Appendix

Tayssir Bouraffa
Chalmers University of Technology
Department of Computer Science and Engineering
Gothenburg, Sweden
tayssir@chalmers.se

Dimitrios Koutsakis
Chalmers University of Technology
Gothenburg, Sweden
koutsakis.d@hotmail.com

Salvija Zelvyte
Chalmers University of Technology
Gothenburg, Sweden
salvija.zelvyte22@gmail.com

Appendix A. Dataset Folds

To accurately evaluate the performance of the models in vehicular environments, the MR-NIRP dataset was utilized for both training and testing purposes. To minimize overfitting and ensure fair predictions, separate train-validation-test splits were created. Since predefined folds for training and testing are not available in the MR-NIRP dataset, we followed the approach described by Gideon et al. [2]. As shown in Table 1, the dataset was partitioned into five folds based on subject IDs, with a unique test set set aside for each fold. Each model was trained and tested on all folds, and the average performance across these folds was reported as the benchmark result.

Table 1. Subject IDs assigned to the training, validation, and test sets for each fold of the cross-validation process.

	Train Set	Validation Set	Test Set
Fold 1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	13, 14, 15	16, 17, 18, 19
Fold 2	5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16	17, 18, 19	1, 2, 3, 4
Fold 3	1, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19	2, 3, 4	5, 6, 7, 8
Fold 4	1, 2, 3, 4, 5, 13, 14, 15, 16, 17, 18, 19	6, 7, 8	9, 10, 11
Fold 5	1, 2, 3, 4, 5, 6, 7, 8, 16, 17, 18, 19	9, 10, 11	13, 14, 15

Appendix B: Excluded Cases

Table 2 highlights the instances from the MR-NIRP dataset that were removed from the study due to potential issues affecting benchmark reliability.

Specifically, five sequences were excluded because poor lighting made it impossible to detect the subject's face, and one sequence was omitted due to corrupted video frames. Additionally, two cases were identified where the ground-truth PPG signals contained prolonged zero values, suggesting errors in the data sampling process. All data from Subject 12 were excluded from the benchmark due to significant noise in the ground-truth data, identified through PSD analysis, which indicated a possible sampling error. Moreover, the heart rate derived from these PPG signals consistently fell below 50 beats per minute, an atypical value for a healthy adult male like Subject 12. This issue was prevalent across most of the subject's recordings, making them unreliable for accurate ground-truth vital sign extraction [1].

Table 2. List of excluded MR-NIRP recordings from the benchmark. The entry subject12* indicates that all recording from this subject were excluded.

Justification	Excluded Cases
5***Dark Frames**	subject5_garage_still_975 subject6_garage_still_975 subject6_garage_small_motion_975 subject6_garage_large_motion_975 subject2_driving_still_940
Corrupted Frames	subject2_garage_small_motion_940
3***PPG Sampling Error**	subject7_driving_small_motion_975 subject7_driving_still_975 subject12*

Appendix C: Evaluation Metrics

To evaluate the effectiveness of rPPG algorithms in realistic automotive environments and accurately predict HR and RR under dynamic vehicular conditions, we selected five metrics that address different aspects of algorithm performance, including error rates and signal quality.

Mean Absolute Error (MAE): Represents the average absolute difference between the ground-truth signal rate

(R_{GT}) and the predicted signal rate (R_{Pred}) for HR or RR across all observation windows (T).

$$\textbf{MAE} = \frac{1}{T}\sum_{i=1}^{T}|R_{GT} - R_{Pred}|$$

Root Mean Square Error (RMSE): Evaluates the size of the prediction error by comparing the ground-truth signal rate (R_{GT}) with the predicted signal rate (R_{Pred}) across all observation windows (T).

$$\textbf{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(R_{GT} - R_{Pred})^2}$$

Mean Absolute Percentage Error (MAPE): Calculates the average absolute percentage difference between the ground-truth signal rate (R_{GT}) and the predicted signal rates (R_{Pred}), expressed as a percentage of the ground-truth, over all observation windows (T).

$$\textbf{MAPE} = \frac{100}{T}\sum_{i=1}^{T}\left|\frac{R_{GT} - R_{Pred}}{R_{GT}}\right|$$

Pearson Correlation Coefficient (ρ)**:** A statistical measure that quantifies the strength and direction of the linear relationship between the ground-truth signal rate (R_{GT}) and the predicted signal rates (R_{Pred}) across all observation windows (T).

$$\rho = \frac{\sum_{i=1}^{T}(R_{GT} - \overline{R_{GT}})(R_{Pred} - \overline{R_{Pred}})}{\sqrt{\sum_{i=1}^{T}(R_{GT} - \overline{R_{GT}})^2 \sum_{i=1}^{T}(R_{Pred} - \overline{R_{Pred}})^2}}$$

Signal-to-Noise Ratio (SNR): Defined as the ratio of the area under the curve of the power spectrum near the first and second harmonics of the ground-truth signal rate frequency to the area under the curve for the rest of the power spectrum.

$$\textbf{SNR} = \frac{1}{T}\sum_{i=1}^{T}\left|10\log_{10}\left(\frac{\sum_{f=lf}^{hf}\left(\hat{S}(f)\cdot U_t(f)\right)^2}{\sum_{f=lf}^{hf}\left(\hat{S}(f)\cdot(1-U_t(f))\right)^2}\right)\right|$$

$\hat{S}$ represents the power spectrum of the predicted signal S, f denotes the frequency, and $U_t(f)$ is a binary template set to 1 around the first and second harmonics of the ground-truth signal, and 0 elsewhere. This method considers only the power spectrum within the frequency ranges of 0.75–2.5 Hz for HR and 0.08–0.5 Hz for RR.

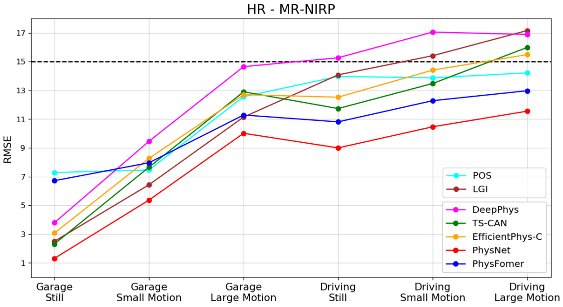

Figure 1. Visualization of qualitative HR estimation RMSE for POS and LGI unsupervised methods, as well as all supervised NN models, trained and evaluated on the MR-NIRP car dataset.

Appendix D: Model Training

To ensure a fair comparison among the rPPG models, we standardized the training parameters. This included using the AdamW optimizer for all NN models, with the exception of PhysNet, which used the Adam optimizer, and using the negative Pearson loss function. We also implemented a one-cycle learning rate scheduler, with a peak learning rate of 0.009 applied across all models. Training was conducted over 30 epochs, with the model exhibiting the lowest validation loss at the end of each epoch being selected. A batch size of 4 was consistently used in all experiments, and a 20% dropout rate was utilized to mitigate overfitting.

The training configuration for PhysFormer deviated from the standard setup due to its distinct architecture, following the experimental specifications specified in the original paper. Specifically, for PhysFormer, the Adam optimizer was used with an initial learning rate of 0.0001 and a weight decay of 0.00005, without applying a learning rate scheduler. Instead of the negative Pearson loss, a dynamic loss function was implemented, combining label distribution loss, frequency cross-entropy loss and negative Pearson loss. The model parameters were configured to match the configuration used in the pretrained models provided in [3] [4], where our codebase benchmark is paired to their toolbox, available here GitHub repository.

References

[1] Yamama Hafeez and Grossman Shamai A. Sinus bradycardia. *StatPearls [Internet]. Treasure Island, FL: StatPearls Publishing. https://www.ncbi.nlm. nih.gov/books/NBK493201/*, 2023. 1

[2] Gideon John and Simon Stent. The wayto myheart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision, Montreal, BC, Canada*, 2021. 1

[3] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak

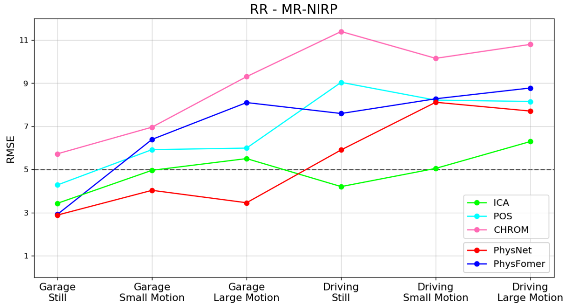

Figure 2. Visualization of qualitative RR estimation RMSE for ICA, POS, and CHROM unsupervised methods, alongside PhysNet and PhysFormer supervised NN models, trained and evaluated on the MR-NIRP car dataset.

Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. In *Advances in Neural Information Processing Systems, Vancouver, Canada*, volume 34, 2024. 2

[4] Zitong Yu, Shen Yuming, Shi Jingang, Zhao Hengshuang, Torr Philip HS, and Zhao Guoying. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA*, pages 4186–4196, 2022. 2