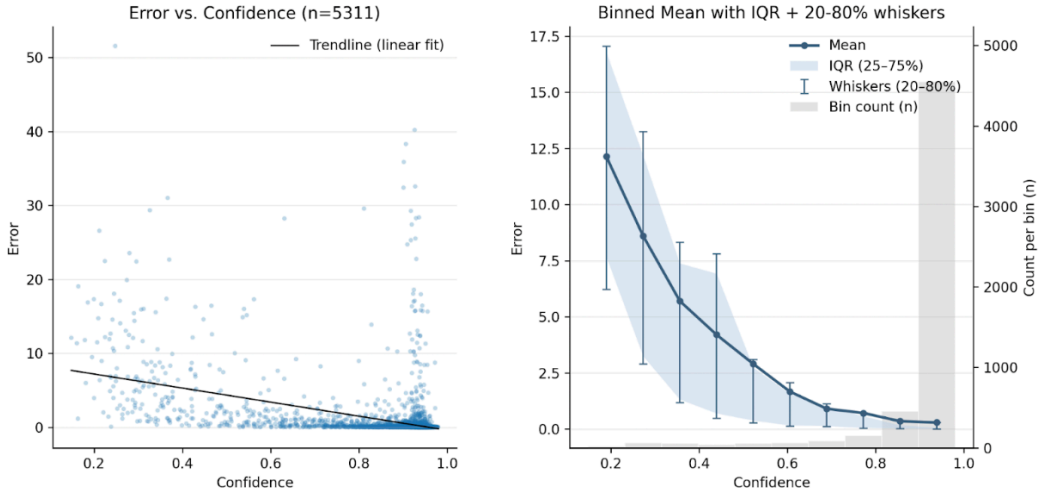


Figure 11B. Heart rate measured by rPPG: Error vs. confidence: scatter with linear trend and Error vs. confidence: binned mean with IQR and 20–80% whiskers (with bin counts).

For heart rate (HR), rPPG predictions showed excellent agreement with the reference, with MAE 0.37 bpm, RMSE 1.85 bpm, and negligible bias 0.05 bpm; the correlation was R = 0.99 (R^2 = 0.98). For SDNN, accuracy was lower but remained clinically useful (MAE 5.85 ms, RMSE 13.63 ms, bias 0.32 ms; R = 0.69, R^2 = 0.48). Points colored by the rPPG quality score cluster around the identity line at high quality and disperse at lower quality. The absolute error demonstrated a consistent negative correlation with quality (Spearman ρ ≈ −0.40 for HR; −0.45 for SDNN). When analyses were restricted to high-quality windows (e.g., Q⩾0.95Q \ge 0.95Q⩾0.95), error decreased markedly (MAE HR: 0.07 bpm vs 0.73 bpm at low Q; MAE SDNN: 2.00 ms vs 10.64 ms). Complementary "error vs confidence" plots (n = 5,311) showed a monotonic decline of the mean error with increasing confidence and progressive narrowing of dispersion (IQR, whiskers), supporting the use of quality thresholds or confidence flags in downstream applications.

3.2 rBCG: Agreement with Ground Truth and Dependence on Quality (HR and SDNN)

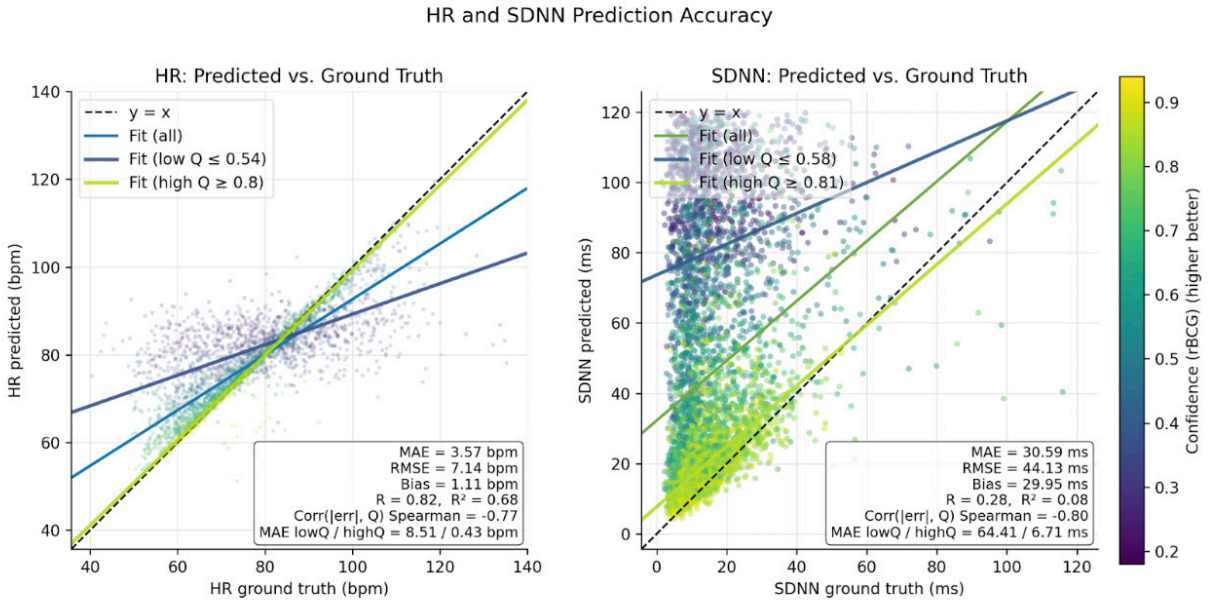


Figure 12. Heart rate and heart rate variability (SDNN) measured by rBPG: predicted vs. ground truth, colored by confidence; fits for all, low Q, and high Q.

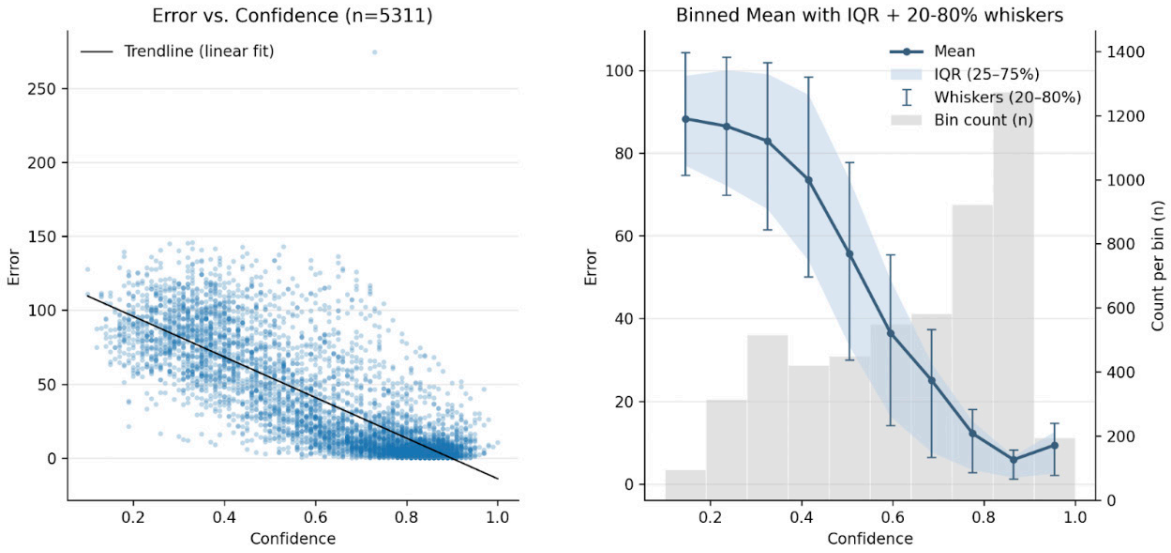


Figure 13A. Heart rate variability (SDNN) measured by rBPG: Error vs. confidence: scatter with linear trend and Error vs. confidence: binned mean with IQR and 20–80% whiskers (with bin counts).

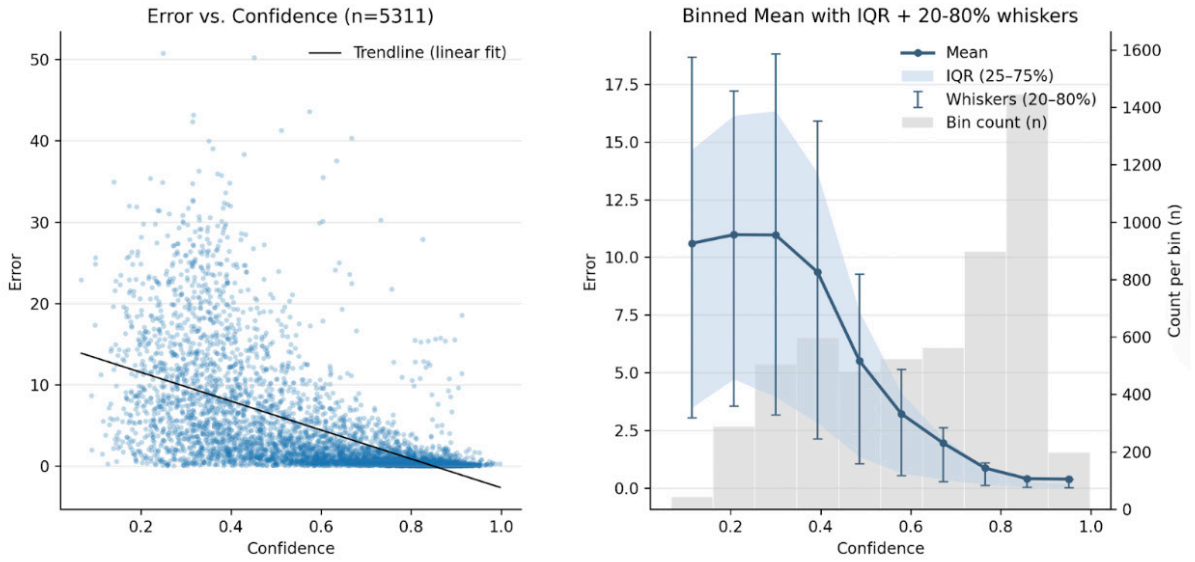


Figure 13B. Heart rate variability (SDNN) measured by rBPG: Error vs. confidence: scatter with linear trend and Error vs. confidence: binned mean with IQR and 20–80% whiskers (with bin counts).

Aggregate rBCG accuracy was lower than rPPG, particularly at low quality. For HR, overall MAE was 3.57 bpm (RMSE 7.14 bpm, bias 1.11 bpm; R = 0.82, R^2 = 0.68). For SDNN, overall MAE was 30.59 ms (RMSE 44.13 ms, bias 29.95 ms; R = 0.28, R^2 = 0.08). Nevertheless, rBCG performance improved sharply at high quality: MAE HR fell from 8.51 bpm (low Q) to 0.43 bpm (high Q), and MAE SDNN from 64.41 ms to 6.71 ms. The error–quality relationship was strongly negatively correlated (Spearman ρ ≈ −0.77 for HR; −0.80 for SDNN). The corresponding "error vs confidence" plots again showed a near-linear decrease of mean error with increasing confidence and reduced dispersion at higher QQQ, indicating that quality-gated rBCG can yield reliable estimates in a substantial fraction of windows despite weaker aggregate performance. These results provide the empirical basis for the subsequent best-of (quality-driven) modality selection, whereby rBCG is preferentially used in conditions where its confidence is high.

3.3. Confidence, error, and outcome of the quality-driven selection

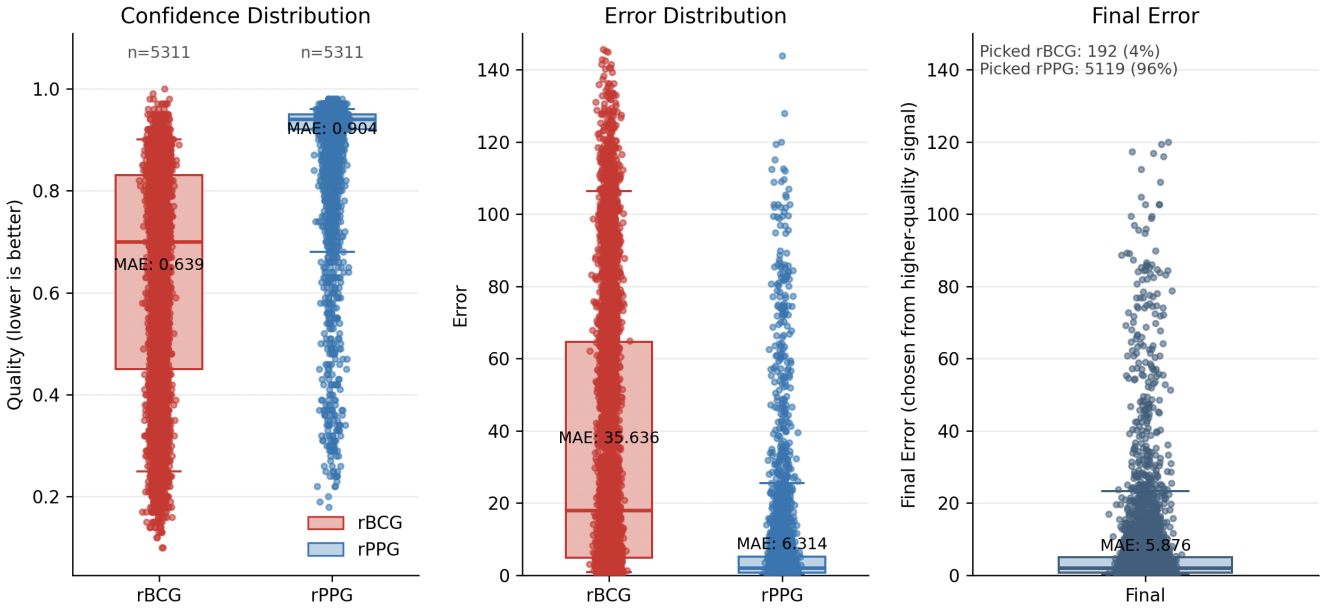


Figure 14A. The distributions of SDNN errors and quality scores for each modality (rPPG and rBCG). Applying the optimal-selection algorithm improves the final SDNN prediction accuracy by approximately 7%.

For SDNN, the confidence distributions differ markedly between modalities: rPPG exhibits a high central tendency (median around 0.90), whereas rBCG is centered lower (median around 0.64). The SDNN error distributions mirror this pattern: rPPG MAE = 6.314 ms, markedly below rBCG MAE = 35.636 ms. Applying the best-of (quality-driven) selector over all windows results in rPPG being chosen in 5,119/5,311 cases (96%) and rBCG in 192/5,311 (4%), yielding a final MAE of 5.876 ms. These summaries indicate that, in aggregate, rPPG attains both higher quality and lower SDNN error, while rBCG contributes selectively in a minority of cases where its quality is sufficient.

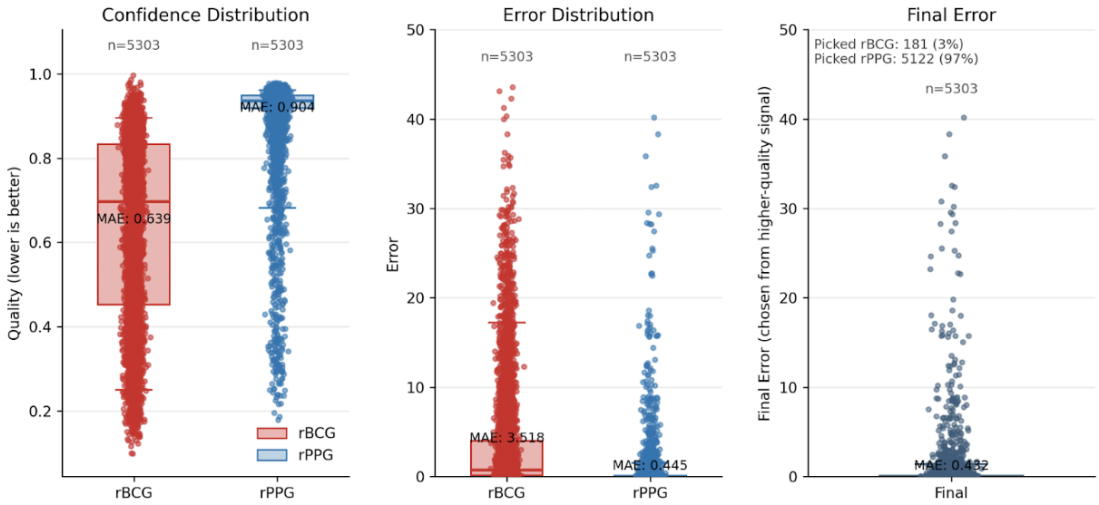


Figure 14B. The distributions of HR errors and quality scores for each modality (rPPG and rBCG). Applying the optimal-selection algorithm improves the final HR prediction accuracy by approximately 3%.

For Heart Rate (HR), the confidence distributions differ between modalities: rPPG exhibits a higher central tendency (median around 0.90), whereas rBCG is centered lower (median around 0.64). The HR error distributions follow the same pattern: rPPG MAE = 0.445 bpm, which is substantially lower than rBCG MAE = 3.518 bpm. Applying the best-of (quality-driven) selector over all windows results in rPPG being chosen in 5,122/5,303 cases (97%) and rBCG in 181/5,303 cases (3%), yielding a final MAE of 0.432 bpm. These results indicate that, overall, rPPG achieves both higher quality and lower HR error, while rBCG contributes selectively in rare cases where its signal quality is comparatively better.

3.4. Modality-specific SDNN and HR error by Fitzpatrick phototype (III–VI)

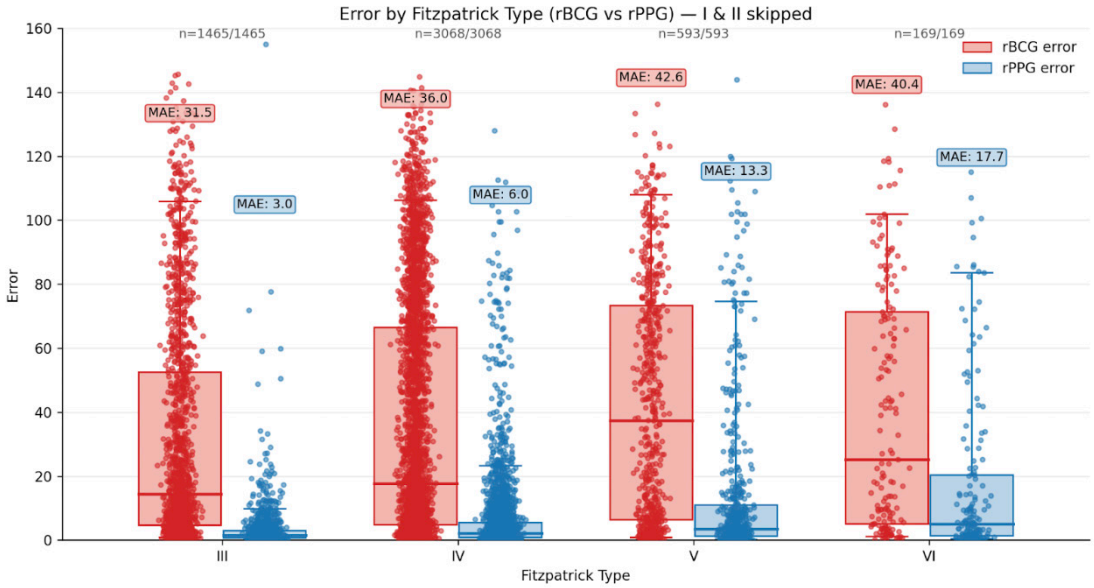


Figure 15A . Modality-specific SDNN error distributions stratified by Fitzpatrick skin type (III–VI). Two findings are evident: (1) rPPG error increases with higher Fitzpatrick types (darker skin), whereas rBCG error shows little dependence on skin type; (2) consequently, the modality-selection algorithm yields larger gains at higher Fitzpatrick types by favoring rBCG. At type VI, the improvement is approximately 18%, compared with an average improvement of about 7% across the cohort.

Stratification by skin phototype shows a monotonic increase of rPPG error with darker phototypes, while rBCG error remains relatively flat. Reported MAE (ms) by type:

- Type III: rPPG 3.0, rBCG 31.5
- Type IV: rPPG 6.0, rBCG 36.0
- Type V: rPPG 13.3, rBCG 42.6
- Type VI: rPPG 17.7, rBCG 40.4

These data confirm pigmentation-related degradation for rPPG and relative invariance for rBCG, supporting the rationale for modality selection conditioned on quality

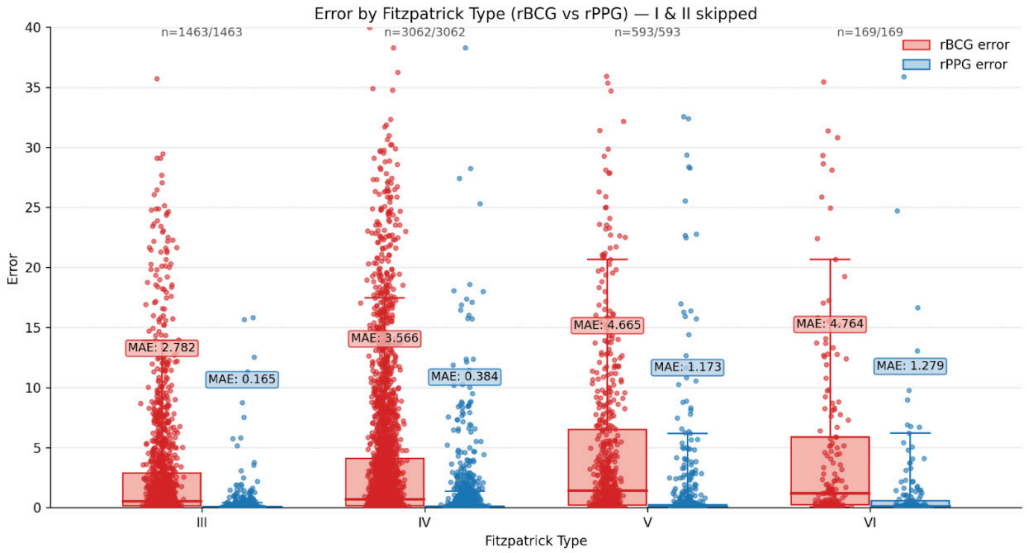


Figure 15B. Modality-specific HR error distributions stratified by Fitzpatrick skin type (III–VI). Two findings are evident: (1) rPPG error increases monotonically with darker skin tones, whereas rBCG error remains comparatively stable across skin types; (2) as a consequence, the modality-selection algorithm is expected to yield larger benefits in darker skin tones by favoring rBCG whenever its quality surpasses rPPG.
Stratification by skin phototype shows this clear trend, with mean absolute error (MAE, bpm) by type:

- Type III: rPPG = 0.165, rBCG = 2.782
- Type IV: rPPG = 0.384, rBCG = 3.566
- Type V: rPPG = 1.173, rBCG = 4.665
- Type VI: rPPG = 1.279, rBCG = 4.764

These data confirm pigmentation-related degradation for rPPG in HR estimation, while rBCG remains relatively invariant across skin types. This pattern supports the rationale for applying quality-conditioned multimodal fusion to maintain accuracy in diverse populations.

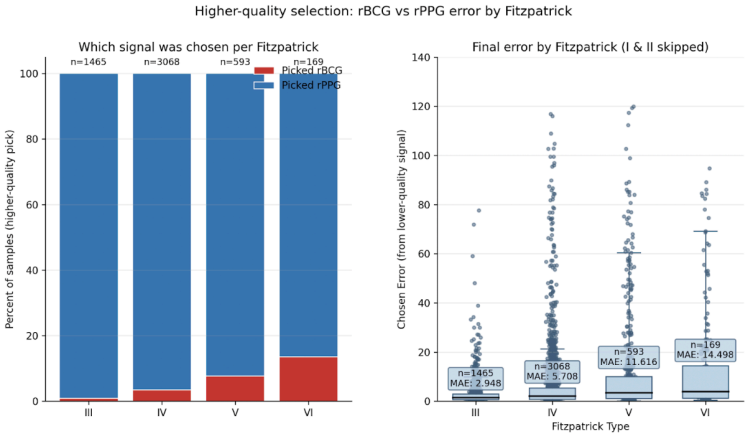


Figure 16A. Selected modality and resulting SDNN error by Fitzpatrick phototype

Across phototypes III–VI, the selection mechanism chooses rPPG in the majority of windows, with the share of rBCG selections increasing at higher phototypes. The final SDNN error (MAE, ms) after selection rises with phototype: 2.9 (III), 5.7 (IV), 11.6 (V), 14.5 (VI). Thus, although overall performance declines with darker skin (driven by rPPG sensitivity), quality-gated selection maintains practical accuracy and limits error growth by deferring to rBCG where appropriate.

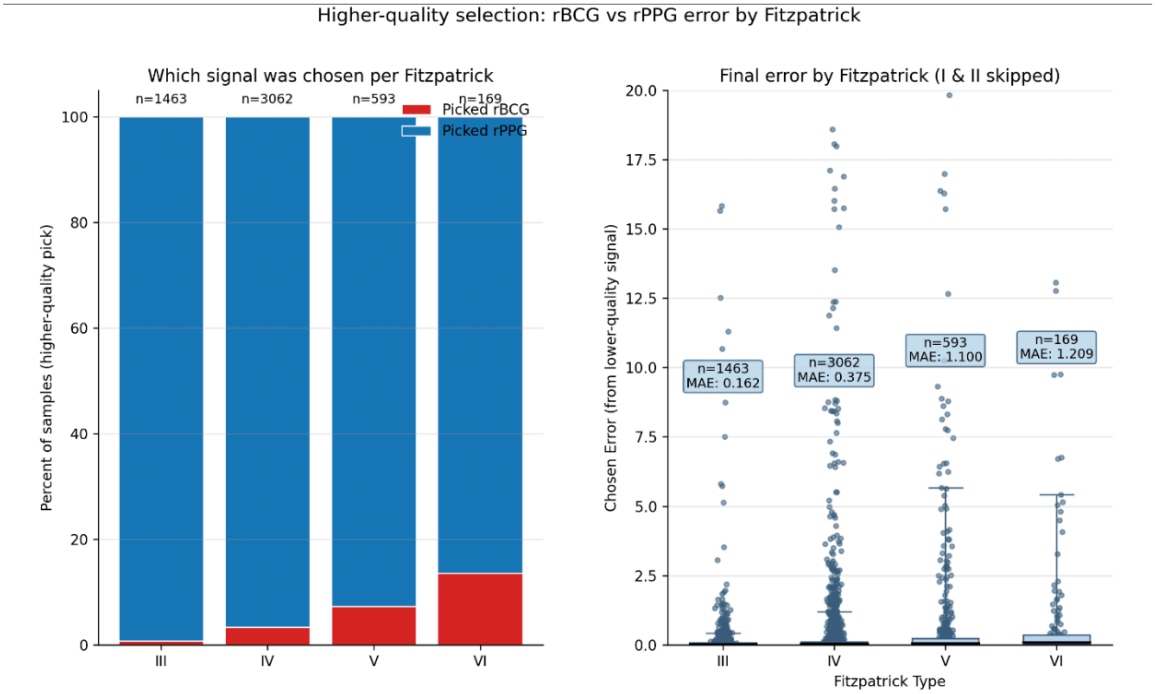


Figure 16B. Selected modality and resulting HR error by Fitzpatrick phototype

Across phototypes III–VI, the selection mechanism chooses rPPG in the vast majority of windows, with the share of rBCG selections gradually increasing at higher phototypes. The final HR error (MAE, bpm) after selection also rises with phototype: 0.162 (III), 0.375 (IV), 1.100 (V), 1.209 (VI). Thus, although overall performance declines slightly with darker skin (driven mainly by rPPG sensitivity), quality-gated selection maintains practical accuracy and limits error growth by deferring to rBCG where appropriate.

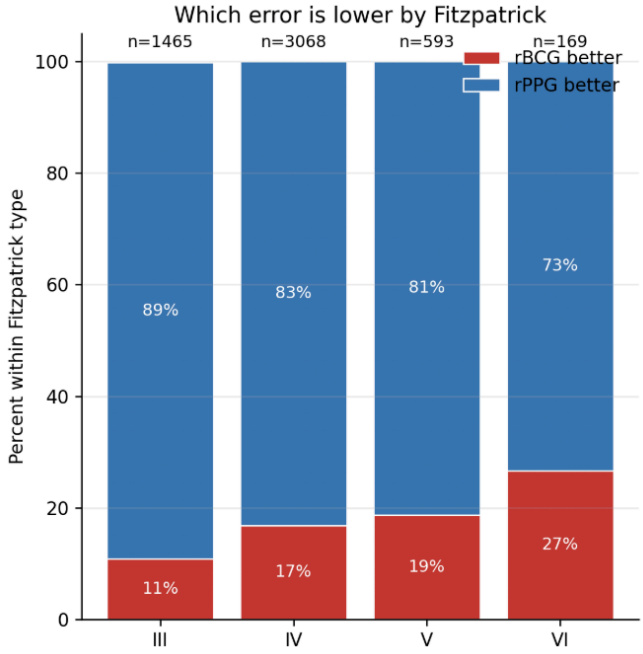


Figure 17. Proportion of windows with lower error by modality (by phototype)

Within each phototype, the proportion of windows where rPPG attains lower error than rBCG is: 89% (III), 83% (IV), 81% (V), 73% (VI); conversely, rBCG is better in 11%, 17%, 19%, and 27%, respectively. The increasing share of rBCG wins with higher phototype quantifies the modality shift implied by the quality metrics and supports the best-of strategy as phototype increases.

4. Discussion

This work extends the evidence base for camera-based cardiophysiology by formalizing a signal-quality–driven selection mechanism between rPPG and rBCG and by analyzing its behavior on the joint quality plane. The near-identity agreement of rPPG with the reference for heart rate (HR) and the moderate agreement for SDNN replicate the well-described pattern in which optical methods recover rate with higher fidelity than variability, the latter being intrinsically more sensitive to timing jitter and windowing. rBCG, while weaker on average, offers complementary robustness when optical quality degrades, reflecting its reliance on mechanical micro-motion rather than chromatic changes. Two measurement factors help explain this pattern. First, timing jitter—small random perturbations in beat timing arising from motion, frame-rate quantization, and peak-detection error—adds variance to inter-beat intervals and disproportionately inflates HRV error relative to HR. Second, windowing—estimating metrics in finite 60-s segments—introduces dependence on window length and placement; short windows increase estimator variance and edge effects, whereas longer windows improve stability at the cost of responsiveness. The decision-surface analyses provide a mechanistic account of why the selection strategy succeeds. After normalization, the rPPG and rBCG quality metrics contribute comparably to error discrimination, and the intersection ridge between the two error surfaces yields an explicit, stable decision boundary. This geometry supports simple linear rules for selection and motivates explicit abstention policies in regions where both qualities are low. In decision-theoretic terms, the selector approximates a risk-sensitive policy that minimizes expected error subject to a constraint on returning results only when quality exceeds a task-specific threshold. In aggregate, the "best-of" strategy improves SDNN accuracy by approximately 7%, with larger gains when one modality's quality systematically degrades—for example, at higher Fitzpatrick phototypes where rPPG quality drops and the boundary shifts to favor rBCG (≈18% improvement at type VI). Practical control of the operating point is possible by tuning the boundary (or adding a small margin, δ) to trade coverage against accuracy in low-quality regions.

Subgroup analyses across Fitzpatrick phototypes demonstrate that rPPG error increases with higher pigmentation, whereas rBCG shows limited dependence on phototype. The quality-aware selector therefore reallocates weight toward rBCG as phototype increases, mitigating a well-documented equity concern of optical sensing. Importantly, the residual increase in error at the highest phototypes indicates that selection alone does not fully close the gap, pointing to opportunities for improved signal modeling, illumination control, sensor/ISP tuning, and diversification of training data.

In addition to the SDNN-focused analyses, the present results for Heart Rate (HR) warrant explicit discussion. HR estimation by rPPG demonstrated near-identity agreement with the reference, with overall MAE well below 0.5 bpm and correlation above 0.98. In contrast, rBCG alone yielded higher aggregate error (MAE ≈ 3.5 bpm), yet its performance improved markedly when quality was high. Importantly, the quality-driven selection mechanism reduced the final HR error to 0.43 bpm, confirming that the same selection logic effective for SDNN also benefits HR, albeit with smaller absolute gains (~3% improvement).

Subgroup analyses by Fitzpatrick phototype further revealed that rPPG HR error rises slightly with increasing pigmentation, from 0.16 bpm (Type III) to 1.28 bpm (Type VI). rBCG error remained more stable across phototypes but consistently higher (≈2.8–4.8 bpm). The best-of selection therefore continued to favor rPPG in the vast majority of windows (>95%), but with a growing share of rBCG contributions at darker skin tones. This adaptive reallocation limited the increase in final HR error across phototypes, which remained below 1.3 bpm even at Type VI.

These HR-specific findings extend the general conclusion that rPPG provides superior accuracy under favorable conditions, while rBCG contributes selectively when rPPG quality degrades. They also highlight that quality-gated selection mitigates, though does not entirely eliminate, the performance gap at higher phototypes. Together with the SDNN results, these observations strengthen the rationale for multimodal, quality-aware cardiophysiological monitoring.

Operationally, coupling each estimate with its quality enables several practical controls aligned with clinical expectations: quality thresholds to constrain error; dynamic re-record prompts when both qualities are low; and user-facing confidence labels to support interpretation. Calibrating the mapping from quality to expected error—e.g., via isotonic or Platt-style calibration on held-out data—would allow quality to be treated as a probabilistic reliabili-

ty score, improving transparency for end users and enabling scenario-specific operating points (e.g., high-sensitivity versus high-specificity settings). Reporting a per-window quality-derived confidence alongside SDNN and heart rate provides transparent, actionable uncertainty estimates.

Methodologically, selection and fusion should be viewed as complementary. The present results show that selection alone yields measurable gains; however, in regions where both modalities have intermediate quality, weighted fusion (with weights derived from calibrated quality) may reduce variance relative to either modality alone. A hierarchical policy—abstain when both qualities are poor; select when one clearly dominates; fuse when both are moderate—offers a principled roadmap for future iterations without sacrificing interpretability.

Finally, robust deployment will depend on generalization beyond the controlled setting used here. Cross-device and cross-environment calibration, sensitivity to image signal-processing pipelines, and temporal stability across repeated measures warrant systematic study. Extending validation to additional HRV endpoints (e.g., RMSSD, frequency-domain metrics) and to clinical cohorts characterized by arrhythmia or low perfusion will clarify the scope of safe use and the residual failure modes of camera-based cardiophysiology.

4.1 Limitations of the Study

This evaluation was conducted under controlled, seated conditions with uniform frontal illumination and short (≈60-s) recordings. Performance in unconstrained environments—variable lighting, head pose changes, background motion, and hand-held capture—was not directly assessed. Consequently, error rates reported here likely underestimate worst-case performance in the wild, and additional validation is required to quantify robustness under everyday use.

Ground truth was derived exclusively from transmissive pulse oximetry; heart rate and inter-beat intervals were computed from the photoplethysmographic waveform with synchronized timestamps. The absence of an ECG reference precludes characterization of timing offsets relative to electrical R-peaks and limits assessment of arrhythmias or ectopy. Although pulse oximetry is an accepted reference for HR and IBI at rest, latency and morphology differences (e.g., variability in pulse transit time) may introduce small systematic biases. A prospective study including a harmonized multi-lead ECG and unified synchronization would reduce this source of error and more fully characterize timing performance.

The demographic composition was skewed toward Fitzpatrick phototypes III–IV, with incidental representation of I–II and more limited representation of V–VI. As a result, precision of subgroup estimates is lower for the rarest categories; the residual elevation of error in darker phototypes should therefore be interpreted with appropriate uncertainty.

Exclusion criteria removed clinically important groups (e.g., significant anemia, heart failure, marked respiratory compromise, pacemaker carriers), limiting generalizability to high-risk populations. Furthermore, participants were recorded at rest; the effect of exercise, talking, or facial expressions—common in telehealth—was not examined.

Only SDNN was analyzed as the HRV endpoint. Other time-domain (e.g., RMSSD, pNN50) and frequency-domain measures (e.g., LF/HF) may respond differently to camera-based sensing and to the selection policy. We did not evaluate beat-to-beat timing errors relative to ECG fiducials, nor did we assess arrhythmia detection.

All data were collected with a single smartphone/lighting configuration. Cross-device and cross-ISP (image signal processing) robustness, frame-rate sensitivity, compression effects, and generalization to different camera geometries were not systematically studied.

Finally, statistical analyses focused on point estimates (MAE/RMSE) without full uncertainty quantification at the subject level. Test–retest reliability, day-to-day within-subject variability, and potential confounders (caffeine, recent exercise, medications) were not captured, precluding stability analyses over time.

5. Conclusions

A quality-driven "best-of" selection between rPPG and rBCG enables reliable camera-based estimation of HR and SDNN from ≈60 s recordings with pulse-oximetry ground truth (shorter measurement times of 30s and 45s are also possible). rPPG provides near-identity HR and accurate SDNN at high confidence, while rBCG contributes when optical quality degrades. Quality scores are consistently negatively correlated with error, and the intersection of the modality-specific error surfaces on the joint quality plane yields a stable, interpretable decision boundary that justifies simple selection. At the population level, the policy improves SDNN by ~7% overall and ~18% at Fitzpatrick VI, mitigating skin-tone–related disparities while preserving rPPG's HR accuracy under favorable conditions. In practice, systems should pair each output with a quality score, abstain or prompt a brief re-recording when quality is low, and

tune a small margin (δ) around the decision boundary to balance accuracy and coverage; future work should calibrate quality-to-error mappings, evaluate selection-plus-fusion policies, and extend validation across devices, environments, and HRV endpoints.

6. Summary

Remote photoplethysmography (rPPG) and ballistocardiography (rBCG) enable contactless monitoring of cardiovascular function but each modality has inherent limitations. rPPG offers high accuracy for heart rate (HR) but is sensitive to lighting and skin pigmentation, while rBCG is less affected by these factors yet noisier overall. Shen.AI developed a multimodal, quality-driven selection algorithm that dynamically chooses the modality with the higher confidence score to optimize performance.

In a cohort of 5,311 participants (mean age 53.8 years, 64.7% female), simultaneous rPPG, rBCG, and pulse oximetry reference recordings were analyzed. rPPG achieved near-identity agreement for HR (MAE ≈ 0.37 bpm, R = 0.99), while rBCG showed higher error (MAE ≈ 3.6 bpm) but improved substantially under high-quality conditions. For HRV (SDNN), rPPG again outperformed rBCG (MAE ≈ 6 ms vs 36 ms). Applying the best-of selector reduced errors by ~3% for HR and ~7% for SDNN. Stratification by Fitzpatrick skin type confirmed rising rPPG error with darker phototypes, while rBCG remained stable; selection mitigated disparities, with SDNN improvement reaching ~18% at type VI.

This multimodal strategy ensures robust and equitable contactless monitoring of HR and HRV across diverse populations.

6.1 Take home message

This study demonstrates that remote photoplethysmography (rPPG) provides highly accurate heart rate estimation, while remote ballistocardiography (rBCG) offers complementary robustness when rPPG quality degrades, particularly in darker skin tones. By leveraging quality metrics, Shen.AI's multimodal "best-of" selection algorithm dynamically chooses the most reliable modality, reducing errors by ~3% for HR and ~7% for HRV (SDNN). In darker phototypes, improvements were even larger (≈18% for SDNN). These results confirm that multimodal, quality-aware cardiophysiology can deliver reliable and equitable contactless monitoring across diverse populations, supporting broader telemedicine and preventive healthcare applications.