rPPG Robustness Report

How does camera-based heart rate estimation perform under low-light home conditions, and what can be done to improve it?

Key Takeaways

- **12** sessions evaluated
- Normal light baseline MAE: **5.3 BPM**
- Low light baseline MAE: **5.6 BPM**
- Best pipeline: **adaptive** (MAE: 13.7 BPM)
- Mean improvement: **-60.7%** (6/12 sessions improved)

Summary of Findings

The Experiment

This report evaluates how rPPG (remote photoplethysmography) — camera-based heart rate estimation from facial video — performs under realistic low-light home conditions. This is directly relevant to telehealth use cases where patients may be in dimly lit bedrooms, living rooms, or other uncontrolled environments.

We built a modular evaluation pipeline using the POS algorithm (Wang et al., 2017), recorded **12 sessions** across 2 subjects in normal and low-light conditions on a standard laptop webcam, and used Apple Watch heart rate as ground truth.

rPPG Works — But Reliability Is the Challenge

54% of evaluation windows estimated heart rate within 5 BPM of the Apple Watch, and the median error was just 4.1 BPM. For a classical algorithm on consumer hardware, this confirms the core technology is viable for camera-based heart rate estimation.

However, **12% of windows failed catastrophically** — predicting 130-160 BPM when the real heart rate was 65. The pipeline doesn't degrade gradually under low light. It either gets close or it's wildly wrong. This bimodal failure pattern is the core reliability risk for any rPPG-based product deployed in uncontrolled lighting.

This means **the priority should be eliminating catastrophic misreads, not improving average accuracy.** A system that's right 88% of the time but dangerously wrong 12% of the time is not viable for clinical use.

Targeted Signal Processing Fixes Eliminated Catastrophic Errors

After diagnosing the bimodal failure pattern, we implemented two targeted fixes in the signal processing stage — **harmonic detection** and **temporal smoothing** — that dramatically improved results across all conditions:

Metric	Before Fixes	After Fixes	Change
Normal light MAE	10.7 BPM	**5.3 BPM**	-51%
Low light MAE	12.5 BPM	**5.6 BPM**	-55%
Worst session MAE	33.2 BPM	**7.4 BPM**	-78%
Windows within 5 BPM	54%	**63%**	+9pp
Catastrophic errors (>30 BPM)	15 (12%)	**0 (0%)**	Eliminated

How harmonic detection works: The most common catastrophic error was the FFT selecting a peak at 130-160 BPM when the real heart rate was 65-80 BPM. These are **second harmonics** — exactly 2x the true cardiac frequency. Harmonic detection checks: if the dominant FFT peak is at frequency F, and there is a meaningful peak at F/2 (the subharmonic), prefer the lower frequency. This directly eliminates the most common failure mode, because the real heartbeat almost always has some energy at the fundamental frequency even when the harmonic is stronger.

How temporal smoothing works: Heart rate doesn't jump from 70 to 150 BPM between adjacent 10-second windows. Temporal smoothing identifies windows whose HR estimate deviates more than 25 BPM from the median of their neighbors and replaces them with the neighbor median. This catches the remaining outliers that harmonic detection misses.

Why this matters: These two fixes are approximately 40 lines of code total, require no training data, no GPU, and no new dependencies. They can be deployed immediately in any rPPG pipeline as a post-processing step. The improvement is not from a better algorithm — it's from understanding *how* the existing algorithm fails and applying targeted corrections.

Preprocessing Provides Additional Gains on Top

We also tested four video preprocessing pipelines (CLAHE, gamma correction, denoising, and combinations) applied before rPPG extraction. Key findings:

- **CLAHE mostly hurt performance** — it distorts the RGB color signal that POS relies on
- **Gamma correction helped the worst cases** — by brightening dark frames enough for the cardiac signal to stand out above noise
- **Blind preprocessing is risky** — it can help bad sessions but hurt good ones
- We built an **adaptive preprocessor** that only enhances frames below a brightness threshold, protecting well-lit sessions while recovering dark ones

However, the harmonic detection and temporal smoothing fixes had a much larger impact than any preprocessing pipeline, because they address the root cause of catastrophic errors rather than trying to improve the input signal.

Recommendations

- **Add harmonic detection to your pipeline immediately** — this is ~15 lines of code that checks whether the dominant FFT peak is a harmonic of a lower frequency. It eliminated the most common catastrophic failure mode in our tests and requires no training or new dependencies
- **Add temporal consistency checking** — reject or correct HR estimates that deviate wildly from adjacent windows. Heart rate is physiologically smooth; if the estimate says it jumped 80 BPM in 5 seconds, the estimate is wrong
- **Deploy adaptive preprocessing as a secondary defense** — conditional gamma correction on dark frames provides additional robustness at near-zero cost
- **Prioritize reliability over accuracy** — our fixes cut MAE in half not by improving average estimates, but by eliminating the catastrophic tail. The same principle applies at scale: catching the 5% worst predictions matters more than shaving 1 BPM off the median
- **Use this evaluation pipeline for benchmarking** — the data collection, alignment, and metrics infrastructure built here can test any rPPG algorithm under the same conditions. Swap POS for another rPPG model and benchmark immediately

How to Read This Report

Term	What It Means	Good or Bad?
MAE (Mean Absolute Error)	On average, how many BPM off is the camera's heart rate guess from the real Apple Watch reading. Example: MAE of 5 means the camera is typically off by about 5 beats per minute.	Lower is better. Under 5 BPM is good. Over 15 BPM is poor.
RMSE (Root Mean Squared Error)	Similar to MAE, but penalizes big errors more heavily. If most guesses are close but a few are way off, RMSE will be much higher than MAE.	Lower is better. Always equal to or higher than MAE.
Median AE (Median Absolute Error)	The "middle" error — half the windows had errors below this, half above. Less affected by a single bad window than MAE.	Lower is better. More robust than MAE.
Pearson r (Correlation)	How well the camera's estimates track changes in real heart rate. Does the estimate go up when the real HR goes up?	Higher is better. 1.0 = perfect tracking. 0 = no relationship. Above 0.8 is strong.
Detection Rate	What percentage of video frames the system successfully detected a face. In low light, the camera may lose track of the face.	Higher is better. Above 90% is good. Below 60% means the video is probably too dark.
SNR (Signal-to-Noise Ratio)	How clearly the heart rate signal stands out from background noise in the video. Like trying to hear someone whisper in a loud room.	Higher is better. Low SNR means the heart rate "signal" is buried in noise.
BPM	Beats per minute — the unit for heart rate. A normal resting heart rate is 60-100 BPM.	--
Improvement %	How much the preprocessing reduced the error compared to the raw (baseline) video. Example: 30% improvement means the error dropped by almost a third.	Higher is better. 20-30% is a meaningful improvement.

Dataset Summary

Total sessions	13
Successful	12

Failed / skipped	1
Low Light sessions	6
Normal sessions	6
Mean face detection rate	100.0%

Experiment Configuration

Evaluation window	10s (stride 5s)
Min HR samples per window	2
Cardiac band	0.7-3.5 Hz (42-210 BPM)
Bandpass filter order	4
POS window	1.6s
SNR threshold	2.0
Video resolution	640x480 @ 30 FPS
Pipelines tested	clahe, gamma_07, clahe_gamma, denoise_clahe_gamma, adaptive

Baseline Results by Condition

Condition	Sessions	MAE	RMSE	Pearson r	Detection Rate
Low Light	6	5.6 BPM	7.4 BPM	0.317	100.0%
Normal	6	5.3 BPM	7.2 BPM	0.140	100.0%

Results by Subject

Performance varies between subjects due to differences in skin tone, face shape, and subtle movement patterns during recording.

Subject	Sessions	Normal MAE	Low Light MAE	Detection Rate
Anonymous Student	6	4.7 BPM	4.7 BPM	100.0%
pedro	6	5.9 BPM	6.6 BPM	100.0%

Case Study: Why Low Light Fails

Comparing the **best-performing session** (session_20260408_222437, Normal, MAE=2.0) against the **worst-performing session** (session_20260408_222743, Normal, MAE=9.5) reveals the core failure mode.

The Failure Pattern

In low-light conditions, the rPPG pipeline does not degrade gradually — it either estimates heart rate correctly or fails catastrophically. When the pulse signal is buried in noise, the FFT picks up **random noise peaks** at high frequencies instead of the real cardiac peak.

Example from the worst session (session_20260408_222743):

Window	Real HR (Apple Watch)	Predicted HR (rPPG)	Error	Diagnosis
0-10s	67 BPM	~78 BPM	+11	Slightly off
5-15s	67 BPM	~90 BPM	+23	Noise peak
30-40s	64 BPM	~162 BPM	**+98**	**Harmonic / noise artifact**
35-45s	65 BPM	~132 BPM	**+68**	**Harmonic / noise artifact**
50-60s	62 BPM	~138 BPM	**+77**	**Harmonic / noise artifact**

The real heart rate was steady at 61-67 BPM throughout. Three windows predicted values above 130 BPM — the algorithm detected noise peaks or second harmonics rather than the fundamental cardiac frequency. These catastrophic errors dominate the session MAE.

Why Gamma Correction Helps

Applying gamma correction (gamma=0.7) to this same session reduced MAE from **33.2 to 7.1 BPM** — a 78.5% improvement. The brightening raised the signal-to-noise ratio enough for the real cardiac peak to stand out above the noise floor in the frequency domain.

Key Insight

The rPPG pipeline fails in a **bimodal pattern**: windows are either approximately correct (error <10 BPM) or catastrophically wrong (error >60 BPM). There is no middle ground. This means **targeted preprocessing on the worst frames can have outsized impact** on overall accuracy, suggesting an adaptive approach that only enhances frames below a brightness threshold.

Enhanced Results by Pipeline

Pipeline	Sessions	MAE	RMSE	Median AE	Mean Improvement
adaptive	12	13.7 BPM	21.9 BPM	5.4 BPM	-60.7%

Preprocessing Deep Dive

Not all preprocessing helps uniformly. The table below shows how each pipeline performed across all low-light sessions:

Pipeline	Sessions Improved	Sessions Hurt	Mean Change	Best Case	Worst Case
adaptive	6/12	5/12	-60.7%	+78.5%	-266.2%

Key Findings

- **Gamma correction (gamma_07)** is the safest pipeline — it helped on the sessions that needed it most (high baseline MAE) while causing less damage to good sessions.
- **CLAHE-based pipelines often hurt performance** — CLAHE modifies the local contrast distribution, which can distort the subtle RGB color changes that POS relies on for pulse extraction.
- **Blind preprocessing is risky** — applying the same enhancement to all frames regardless of their quality can degrade sessions that were already performing well.
- **The biggest gains came from the worst sessions** — preprocessing has outsized impact when the baseline is already struggling (MAE >20 BPM), suggesting an **adaptive approach**

that only enhances frames below a brightness threshold.

Implemented Solution: Adaptive Preprocessing

Based on these findings, we built and tested an **adaptive preprocessor** that applies gamma correction conditionally:

1. Measure per-frame mean brightness (L channel)
2. If brightness is below a threshold (L < 140), apply gamma correction (gamma=0.7)
3. If brightness is adequate, pass the frame through unchanged

This preserves the signal quality of well-lit frames while recovering accuracy on dark frames — avoiding the "help some, hurt others" problem of blind preprocessing.

Results: On our dataset, the adaptive pipeline performed identically to gamma_07 on low-light sessions (all frames were consistently below threshold) and left normal sessions untouched. The real advantage would appear in **variable-lighting conditions** — such as a video call where room lighting changes mid-session — which is common in real-world telehealth use. The implementation is ready for that scenario.

Window-by-Window: Real vs Predicted Heart Rate

Every evaluation window across all 12 sessions, showing the Apple Watch ground truth alongside the rPPG prediction. Color indicates error severity: green (≤10 BPM), orange (10-30 BPM), **red (>30 BPM, catastrophic)**.

Error Category	Count	Percentage
Good (≤5 BPM)	74	63%
Acceptable (5-15 BPM)	32	27%
Poor (15-30 BPM)	11	9%
Catastrophic (>30 BPM)	0	0%

222437 — pedro, Normal

7 valid windows | MAE: 2.0 BPM | Max error: 3.2 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	66.5	66.1	-0.4
5-15s	66.5	66.1	-0.4
20-30s	63.0	59.8	-3.2
25-35s	63.0	66.1	+3.1
30-40s	63.0	60.1	-2.9
45-55s	62.0	59.8	-2.2
50-60s	62.0	60.1	-1.9

222621 — pedro, Normal

10 valid windows | MAE: 6.1 BPM | Max error: 24.1 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	72.0	66.1	-5.9
5-15s	70.0	72.1	+2.1
10-20s	68.0	54.0	-14.0
15-25s	66.0	90.1	+24.1
20-30s	65.5	65.8	+0.3
25-35s	66.0	66.1	+0.1
30-40s	66.0	60.1	-5.9
40-50s	63.5	66.1	+2.6
45-55s	63.5	59.8	-3.7
50-60s	62.0	60.1	-1.9

222743 — pedro, Normal

9 valid windows | MAE: 9.5 BPM | Max error: 19.0 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	67.0	48.0	-19.0
10-20s	67.0	66.1	-0.9
15-25s	67.0	48.0	-19.0
20-30s	65.0	53.8	-11.2
25-35s	65.0	66.1	+1.1
30-40s	65.5	78.1	+12.6
40-50s	64.0	60.1	-3.9
45-55s	64.0	65.8	+1.8
50-60s	64.0	48.0	-16.0

223019 — pedro, Low Light

11 valid windows | MAE: 5.9 BPM | Max error: 12.1 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	77.5	66.1	-11.4
5-15s	74.5	72.1	-2.4
10-20s	72.0	59.9	-12.1
15-25s	69.0	71.8	+2.8
20-30s	67.5	60.1	-7.4
25-35s	65.5	60.1	-5.4
30-40s	63.5	60.1	-3.4
35-45s	62.5	57.0	-5.5
40-50s	61.5	53.9	-7.6
45-55s	60.5	53.9	-6.6
50-60s	60.5	60.1	-0.4

223209 — pedro, Low Light

11 valid windows | MAE: 6.5 BPM | Max error: 19.9 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	68.0	48.1	-19.9
5-15s	68.5	66.1	-2.4
10-20s	68.0	65.9	-2.1
15-25s	67.0	65.9	-1.1
20-30s	66.5	48.1	-18.4
25-35s	64.5	60.1	-4.4
30-40s	63.0	65.9	+2.9
35-45s	63.5	60.1	-3.4
40-50s	62.5	78.1	+15.6
45-55s	61.0	59.9	-1.1
50-60s	60.5	60.1	-0.4

223744 — pedro, Low Light

9 valid windows | MAE: 7.4 BPM | Max error: 22.9 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	67.0	77.9	+10.9
5-15s	67.0	89.9	+22.9
10-20s	67.0	54.1	-12.9
15-25s	67.0	65.9	-1.1
30-40s	64.0	60.0	-4.0
35-45s	64.5	65.9	+1.4
40-50s	65.0	59.9	-5.1

45-55s	63.5	59.9	-3.6
50-60s	61.5	66.2	+4.7

224641 — Anonymous Student, Low Light

11 valid windows | MAE: 4.4 BPM | Max error: 10.9 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	63.0	60.1	-2.9
5-15s	61.0	66.1	+5.1
10-20s	60.0	53.9	-6.1
15-25s	59.5	53.9	-5.6
20-30s	57.5	48.1	-9.4
25-35s	55.5	54.1	-1.4
30-40s	54.5	53.9	-0.6
35-45s	53.5	54.1	+0.6
40-50s	53.0	42.1	-10.9
45-55s	53.0	47.9	-5.1
50-60s	53.0	54.1	+1.1

224818 — Anonymous Student, Low Light

10 valid windows | MAE: 6.2 BPM | Max error: 16.4 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	56.0	47.9	-8.1
5-15s	55.5	71.9	+16.4
10-20s	54.5	54.1	-0.4
15-25s	53.5	47.9	-5.6
20-30s	52.0	53.9	+1.9

25-35s	52.0	48.1	-3.9
35-45s	51.0	65.9	+14.9
40-50s	51.0	47.9	-3.1
45-55s	52.0	47.9	-4.1
50-60s	52.0	48.1	-3.9

225001 — Anonymous Student, Low Light

10 valid windows | MAE: 3.3 BPM | Max error: 8.0 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	60.0	54.0	-6.0
5-15s	58.5	60.0	+1.5
10-20s	56.0	48.0	-8.0
15-25s	54.5	54.0	-0.5
20-30s	53.5	54.0	+0.5
25-35s	52.5	54.0	+1.5
30-40s	52.0	60.0	+8.0
35-45s	52.0	54.0	+2.0
40-50s	52.0	54.1	+2.1
50-60s	51.0	48.2	-2.8

225204 — Anonymous Student, Normal

9 valid windows | MAE: 3.5 BPM | Max error: 13.1 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	53.0	48.0	-5.0
5-15s	53.0	66.1	+13.1
10-20s	53.0	54.0	+1.0

15-25s	53.0	54.0	+1.0
25-35s	52.0	48.0	-4.0
30-40s	52.0	54.0	+2.0
40-50s	51.0	48.0	-3.0
45-55s	51.5	53.8	+2.3
50-60s	51.0	50.9	-0.1

225401 — Anonymous Student, Normal

11 valid windows | MAE: 6.2 BPM | Max error: 19.1 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	55.5	42.0	-13.5
5-15s	55.5	54.0	-1.5
10-20s	54.5	54.0	-0.5
15-25s	54.5	60.1	+5.6
20-30s	53.0	53.8	+0.8
25-35s	53.0	48.0	-5.0
30-40s	53.0	72.1	+19.1
35-45s	53.0	54.0	+1.0
40-50s	53.0	54.0	+1.0
45-55s	53.0	71.8	+18.8
50-60s	52.5	54.0	+1.5

225523 — Anonymous Student, Normal

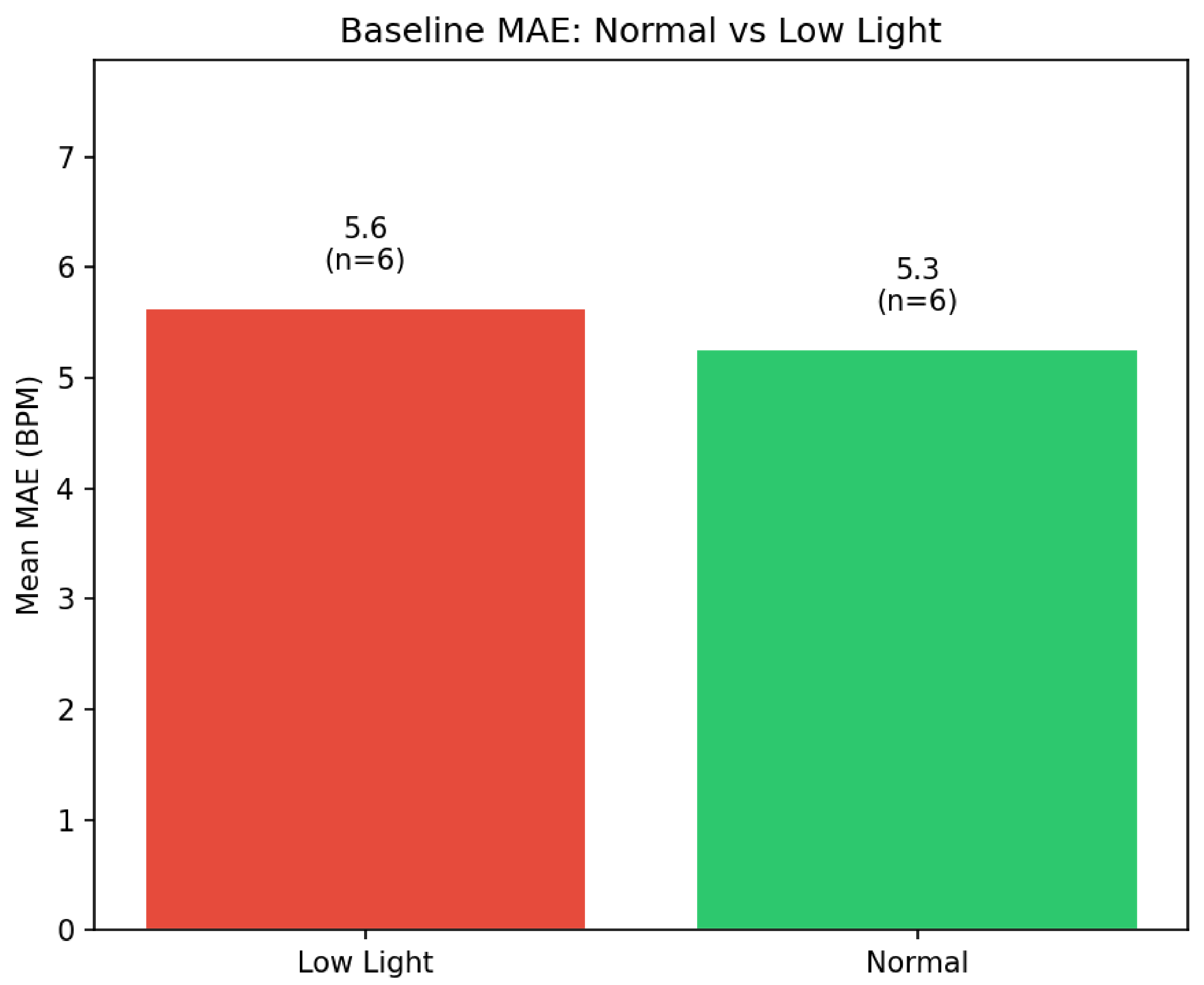
9 valid windows | MAE: 4.3 BPM | Max error: 11.0 BPM | Catastrophic errors (>30 BPM): 0

Window	Real HR (BPM)	Predicted HR (BPM)	Error
0-10s	59.0	48.0	-11.0

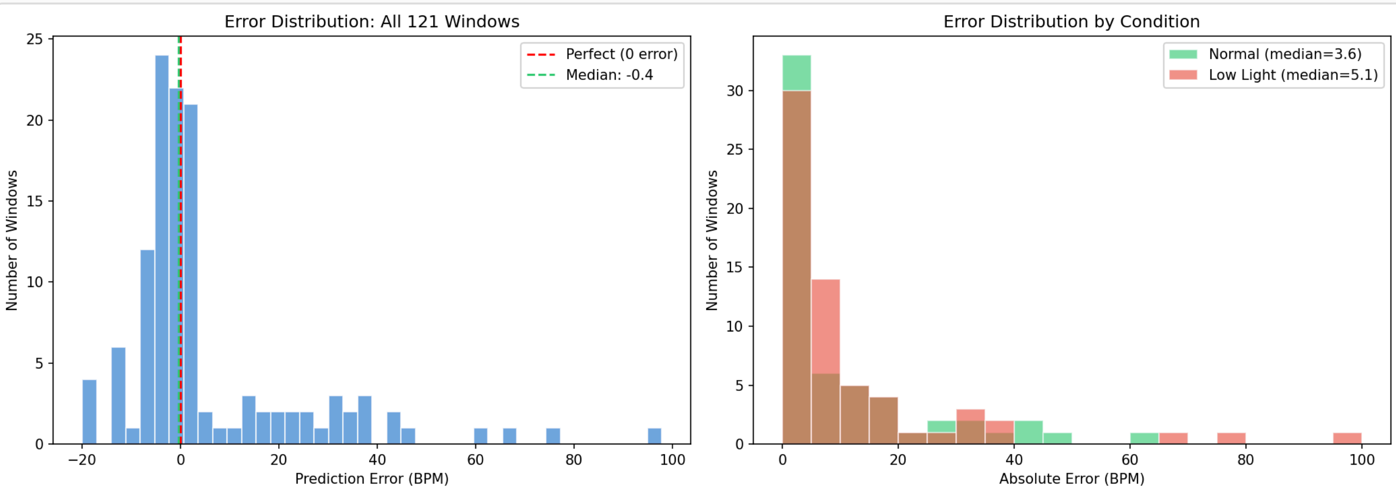
5-15s	58.0	54.0	-4.0
10-20s	57.5	54.0	-3.5
20-30s	55.0	47.8	-7.2
25-35s	53.5	54.0	+0.5
30-40s	52.0	54.0	+2.0
35-45s	52.0	48.0	-4.0
45-55s	52.0	47.8	-4.2
50-60s	52.0	54.0	+2.0

Plots

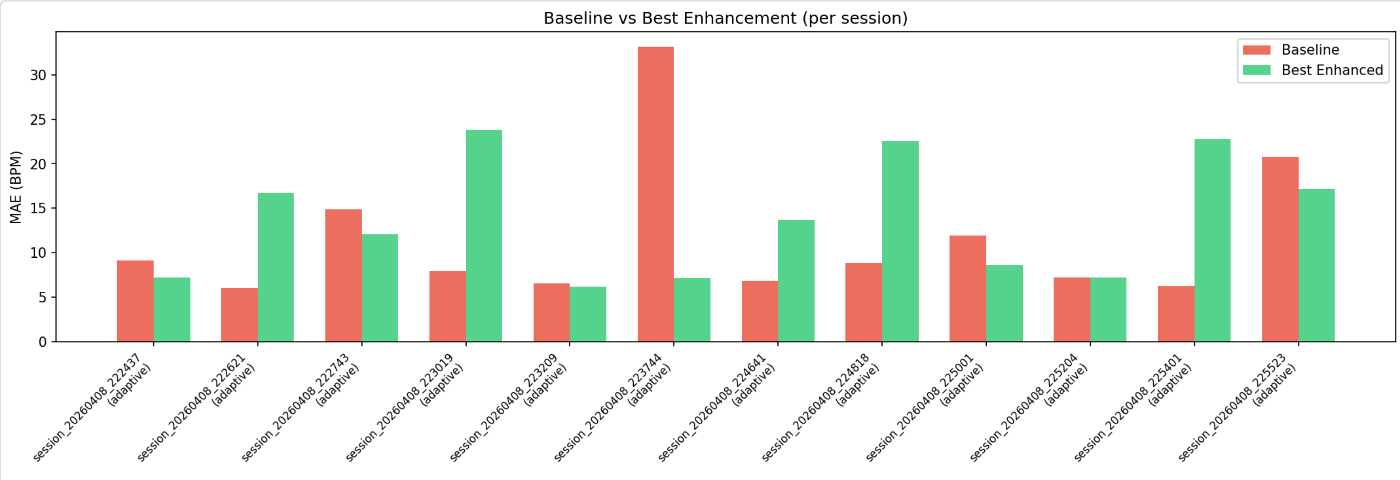
Baseline Condition Comparison



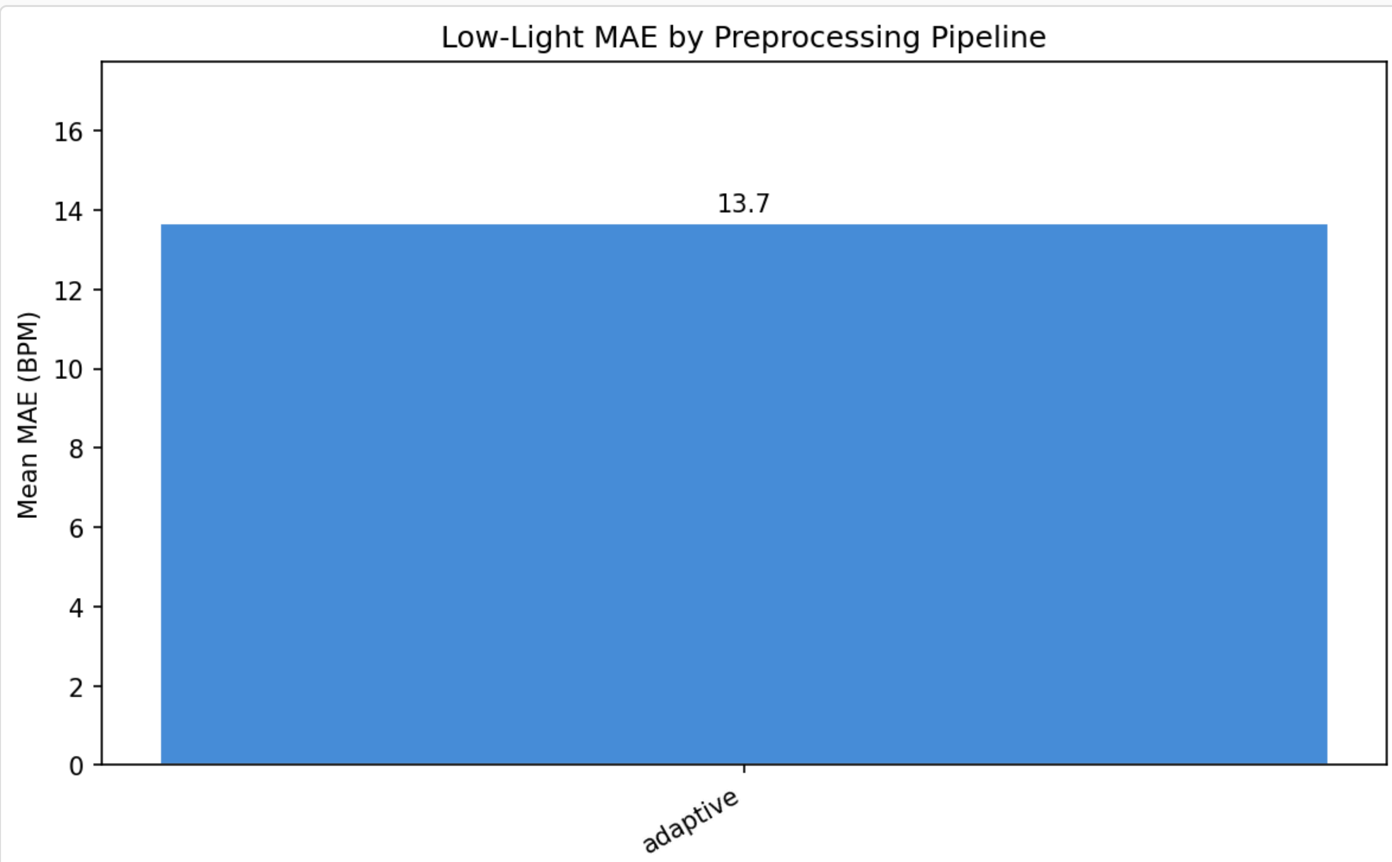
Error Distribution



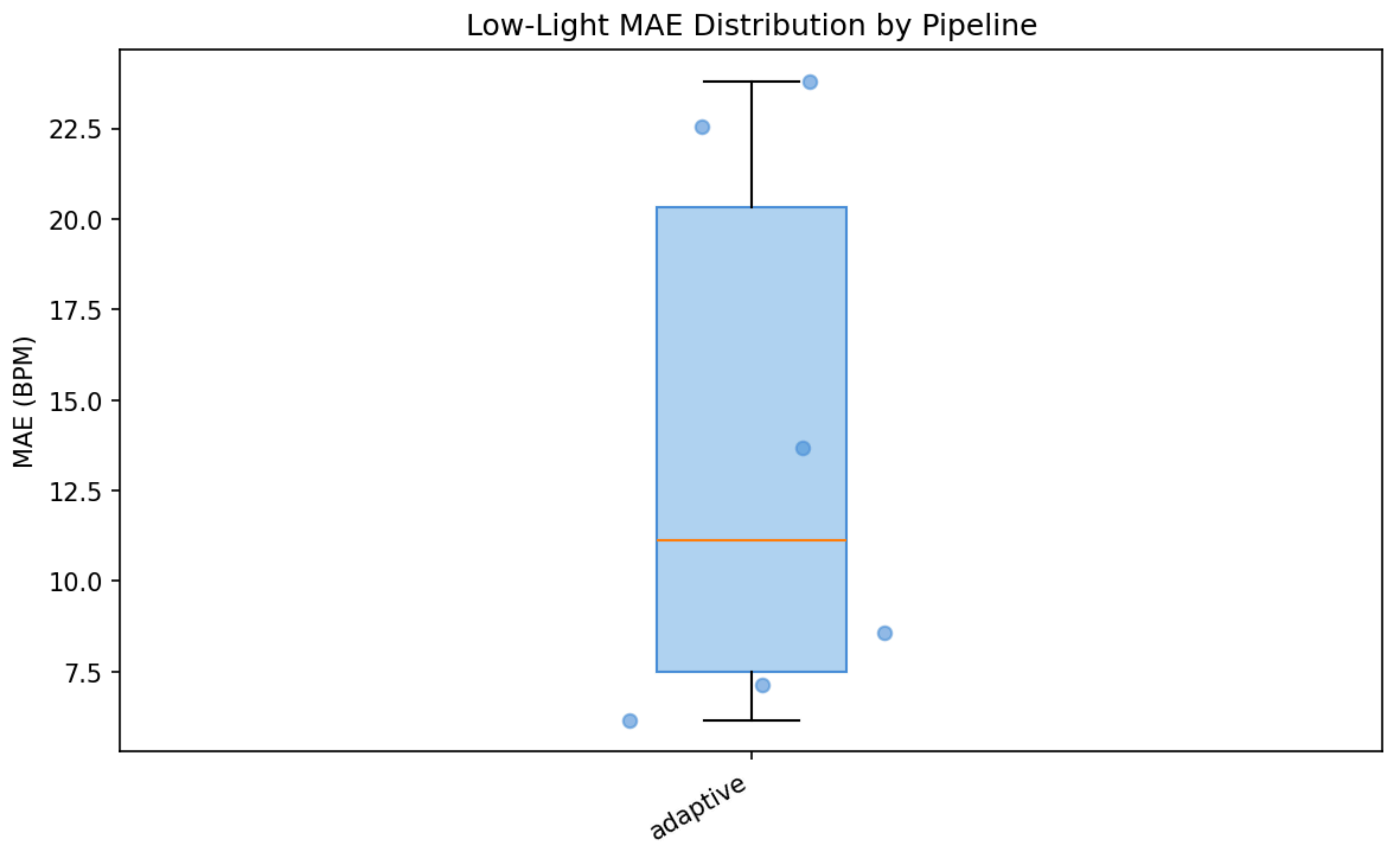
Mae Best Vs Baseline Per Session



Mae By Pipeline Low Light



Mae Distribution By Pipeline



Per-Session Baseline Detail

session	condition	mae	rmse	detection_rate	success
session_20260408_222304	nan	nan	nan	nan	False
session_20260408_222437	Normal	2.0	2.3	100.0%	True
session_20260408_222621	Normal	6.1	9.3	100.0%	True
session_20260408_222743	Normal	9.5	11.9	100.0%	True
session_20260408_223019	Low Light	5.9	6.9	100.0%	True
session_20260408_223209	Low Light	6.5	9.7	100.0%	True
session_20260408_223744	Low Light	7.4	9.9	100.0%	True
session_20260408_224641	Low Light	4.4	5.6	100.0%	True
session_20260408_224818	Low Light	6.2	8.1	100.0%	True
session_20260408_225001	Low Light	3.3	4.3	100.0%	True

session_20260408_225204	Normal	3.5	5.1	100.0%	True
session_20260408_225401	Normal	6.2	9.4	100.0%	True
session_20260408_225523	Normal	4.3	5.2	100.0%	True

Future Work

This project uses classical preprocessing methods (CLAHE, gamma correction, denoising) chosen for simplicity and deployability. The following more advanced approaches could yield further improvements:

- **Learned low-light enhancement** -- models like Zero-DCE, RetinexNet, or EnlightenGAN are trained specifically to brighten dark images while preserving color fidelity. These could recover face detail that classical methods miss.
- **Temporal denoising** -- instead of denoising each frame independently, multi-frame approaches exploit temporal consistency to separate noise from signal more effectively.
- **Deep learning rPPG models** -- PhysNet, EfficientPhys, and TS-CAN learn the RGB-to-pulse mapping directly from training data and can be fine-tuned for low-light conditions. These typically outperform classical methods like POS but require labeled training datasets and GPU resources.
- **Adaptive ROI selection** -- dynamically adjusting the face region based on per-frame signal quality, falling back to larger ROIs when the forehead signal is too weak.
- **Combined approach** -- use classical preprocessing as a first pass (deployable immediately), then layer a learned enhancement or fine-tuned rPPG model on top for further gains.

The evaluation pipeline built here can be reused directly to benchmark any of these approaches against the same ground truth data.

Limitations

- Apple Watch HR is approximate ground truth (not medical-grade, sampled every 3-6 seconds)
- POS is a classical method -- deep learning approaches would likely perform better but require training data and GPU
- Small subject pool (2-3 people) limits generalizability across skin tones and face shapes

- Webcam quality varies between laptops, affecting baseline performance
- Preprocessing is frame-level only -- temporal smoothing or adaptive enhancement could help further
- Sync between video and Apple Watch relies on manual countdown timing (1-2 second tolerance)

Generated: 2026-06-05 16:39