CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Camera-based Vital Sign Detection in Autonomous Vehicles using Deep Learning

A Benchmark Study

Master's thesis in Computer science and engineering

DIMITRIOS KOUTSAKIS & SALVIJA ZELVYTE

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Master's thesis 2024

Camera-based Vital Sign Detection in Autonomous Vehicles using Deep Learning

A Benchmark Study

Dimitrios Koutsakis & Salvija Zelvyte

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2024

Camera-based Vital Sign Detection in Autonomous Vehicles using Deep Learning
A Benchmark Study
Dimitrios Koutsakis & Salvija Zelvyte

Supervisor: Tayssir Bouraffa, Department of Computer Science and Engineering
Examiner: Hans-Martin Heyn, Department of Computer Science and Engineering

Master's Thesis 2024
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2024

Camera-based Vital Sign Detection in Autonomous Vehicles using Deep Learning
A Benchmark Study
Dimitrios Koutsakis & Salvija Zelvyte
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This study explores the feasibility of employing camera-based, deep learning algorithms for detecting vital signs in autonomous vehicles, with a focus on enhancing driver safety. By evaluating various remote photoplethysmography techniques in dynamic vehicular environments, challenges such as motion artifacts and varying lighting conditions were addressed. Findings suggest that machine learning models, particularly neural network based approaches, hold promise in accurately estimating heart rate and respiratory rate from video data in such settings. The study emphasizes the potential of deep learning methodologies to improve driver safety through the integration of non-invasive vital sign monitoring systems in autonomous vehicles. Future research should address dataset imbalances and broaden the benchmark scope to include additional vital signs and algorithms, while also exploring alternative methods such as optical-flow based approaches to enhance respiration rate detection.

Keywords: Camera-based, Vital Sign Detection, Autonomous Vehicles, Software Engineering, Benchmark, Remote Photoplethysmography, Deep Learning, Neural Network, Machine Learning.

Acknowledgements

We would like to express our sincere appreciation to our supervisor, Tayssir Bouraffa, who assisted us with continuous feedback and support through the entire research from the proposal to the final submitted report. Her invaluable contributions have not only directed but also refined our thesis project, and we are deeply grateful for her tireless dedication and mentorship.

We would also like to extend our gratitude to our examiner, Hans-Martin Heyn, for his thorough review of our proposal, active involvement in our mid-term presentation, and the invaluable suggestions he provided, which greatly enhanced our research.

Dimitrios Koutsakis & Salvija Zelvyte, Gothenburg, June 2024

List of Acronyms

List of acronyms used in this thesis in alphabetical order:

BCG	Ballistocardiography
Beats/min	Beats per Minute
Breaths/min	Breaths per Minute
BVP	Blood Volume Pulse
CNN	Convolutional Neural Network
ECG	Electrocardiogram
FFT	Fast Fourier Transform
FPS	Frames per Second
HR	Heart Rate
HRV	Heart Rate Variability
IBI	Inter-Beat Interval
ICA	Independent Component Analysis
LGI	Local Group Invarince
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MR-NIRP	MERL-Rice Near-Infrared Pulse Car Dataset
MSE	Mean Squared Error
NIR	Near-Infrared
NN	Neural Network
PPG	Photoplethysmography
PSD	Power Spectral Density
PURE	Pulse Rate Detection Dataset
RMSE	Root Mean Square Error
ROI	Region of Interest
rPPG	remote Photoplethysmography
RR	Respiration Rate
RSA	Respiratory Sinus Arrhythmia
SCAMPS	Synthetics for Camera Measurement of Physiological Signals Dataset
SNR	Signal-to-Noise Ratio
SpO2	Blood Oxygen Saturation
TSM	Temporal Shift Module
UBFC-rPPG	University Bourgogne Franche-Comté rPPG Dataset
ρ	Pearson Correlation Coefficient

Contents

List of Acronyms **ix**

List of Figures **xiii**

List of Tables **xvii**

1 Introduction **1**
1.1 Statement of the Problem . . . 1
1.2 Purpose of the Study . . . 2
1.3 Research Questions and Hypotheses . . . 2
1.4 Limitations and Delimitations . . . 3

2 Background and Related work **5**
2.1 Contact-based vital sign detection in vehicles . . . 5
2.2 Remote vital sign detection in vehicles . . . 6
2.2.1 Radar-based methods . . . 7
2.2.2 Optical flow-based methods . . . 8
2.2.3 rPPG-based methods . . . 8
2.3 Deriving vital signs from PPG signal . . . 9

3 Methodology **11**
3.1 Benchmark Study . . . 11
3.2 rPPG Algorithms . . . 12
3.2.1 Unsupervised methods . . . 13
3.2.2 Supervised Neural Network models . . . 14
3.3 rPPG Datasets . . . 15
3.3.1 General Datasets . . . 16
3.3.2 Dataset in vehicle environment . . . 17
3.4 Experimental Setup . . . 18
3.4.1 Preprocessing . . . 19
3.4.2 Postprocessing . . . 20
3.4.3 Excluded Cases . . . 21
3.5 Training Setup . . . 22
3.5.1 Dataset Folds . . . 23
3.5.2 Training Parameters . . . 23
3.6 Evaluation Metrics . . . 24
3.6.1 Performance Metrics . . . 24

3.6.2 Efficiency Metric 25

4 Results **27**
4.1 Heart Rate estimation 27
4.1.1 Unsupervised methods 27
4.1.2 Neural Network models 30
4.1.2.1 DeepPhys 31
4.1.2.2 TS-CAN 33
4.1.2.3 EfficientPhys-C 34
4.1.2.4 PhysNet 35
4.1.2.5 PhysFormer 37
4.2 Respiration Rate estimation 39
4.2.1 Unsupervised methods 39
4.2.2 Neural Network models 41
4.3 Efficiency of Neural Network models 42

5 Discussion **45**
5.1 Discussion of Results 45
5.1.1 Heart Rate estimation in vehicles 45
5.1.2 Respiration Rate estimation in vehicles 46
5.1.3 Generalization Performance 47
5.2 Influencing Factors 48
5.2.1 Vehicular Dataset limitations 49
5.2.2 ROI extraction for Unsupervised methods 49
5.2.3 Excluding Cases vs. Pseudo-Labeling 50
5.2.4 Selection of Loss function 50
5.2.5 Accuracy of extraction methods 50
5.3 Ethical Considerations 51

6 Conclusion **53**
6.1 Future Work 53

Bibliography **55**

A Appendix **I**
A.1 Additional Respiration Rate results I
A.2 Results per Fold III
A.2.1 Heart Rate results III
A.2.2 Respiration Rate results V

List of Figures

2.1 The skin reflection model that contains specular and diffuse reflections. The specular reflections are affected by motion variations while the diffused reflections contain pulsatile information [51]. 7
2.2 Attenuation of light while transmitted through different layers of tissue and blood [18]. 9

3.1 Typical processing components of conventional unsupervised algorithms for rPPG [61]. 13
3.2 Basic design of CNN-based NN models for rPPG [61]. 14
3.3 Frame sample from the PURE dataset [57]. 16
3.4 Frame sample from the UBFC-rPPG dataset [58]. 17
3.5 Frame sample from the SCAMPS dataset [59]. 17
3.6 Frame sample from the MR-NIRP dataset [56]. 18
3.7 Preprocessing Pipeline for the input videos of the NN models and the unsupervised methods. 19
3.8 Postprocessing Pipeline for the predicted and ground-truth PPG signals to extract the corresponding HR and RR. 20
3.9 Ground-truth PPG signals with large spans of zeros that indicate sampling error 22
3.10 Example PSD plots from Subject 12. The depicted samples were extracted from the recordings "subject12_garage_small_motion_975" and "subject12_driving_still_940" respectively. 22

4.1 Example plot of the predicted HR signal from each unsupervised method (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. 29
4.2 Example plot of the predicted HR signal from DeepPhys (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the DeepPhys model trained using fold 1 (3.4). The raw predicted signal is depicted in plot (a) and the standardized predicted signal in (b). 32

4.3 Example plot of the predicted HR signal from TS-CAN (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the TS-CAN model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels. 34

4.4 Example plot of the predicted HR signal from EfficientPhys-C (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the EfficientPhys-C model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels. . 35

4.5 Example plot of the predicted HR signal from PhysNet (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the PhysNet model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels. 37

4.6 Example plot of the predicted HR signal from PhysFormer (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the PhysFormer model trained using fold 1 (3.4). The raw predicted signal is depicted in plot (a) and the standardized predicted signal in (b). 38

4.7 Example plot of the predicted RR signal against the ground-truth signal (black) for ICA, POS, CHROM unsupervised methods. The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. 40

4.8 Example plot of the predicted RR signal against the ground-truth signal (black) for PhysNet, PhysFormer NN models. The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the models trained using fold 1 (3.4) and standardized. 41

5.1 Visualization of case specific HR estimation RMSE of POS, LGI unsupervised methods and all the supervised NN models when trained and evaluated on the the MR-NIRP dataset (Tables 4.2, 4.4). 46

5.2 Visualization of case specific RR estimation RMSE of ICA, POS, CHROM unsupervised methods and PhysNet, PhysFormer supervised NN models when trained and evaluated on the the MR-NIRP dataset (Tables 4.16, 4.18). 47

5.3 Visualization of the overall HR estimation RMSE of all the supervised NN models when trained on MR-NIRP, PURE, UBFC-rPPG, or SCAMPS and evaluated on the the MR-NIRP dataset (Table 4.3). 48

List of Tables

3.1 Basic architecture of each NN model. 15
3.2 Overview of the datasets utilized in this benchmark. 16
3.3 Table of excluded MR-NIRP recordings from this study. The entry subject12* indicates that all recording from this subject were omitted. 21
3.4 The subject IDs in every train-validation-test set for each fold. 23

4.1 Overall HR performance metrics of each unsupervised method evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. 28
4.2 Case specific HR performance metrics of each unsupervised method evaluated on the MR-NIRP subsets. 28
4.3 Overall HR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset and evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. 30
4.4 Case specific HR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset using 5-Fold cross-validation and evaluated on the MR-NIRP subsets. 30
4.5 Overall HR performance metrics of the DeepPhys model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 31
4.6 Case specific HR performance metrics of the DeepPhys trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 32
4.7 Overall HR performance metrics of the TS-CAN model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 33
4.8 Case specific HR performance metrics of the TS-CAN trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 33
4.9 Overall HR performance metrics of the EfficientPhys-C model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 34
4.10 Case specific HR performance metrics of the EfficientPhys-C trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 35

4.11 Overall HR performance metrics of the PhysNet model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 36
4.12 Case specific HR performance metrics of the PhysNet trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 36
4.13 Overall HR performance metrics of the PhysFormer model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 37
4.14 Case specific HR performance metrics of the PhysFormer trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. 38
4.15 Overall RR performance metrics of the unsupervised methods evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. The unsupervised methods GREEN, LGI, and PBV did not yield accurate RR predictions. 39
4.16 Case specific RR performance metrics of the unsupervised methods evaluated on the MR-NIRP dataset. The unsupervised methods GREEN, LGI, and PBV did not yield accurate RR predictions in any test case. 40
4.17 Overall RR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset and evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. DeepPhys, TS-CAN and EfficientPhys-C did not yield accurate RR predictions. . 41
4.18 Case specific RR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset using 5-Fold cross-validation and evaluated on the MR-NIRP subsets. DeepPhys, TS-CAN and EfficientPhys-C did not yield accurate RR predictions in any test case. 42
4.19 Average Throughput of NN models. 43

A.1 Overall RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset. I
A.2 RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset over the Driving test cases. I
A.3 RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset over the Garage test cases. II
A.4 Overall HR performance metrics of all the NN methods for each Fold. III
A.5 HR performance metrics of all the NN methods for each Fold over the Driving test cases. III
A.6 HR performance metrics of all the NN methods for each Fold over the Garage test cases. IV
A.7 Overall RR performance metrics of all the NN methods for each Fold. V

A.8 RR performance metrics of all the NN methods for each Fold over the Driving test cases. V
A.9 RR performance metrics of all the NN methods for each Fold over the Garage test cases. VI

1
Introduction

The majority of car accidents stem from drivers' sudden inability to control their vehicles, often caused by factors like fatigue, drowsiness, stress, distraction, or health related incidents [3][4]. Implementing systems that can detect these factors early on is crucial for preventing those accidents. A promising solution to address this issue involves integrating vital sign monitoring into vehicles to detect driver incapacitation and avert collisions. Furthermore, ensuring that the monitoring solution does not require direct physical contact with the driver is essential to avoid disrupting driving capabilities, thereby ensuring safety for both the driver and others on the road. In order to holistically asses a patient's condition, five vital signs are required: heart rate (HR), respiration rate (RR), blood pressure, body temperature, and blood oxygen saturation (SpO2) [19].

One of the most common approaches to achieve this task is through remote Photoplethysmography (rPPG). This non-invasive technique utilizes optical methods to detect subtle variations in blood volume within skin tissue, thereby enabling the monitoring of vital signs such as HR and RR using camera-based sensors. The application of rPPG in vehicles, however, faces unique challenges. Dynamic vehicle environments, characterized by motion artifacts and fluctuating illumination conditions can significantly impair the accuracy of rPPG readings [1]. Addressing these challenges is imperative for ensuring drivers' safety and health, particularly in situations where sudden health conditions might impair driving abilities. Machine learning-based algorithms for rPPG have shown promising results in detecting vital signs, indicating a potential pathway to improve vital sign detection in a laboratory setting [44] [45].

This thesis was conducted as part of a pre-study for developing Video-based Driver Condition Monitoring for Safe Driving (ViDCoM).

1.1 Statement of the Problem

The primary problem lies in the utilization of rPPG techniques within the dynamic vehicle environments. Since rPPG is a non-invasive method using camera sensors to detect blood volume changes in the skin tissue, it is highly susceptible to various external factors prevalent in vehicle settings.

Motion Artifacts: Vehicles inherently involve motion, both from the vehicle itself, like vibrations, accelerations, and decelerations and from the driver, such as head movements, facial expressions, and other physical activities. These motions can introduce significant artifacts into the rPPG signal, leading to inaccurate predictions.

Fluctuating Illuminations: The illumination within a vehicle is not consistent, impacting the accuracy of vital sign detection. Changes in external lighting conditions such as transitioning from tunnels to open roads, day-to-night changes as well as variations in internal cabin lighting can affect the quality of the optical signals captured by camera-based systems.

Environmental and Contextual Variability: Each vehicle and driving scenario can present a unique set of environmental factors, such as window tints, cabin designs, and driver positioning and skin complexity. These varying factors can further complicate the process of extracting accurate physiological signals from rPPG data.

Machine Learning (ML) models, particularly Neural Network (NN) approaches, have shown promising results in vital sign detection using rPPG in different environments, like indoors and hospitals [44][45]. However, their effectiveness in the context of dynamic vehicle environments remains largely unexplored and is yet to be extensively studied and validated.

1.2 Purpose of the Study

The contribution of this thesis lies in advancing the domain of camera-based, real-time vital sign monitoring within dynamic vehicle environments using ML approaches. The central objective is to investigate the effectiveness of ML techniques, particularly NN models, in accurately estimating Heart Rate (HR) and Respiration Rate (RR) using rPPG in autonomous vehicles. Specifically, our aim is to evaluate the performance of these models in a vehicle environment and compare the outcomes with the current state-of-the-art.

This thesis will focus on measuring their effectiveness in mitigating common vehicular challenges such as motion artifacts from driver activities and external vibrations, as well as variations in illumination due to changing external lighting and internal cabin lights.

1.3 Research Questions and Hypotheses

This thesis aims to explore the application of NN models for camera-based HR and RR detection in dynamic vehicle environments. The study is guided by the following

research questions (RQs) and hypotheses (Hs):

RQ1: How accurately can ML models for rPPG estimate HR and RR in the dynamic vehicle environment? This question seeks to investigate the feasibility of using rPPG ML models to detect HR and RR under the unique conditions of a vehicle environment.

RQ2: To what extent can state-of-the-art rPPG NN models, which are trained on generalized environments, maintain their accuracy and performance in automotive settings? This question aims to assess the generalizability of existing ML models trained on non-vehicular environments when evaluated in a vehicle setting.

H1: The NN models designed for HR and RR estimation in generalized environments demonstrate applicability in dynamic vehicle conditions, showcasing comparable performance.

H2: The pulse oximeter data provide the raw Photoplethysmography (PPG) waveform recordings such as those provided in the MR-NIRP dataset can be used to accurately estimate the HR and RR of the subject which can act as the ground-truth data for evaluating the algorithms.

1.4 Limitations and Delimitations

The study relies on existing datasets, which may not cover all possible driving scenarios or environmental conditions. The variety of vehicles, driver behaviors, and external factors like weather conditions in these datasets may limit the generalizability of the findings. Additionally, the quality and resolution of the video data in the dataset might affect the accuracy of HR and RR detection, particularly in low-light conditions or with low-resolution cameras. Furthermore, the quality of the ground-truth vital sign measurements can greatly impact the accuracy and reliability of the evaluation.

The scope of the study is narrowed down to focus exclusively on HR and RR extraction from rPPG, neglecting the exploration of other vital signs or non rPPG-based methods. This delimitation may restrict the holistic assessment of a driver's health status.

2
Background and Related work

The following chapter explores the various existing methods for monitoring vital signs within a vehicle environment, focusing on both contact-based and remote approaches. It also examines the specific processes involved in extracting specific vital signs from the signals collected via these methods. Moreover, it includes a review of the literature on various algorithms developed and tested for detecting vital signs remotely within vehicles.

2.1 Contact-based vital sign detection in vehicles

There are several methods that can be applied in a vehicle environment for detecting a driver's vital signs, including Electrocardiogram (ECG), ballistocardiography Ballistocardiography (BCG), and Photoplethysmography (PPG). ECG is a biopotential diagnostic procedure that with the help of electrodes, monitors the heart's electrical activity over a period of time [28]. For ECG monitoring in vehicle environments, the electrodes can typically be placed in steering wheels, car seats and backrest [19][26]. BCG is also a medical sensing method, typically used in car seats and backrests when it comes to vehicle environments. It utilizes pressure sensors that captures vibrations generated by the mechanical actions of the heart and lungs, along with the drive from the blood pulse moving through the aorta [19].

PPG is a method used for continuously monitoring different cardiovascular parameters, such as HR and RR. This is done by optically measuring the blood volume changes within tissue microvasculature with every heartbeat, also known as Blood Volume Pulse (BVP) [8][9][10][18]. This technology is affordable and easily accessible within consumer devices like smart wearables. These devices are equipped with sensors that gather PPG signals, which are linked to cardiac activity. This is accomplished through invasive techniques, specifically skin contact, involving several areas of the skin such as the wrist or fingertip [11].

Leonhardt et al. conducted a review of different methods that can be used for detecting vital signs in vehicles [19]. The main focus of the review was on placement of sensors in areas such as steering wheels, car seats, safety belts and cockpits. Sidikova et al. conducted a similar review over different methods for detecting vital signs in vehicle environments [26]. The study covered various systems including car

seat, steering wheel, helmet, camera, radar and other. The following paragraphs will detail some of the different techniques presented in these reviews.

Steering wheel sensors: Early techniques utilized electrodes on the surface of the steering wheel to measure ECG readings, enabling the assessment of drivers' stress levels [20][21]. Recent advancements include integrating sensors like conductive fabric for ECG monitoring in steering wheels [22][25]. Since frequent steering adjustments and steering with one hand is a common practice between drivers, steering wheel-based ECG monitoring is challenging. Due to its biopotential nature, ECG requires both hands to touch different conductive parts.

Another study measured vital signs in vehicle environments using a PPG sensor on the steering wheel. This sensor was placed on the left side of the horn, for the driver to be able to position their left thumb on it. The authors point out that this could also be a challenge when the driver changes the position of their hands on the steering wheel, since the contact between the thumb and the PPG sensor can be lost. This method also requires the driver to place their finger on the sensor manually meaning that it does not provide constant monitoring [29].

Car seat sensors: Methods based on ECG sensing in car seats and backrests for vital sign monitoring have evolved over the years [23][24]. Various designs have been proposed that integrate electrodes into car seats to track vital signs such as HR and RR. While these methods offer promising capabilities, challenges such as sub-optimal electrode placement, concerns about signal quality, and limitations due to motion artifacts and triboelectricity still remain [19].

Apart from ECG monitoring integrated into car seats, efforts have also been made to incorporate BCG sensors in car seats [30][31]. However, the main challenge for BCG monitoring is motor vibrations while driving, which can negatively affect the sensor's measurements.

2.2 Remote vital sign detection in vehicles

An alternative method for PPG signal collection is rPPG, which involves extracting plethysmographic information from conventional video cameras through the RGB channels. As the skin is illuminated, its color changes over time due to motion-induced intensity variations and subtle color changes caused by the pulse (Figure 2.1). By analyzing color variations in the skin captured by the camera, rPPG can extract BVP in a non-invasive manner [51]. However, the accuracy of rPPG relies heavily on stable skin locations or precise tracking, as even minor movements can significantly affect performance [12].

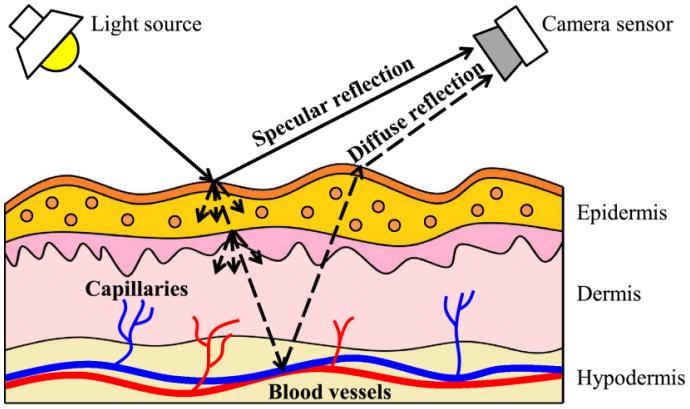


Figure 2.1: The skin reflection model that contains specular and diffuse reflections. The specular reflections are affected by motion variations while the diffused reflections contain pulsatile information [51].

Other non-intrusive techniques for vital sign monitoring in vehicle environments are radar and optical flow-based methods. The term radar originates from "radio detecting and ranging" representing a technique that uses high-frequency electromagnetic waves emitted from a transmitter. These waves bounce of the surface of the human chest and carries back the information of chest displacements and the inner organ movements to a receiver [19]. The optical flow is camera-based method used to estimate motion within a video sequence using a Convolutional Neural Network (CNN). In the case of vital sign detection, this can be used to measure RR by monitoring the chest motion from the respiration cycles [39]. A distinct advantage of this method over rPPG, that is mainly based on color channels, is that it is adaptable to various camera types ,including infrared and thermal, making it impervious to illumination changes.

In the subsequent sections, radar, optical flow and rPPG based methods will be explored, including specific examples of studies that employ these techniques in a vehicular environment.

2.2.1 Radar-based methods

A study conducted by Vinci et al. introduced the idea of employing continuous wave radar sensors to monitor HR and RR while driving. They explored two measurement setups: firstly, positioning the radar system within the steering wheel, directed towards the thorax. Secondly, they placed the same radar sensor on the car seat's backrest, directed towards the back [41]. Another study positioned the radar sensor towards the driver's back in the backrest for HR detection [42]. Similarly in 2018, Schires et al. mounted a miniaturised ultra-wideband pulsed radar in the backrest for monitoring the motion of the heart and lungs, in order to measure HR and RR [43]. All three studies concluded that motion artifacts could pose challenges in extracting displacement signals using radars, which would impact the accuracy of the vital sign measurements.

2.2.2 Optical flow-based methods

In 2023, Othman et al., proposed an optical flow-based method to estimate RR in a vehicle environment, by utilizing the Openpose human pose estimation model and SelFlow NN. This approach managed to measure RR effectively at low vehicle speeds and displayed remarkable robustness against illumination changes. However, its accuracy declined significantly at higher velocity due to noise introduced from increased driver motion and external vibrations [6].

2.2.3 rPPG-based methods

In 2017, Wu et al. developed a rPPG algorithm for monitoring HR in vehicle environments using RGB videos [32]. For HR estimation, a k-nearest neighbor classifier was employed based on frequency domain features. A few years later the same authors further developed their previous study to improve the algorithm for outdoor driving HR estimation by including more outdoor driving data [33]. Another research based on RGB data was conducted by Huang et al., they introduced a novel transformer-based algorithm with promising results, but they mentioned that it is prone to overfitting [35]. Additionally, they created a custom dataset, taking into account motion and illumination artifacts. However, this dataset is not available to the public yet.

In 2019, Hernandez-Ortega et al. tried to improve the extraction of the rPPG signal from a baseline HR estimation system based on Near-Infrared (NIR) imaging by integrating a quality-based approach [34]. In 2022, Nowara et al. proposed the AutoSparsePPG algorithm that leverages quasi-periodicity to improve rPPG signal estimation over its predecessor, SparsePPG [56]. Additionally, the paper introduced the publicly available MR-NIRP Car dataset with synchronized NIR and RGB video recordings, along with ground-truth pulse oximeter measurements. The MR-NIRP Car dataset was used to evaluate the performance of the proposed algorithm in detecting HR along with a range of other unsupervised rPPG methods. In 2023, Xu et al. developed Ivrr-PPG, a method aimed to address illumination challenges and extract a clean signal from rPPG [36]. The core idea of their approach is centered on capturing the nonlinear relationship between changes in illumination and the rPPG signal using NIR cameras.

Wang et al. conducted a benchmark study that also utilized the MR-NIRP Car dataset to evaluate four types of rPPG methods in HR estimation [40]. Specifically, the study explored the performance of two unsupervised physiological based methods and two deep learning models. For the deep learning models they used modified versions of the PhysNet model, proposed by Gideon et al. [64]. This study did not extensively investigate other algorithms, particularly limiting its coverage of deep learning methodologies. Another study proposed by Chiu et al. used two distinct CNN-based models for rPPG construction and HR estimation, with evaluation conducted using MR-NIRP car dataset and other custom datasets that are not publicly available [38]. However, the results reported for most algorithms do not match those

reported in other studies indicating some irregularity in the testing process.

NIR-based methods have demonstrated superior handling of illumination artifacts compared to RGB-based methods, due to their ability to capture signals unaffected by visible light variations. However, NIR methods inherently carry less information than RGB-based approaches due to their single-channel nature. Finally, it is worth noting that all the studies presented in this subsection focus exclusively on evaluating the algorithms on HR detection and did not explore other vital signs.

2.3 Deriving vital signs from PPG signal

PPG sensors use optical methods to monitor fluctuations in blood flow volume within the tissue's microvascular bed. This PPG signal can be used to extract various vital signs, serving as a valuable tool for identifying cardiovascular diseases and other medical conditions. Such metrics include HR, Heart Rate Variability (HRV), RR and SpO2.

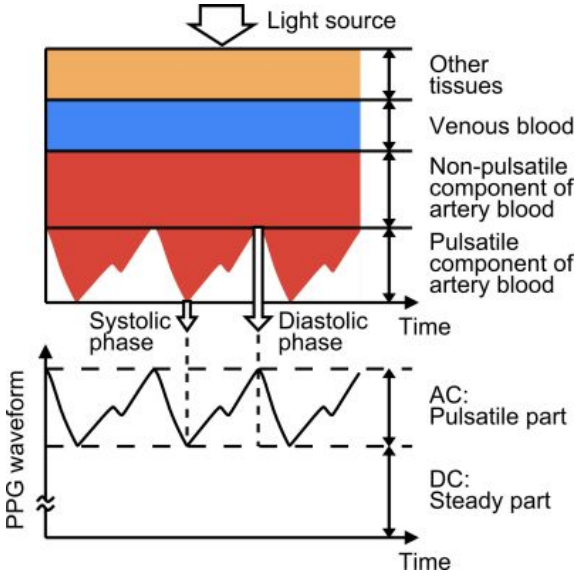


Figure 2.2: Attenuation of light while transmitted through different layers of tissue and blood [18].

Heart Rate and Heart Rate Variability: When the heart beats, blood gets pumped into the arteries and capillaries, increasing their blood volume. As a result, PPG waveform features peaks and valleys corresponding to the systolic and diastolic phases of the cardiac cycle (Figure 2.2). By analyzing the time interval between successive peaks in the PPG waveform, it is possible to calculate the HR also known as Inter-Beat Interval (IBI) [10]. Furthermore, variations between consecutive IBIs measure the HRV [16].

Respiration Rate: Respiration influences the PPG signal in three distinct manners. Firstly a phenomenon known as respiratory sinus arrhythmia (RSA) causes HR to align with the breathing cycle due to an autonomic reaction to breathing. This results in the HR rising with inhalation and falling with exhalation, leading to variations in the PPG signal's frequency, known as respiratory-induced frequency variations. Secondly, fluctuations in intrathoracic pressure during the respiratory cycle prompt a shift of blood between the pulmonary and systemic circulations, altering the signal's baseline intensity, a mechanism called respiratory-induced intensity variations. Lastly, a decrease in cardiac output, linked to diminished ventricular filling, affects the strength of the peripheral pulse. This effect, known as respiratory-induced amplitude variations, reflects changes in the amplitude of the PPG waveform. Different algorithms have been proposed for extracting RR from PPG by leveraging one of or a combination of these phenomena [13].

Blood Oxygen Saturation: The measurement of SpO2 through pulse oxymetry is based on the principle that haemoglobin absorbs more red light while oxyhaemoglobin has higher infrared absorbtion. The PPG signal is split into two elements: the AC component, which originates from the pulsatile arterial blood, and the DC component, attributed to tissue background, venous blood, and the steady portion of arterial blood flow (Figure 2.2). These AC and DC segments are then utilized to determine the SpO2 level in the blood. It should be noted that to measure SpO2, two PPG sensors are required, one red and one infrared [14][15][17].

In this thesis, we are conducting a benchmark study of NN-based rPPG methods, focusing only on HR and RR estimation. The majority of current literature are only focusing on extracting HR from rPPG, while RR and other vital signs are very under-represented. Moreover, within the context of vital sign detection in vehicle environments, there has been minimal attention given to NN-based approaches. The existing literature lacks comprehensive benchmarking, particularly in comparing diverse NN-based methodologies for vital sign monitoring in a vehicle environment. Consequently, this thesis aims to address this gap by providing a thorough evaluation and comparison of various NN-based approaches using rPPG for HR and RR detection in vehicle environments.

3
Methodology

The aim of this benchmark study is to evaluate the adaptability and performance of various machine learning models in estimating HR and RR from rPPG signals in a vehicular environment. This environment introduces unique challenges, such as motion artifacts, lighting variations, and background noise, which can significantly impact the accuracy and reliability of rPPG signal processing. By comparing different types of unsupervised physiological-based algorithms and supervised NNs this study seeks to uncover which methodologies are most effective in overcoming these challenges and maintaining high performance in vital sign estimation in automotive settings. The comparison aims not only to benchmark the current state-of-the-art but also to identify key factors that contribute to the success or limitations of each approach in real-world settings.

Furthermore, a crucial component of this study is the establishment of a comprehensive preprocessing framework tailored for rPPG signal analysis. This framework aims to standardize the initial stages of data handling, ensuring that the input data across all tested models maintain a consistent quality and format that facilitates fair comparison. This comparability is essential for accurately assessing the relative performance of each algorithm under identical conditions, thereby ensuring that any observed differences in performance are attributable to the models' inherent characteristics rather than variations in data preparation. A well-defined preprocessing framework, also, enhances the reproducibility of the research by providing clear guidelines for data preparation. This ensures that researchers and software engineers can reliably replicate and validate studies, assess methodologies, and benchmark against the current state-of-the-art. Ultimately, this increases confidence in applying these findings to real-world scenarios while maintaining consistent methodological standards.

3.1 Benchmark Study

A benchmark study within technical research compares the performance of different tools or techniques using representative tests and performance measures. It comprises three essential components: motivating comparison, task sample, and performance measures. Motivating comparison outlines the comparison to be made and the research agenda to be advanced through the benchmark. Task sample involves selecting representative tasks relevant to the tool or technique's intended use.

Performance measures gauge the effectiveness of the technology in meeting its purpose, whether quantitatively or qualitatively [60].

The benchmark research methodology is well-suited for this thesis as it provides a structured framework for comparing the performance of various algorithms against established state-of-the-art methods.

In this case, the motivating comparison lies in evaluating the performance of ML algorithms in predicting rPPG signals that can be used to estimate HR and RR within the dynamic and challenging environment of autonomous vehicles. This comparison aims to advance the research agenda by exploring the feasibility and efficacy of integrating advanced ML algorithms, designed for controlled laboratory settings, into autonomous vehicle systems for vital sign monitoring purposes. The task sample selection is crucial in benchmark studies, as it ensures that the tasks chosen are representative of real-world scenarios and challenges. In the context of this project, these tasks should encompass variations in lighting conditions, vehicle motion, subject movement, and other factors that may affect the accuracy of rPPG-based HR and RR estimation. Finally, the performance measures need to capture the algorithms' effectiveness, accuracy, and adaptability in predicting heart and respiration signals in realistic vehicular environments.

This approach will enable a comprehensive understanding of how a deep learning rPPG-based HR and RR detection in a vehicle environment framework is (a) designed, (b) realized, and (c) systematically evaluated within professional settings. The research will generate empirical data and insights crucial for the design and development of robust and effective camera-based health monitoring systems, thereby advancing the integration of real-time physiological monitoring in vehicle software. Moreover, this study provides developers with systematic methodologies and introduces performance benchmarks. In software engineering, a benchmark study is crucial as it establishes a standard against which the performance of various software systems can be evaluated and compared.

3.2 rPPG Algorithms

The rPPG algorithms can be broadly categorized into two main approaches: supervised NN models and conventional unsupervised methods. The former employ deep learning techniques to interpret subtle changes in facial videos, adapting to various conditions such as motion and lighting variations. The later utilize statistical and signal processing methods to extract vital signs from video by analyzing the color changes caused by blood volume fluctuations in the skin. Different methods perform a combination of color transformations and signal decomposition techniques on the RGB signals to construct an estimate of the BVP (Figure 3.1).

3.2.1 Unsupervised methods

Conventional unsupervised rPPG algorithms generate temporal RGB signals by calculating the average red, green, and blue values over time from spatially averaged RGB values of skin pixels. These signals are often transformed from RGB into different color spaces to enhance the BVP signal and suppress noise. After transformation, advanced signal processing techniques are applied to further refine the signals. These might include methods like Independent Component Analysis (ICA) or bandpass filtering. These techniques help to decompose the signals into components, isolating the one that most closely represents the BVP [51][61].

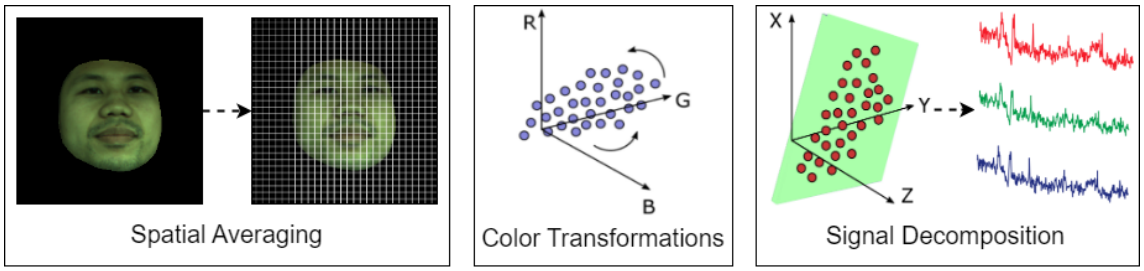


Figure 3.1: Typical processing components of conventional unsupervised algorithms for rPPG [61].

This benchmark study explores the performance of six unsupervised methods for rPPG with different architectures.

ICA: ICA is a computational method used to separate a multivariate signal into independent components [49]. Poh et al. introduced a method that leverages ICA to separate temporal RGB color signals into either uncorrelated or independent signal sources in order to extract the BVP [50].

POS: The Plane-Orthogonal-to-the-Skin method computes a projection plane that is orthogonal to the skin tone, utilizing physiological and optical principles. This approach involves applying a fixed matrix projection to the spatially normalized and averaged pixel values, which facilitates the extraction of the rPPG signal [51].

CHROM: A method proposed by de Haan et al. that estimated the BVP by a linear combination of the chrominance signals obtained from the RGB video [52].

GREEN: This method is based on the principle that the green channel, due to its mid-spectrum wavelength, has a higher absorption contrast with blood than the red and blue channels. This makes it more sensitive to changes in blood volume under the skin. As a result, BVP can be extracted from the green channel after spatial averaging of the RGB video [53].

LGI: Local Group Invarince (LGI) is a feature representation technique designed to be invariant to motion by employing differentiable local transformations. Pilz et al. proposed a method that leverages LGI to extract BVP reliably in dynamic

environments despite motion artifacts [54].

PBV: Another method introduced by de Haan et al., the Blood Volume Pulse Vector (PBV) is based on the understanding that blood pulsations affect the RGB channels in a PPG signal to varying degrees. By representing these changes with a unique signature or vector, allows for a more accurate isolation of the pulsatile component of the blood volume signal. This process helps in removing any alterations in color that are not related to blood movement enabling accurate extraction of the BVP signal [55].

3.2.2 Supervised Neural Network models

The vast majority of existing supervised NN models for rPPG adopt a CNN-based architecture. These models often utilize an attention mechanism to learn and extract the necessary regions of interest (ROIs) which then pass through a collection of convolutional and fully connected layers to extract the BVP (Figure 3.2).

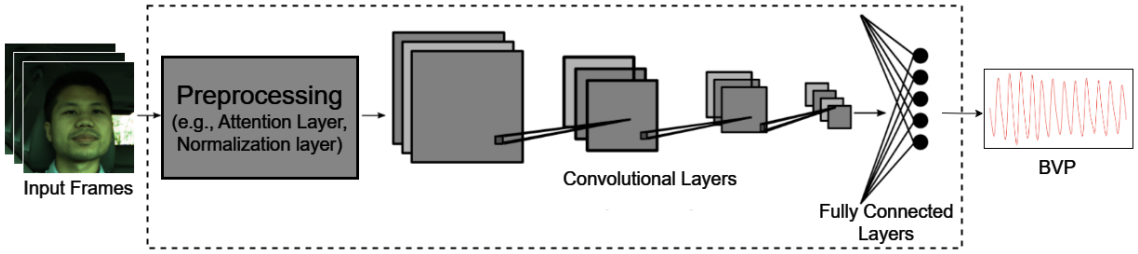


Figure 3.2: Basic design of CNN-based NN models for rPPG [61].

This benchmark explores the performance of four CNN-based NN models and one transformer-based architecture.

DeepPhys: Chen et al. introduced DeepPhys, the first end-to-end system for non-contact video-based physiological measurement for HR and RR. DeepPhys is a two-branch 2D convolutional attention network that incorporates a motion representation based on a skin reflection model and an attention mechanism utilizing appearance information [44].

TS-CAN: Liu et al. designed a 2D temporal shift convolutional attention network called TS-CAN that utilized a Temporal Shift Module (TSM) to better capture spatio-temporal information. The architecture of TS-CAN consists of two branches that handles a distinct representation: one for motion modeling and the other for capturing meaningful spatial features (i.e., appearance). These branches operate concurrently, with the appearance branch directing the motion branch through a gated attention mechanism [45].

EfficientPhys-C: Liu et al. introduced an additional 2D-CNN with a TSM de-

signed for on-device computations in real-time. In contrast to TS-CAN and DeepPhys, the architecture of EfficientPhys-C is built on a single branch that incorporates a normalization module for computing differences in frames and learnable normalization. Additionally, it includes a self-attention module which enables the model to focus on skin pixels relevant to the PPG signal [47].

PhysNet: Yu et al. introduced an end-to-end spatio-temporal network called PhysNet, which employs a 3D-CNN that captures semantic rPPG features across both spatial and temporal dimensions concurrently. This approach enhances the learning of robust contextual features and aids in the restoration of rPPG signals with reduced temporal fluctuations [46].

PhysFormer: PhysFormer employs a video transformer-based design, focusing on the dynamic combination of both local and global spatio-temporal elements to improve rPPG signal representation. This architecture prioritizes the extraction of long-term, global characteristics, which results in notable enhancements in performance when measured against various alternative methods [48].

Table 3.1: Basic architecture of each NN model.

NN Model:	**Architecture**
DeepPhys	two-branch 2D-CNN
TS-CAN	two-branch 2D-CNN with TSM
EfficientPhys-C	single-branch 2D-CNN with TSM
PhysNet	single-branch 3D-CNN
PhysFormer	Transformer

The mentioned NN models utilized distinct architectures (Table 3.1) and demonstrated performance comparable to the state-of-the-art for rPPG detection in generalized settings. However, the datasets used in these studies did not involve real-world automotive settings, where other challenges and conditions could impact their performance.

3.3 rPPG Datasets

Datasets play a pivotal role in training and evaluating rPPG algorithms. These collections feature video recordings of human subjects under controlled or realistic environments with varied lighting, movements, and physiological states. Each recording is paired with accurate physiological measurements obtained from conventional contact-based sensors enabling precise validation of rPPG algorithms. Table 3.2 presents a brief overview of the datasets utilized in this benchmark study, with detailed description provided in the subsequent sections.

Table 3.2: Overview of the datasets utilized in this benchmark.

Dataset	Subjects	Videos	Ground-Truth	Context
PURE [57]	10	59	PPG, SpO2	Indoor setting Natural lighting Motion (head rotation, talking, etc.)
UBFC-rPPG [58]	42	42	PPG, HR	Indoor setting Natural and artificial lighting Motion (head rotation, talking, etc.)
SCAMPS [59]	-	2800	PPG, HR, RR, HRV	Synthetic dataset Simulated realistic setting Simulated dynamic lighting Simulated motion (head rotation, facial actions)
MR-NIRP [56]	19	190	PPG	Vehicle setting Natural dynamic lighting Motion (head rotation, talking, vehicle motion, etc.)

3.3.1 General Datasets

The vast majority of datasets are composed in indoor settings with stable lighting to maintain high video fidelity and quality of the ground-truth physiological measurements. Subjects often perform small movements to simulate realistic conditions such as speech and head motion. Additionally, there exist some synthetic datasets with computer generated videos of human-like subjects and corresponding artificially generated physiological data.

PURE: The Pulse Rate Detection Dataset (PURE) contains recordings from 10 individuals, which include 8 males and 2 females. These recordings were made using an RGB camera, capturing video at a frame rate of 30 Hz and a resolution of 640x480 pixels. The subjects were situated in front of a realistic non-uniform indoor setting and were lit by ambient natural light coming through a window (Figure 3.3). The ground-truth measurements for PPG and SpO2 levels were collected at a frequency of 60 Hz using a pulse oximeter attached to the finger of the subject. Each participant underwent six recording sessions under different motion scenarios, providing a diverse set of data across various physical conditions [57].

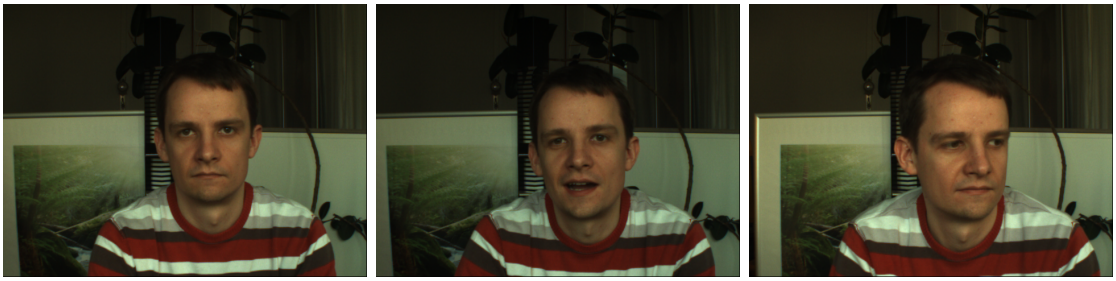


Figure 3.3: Frame sample from the PURE dataset [57].

UBFC-rPPG: The University Bourgogne Franche-Comté rPPG (UBFC-rPPG) Dataset consists of RGB videos captured with a webcam at a frame rate of 30 Hz. The videos have a resolution of 640x480 pixels and are saved in an uncompressed

8-bit RGB format. Ground-truth PPG data was acquired using a pulse oximeter. During the recording sessions, subjects were situated roughly one meter from the camera. The videos were recorded indoors in front of a uniform and static background, with lighting provided by a mix of natural sunlight and artificial light sources (Figure 3.4). The subjects were recorded while playing a game to simulate realistic motion patterns [58].

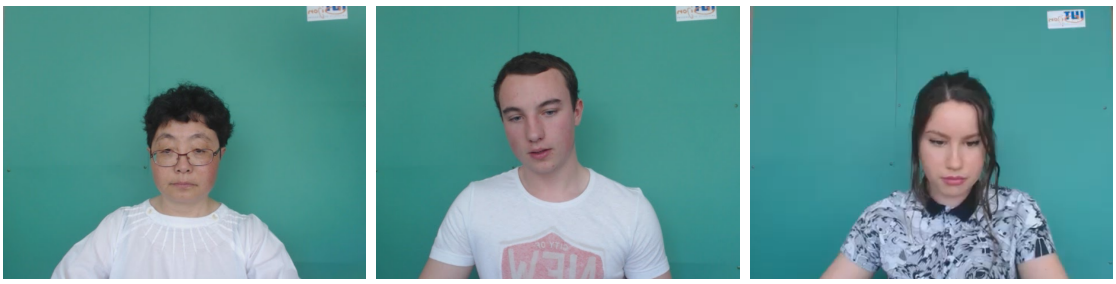


Figure 3.4: Frame sample from the UBFC-rPPG dataset [58].

SCAMPS: Synthetics for Camera Measurement of Physiological Signals Dataset (SCAMPS) is a synthetic dataset that contains 2800 videos, totaling 1.68 million frames, synchronized with cardiac and respiratory signals. The videos and ground-truth waveforms were produced using an advanced facial processing pipeline, achieving high-fidelity and near photorealistic visuals. To ensure robust testing conditions, the videos include a range of variables such as head movements, facial expressions, and shifts in ambient lighting, as well as, a simulated realistic background environment (Figure 3.5) [59].

Figure 3.5: Frame sample from the SCAMPS dataset [59].

3.3.2 Dataset in vehicle environment

Remote PPG detection while driving faces distinct challenges from those encountered in other settings. Consequently, datasets gathered in different scenarios do not provide relevant insights for analyzing rPPG detection in vehicles. The vast majority of rPPG datasets available to the public were obtained in indoor environments using RGB cameras under controlled lighting conditions, which are rarely present in a realistic vehicular setting. These discrepancies render existing general purpose

datasets inadequate for exploring the specific challenges associated with rPPG measurement in automotive environments. There have been a limited number of studies attempting to capture rPPG data in vehicles using camera technology, and to date, only one provides a publicly available dataset [58][57].

MR-NIRP: The MERL-Rice Near-Infrared Pulse Car Dataset (MR-NIRP) consists of recordings from 18 healthy individuals, featuring 16 males and 2 females, aged between 25 and 60 years, with diversity in facial features and skin tones. The recordings were captured using both NIR and RGB (10-bit raw) cameras at a frame rate of 30 Hz and resolution of 640x640 pixels. Ground-truth PPG signals were obtained using a finger pulse oximeter, recorded at 60 Hz. It should be noted that for safety reasons and to maintain the quality of the PPG signal the subjects were situated in the passenger seat during recording [56].

The videos in the MR-NIRP dataset were captured under two primary settings: inside a garage with the engine running, and while driving in city conditions, performing stops at traffic lights and maneuvers such as sudden stopping, accelerating, and turning. For the driving scenario, videos were recorded during different natural lighting conditions, including daytime with sunny and overcast weather, as well as nighttime recordings. Additionally, the dataset is separated based on three motion conditions: still, where the subjects were sitting quietly with minimal head motion, small motion, with the subjects performing natural head movements and talking, and large motion, with additional and more abrupt head movements (Figure 3.6).

Figure 3.6: Frame sample from the MR-NIRP dataset [56].

3.4 Experimental Setup

For this benchmark, in vehicle setting, we employed the MR-NIRP car dataset, since it is the only dataset publicly accessible that is specific to vehicular environments. The analysis encompassed still, small and large motion scenarios, conducted within garage and driving contexts.

To train and evaluate the performance of the described models on the MR-NIRP car dataset, we used the rPPG-Toolbox, an open-source platform for training and benchmarking NN-based and unsupervised rPPG algorithms [61]. Additionally, the rPPG-Toolbox provides pretrained weights for the supervised NN models in this benchmark (Section 3.2.2), that are trained on the general-purpose datasets (Section 3.3.1). These pretrained models were used for the evaluation on MR-NIRP.

3.4.1 Preprocessing

To ensure a fair evaluation of all tested algorithms, we implemented a simple uniform preprocessing pipeline with minimal adjustments to meet the specific input requirements of each method. A visualization of the preprocessing pipeline can be found in Figure 3.7.

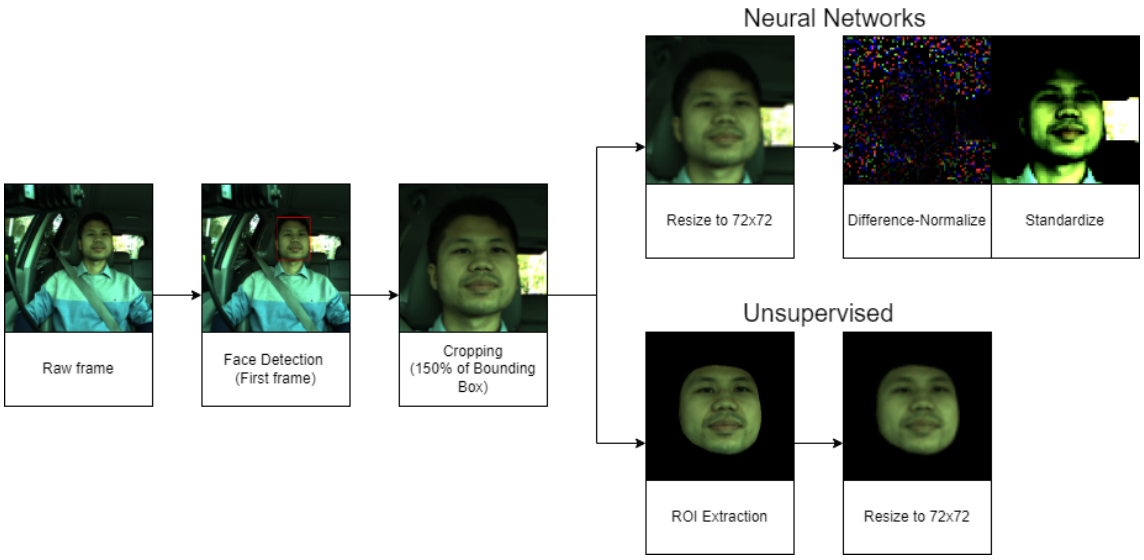


Figure 3.7: Preprocessing Pipeline for the input videos of the NN models and the unsupervised methods.

For each video, face detection was performed on the first frame using RetinaFace, a robust single-stage face detection framework that leverages multi-task learning to perform pixel-wise face localization [62]. This was followed by cropping a rectangular region scaled to 150% of the detected bounding box. Given that the subjects exhibited limited movement, the 50% scale increment was sufficient to ensure the subject's face remained within the cropped frame across all video frames. Subsequently, the cropped video frames were resized to 72x72 pixels. For Physformer, the frames were resized to 128x128 pixels in order to match the input parameters used by the model's authors as well as the pretrained versions of the rPPG-Toolbox [48][61]. For the labels, the ground-truth PPG signal was downsampled from 60 Hz to 30 Hz in order to match the frame rate of the videos.

For the Unsupervised methods, the MediaPipe FaceMesh was also applied to detect landmarks of the subject's face on each image frame [63]. For this study the

whole face was selected as the Region of Interest (ROI) meaning that all 478 landmark indices were included. The area within the convex hull polygon enclosing the chosen landmarks was extracted and used as the ROI. For the NN methods, there was no need for ROI extraction since all tested algorithms contain a trainable attention mechanism which can independently learn to extract the desired ROIs.

After that, the input of the NN models was difference-normalized (c_{norm}), meaning that the difference of every two consecutive frames and labels (c) was calculated, and normalized by their standard deviation (σ).

$$c_{norm} = \frac{c(t+1) - c(t)}{\sigma(c(t+1) - c(t))}$$

The use of difference-normalized frames as the models' input helps minimize reliance on the overall brightness of frames and the subject's skin tone [45]. EfficientPhys-C was exempted from this preprocessing step as it already contains a normalization module that calculates frame differences in its architecture. Instead, the standardized frames were passed as its input while the labels were difference-normalized as usual. Additionally, since DeepPhys and TS-CAN have two branches, they accept difference-normalized frames on the motion branch and standardized frames on the appearance branch.

3.4.2 Postprocessing

The predicted rPPG signal was processed identically for all algorithms to extract the vital signs. It is important to note that if the signal is difference-normalized, it is the first derivative of the PPG signal. In that case, the cumulative sum of its values needs to be calculated to retrieve the rPPG signal before the postprocessing. A visualization of the postprocessing pipeline can be found in Figure 3.8 [61].

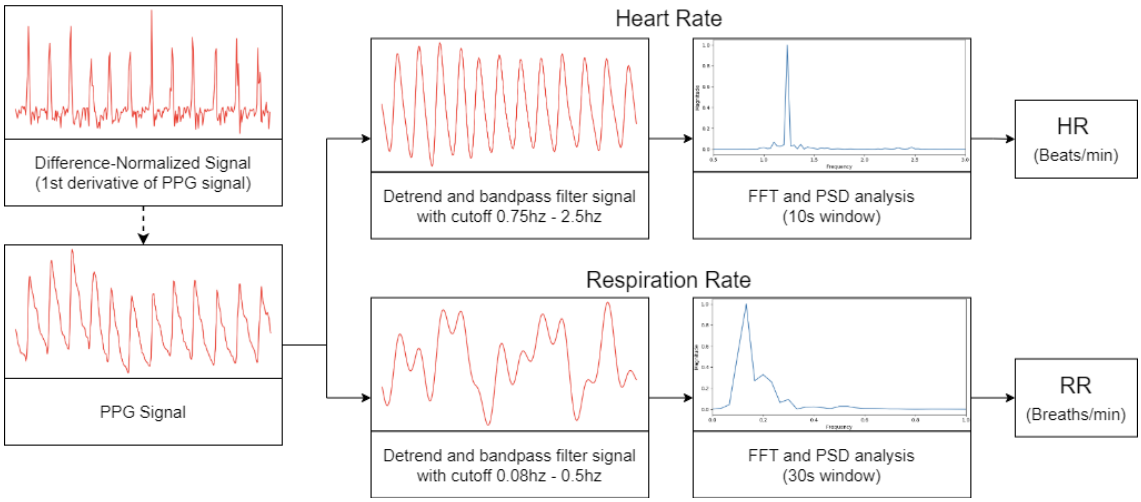


Figure 3.8: Postprocessing Pipeline for the predicted and ground-truth PPG signals to extract the corresponding HR and RR.

Both the predicted and ground-truth signals were detrended with a fixed lambda value of 100, then filtered using a second order butterworth bandpass filter to remove noise. The selection of the filter frequencies is dependent on the application and nature of the data. Therefore, to extract HR the lower and upper cutoff frequencies of the filter were set to 0.75 Hz and 2.5 Hz respectively corresponding to the normal pulse frequency in healthy adults of 45 to 150 Beats per Minute (Beats/min). For RR extraction the cutoff frequencies were set to 0.08 Hz and 0.5 Hz covering the normal adult RR of 5 to 30 Breaths per Minute (Breaths/min). It should be noted that to achieve the stated cutoff, the actual frequencies used in the filter were doubled in order to prevent aliasing, adhering to the Nyquist–Shannon sampling theorem [45][65].

The filtered signal was then transformed into the frequency domain via Fast Fourier Transform (FFT). Power Spectral Density (PSD) analysis was used to identify the highest power frequency which corresponds to the predicted vital sign. The extraction of HR was done in 10 second windows with no overlap while for RR a 30 second window with no overlap was used [61][45].

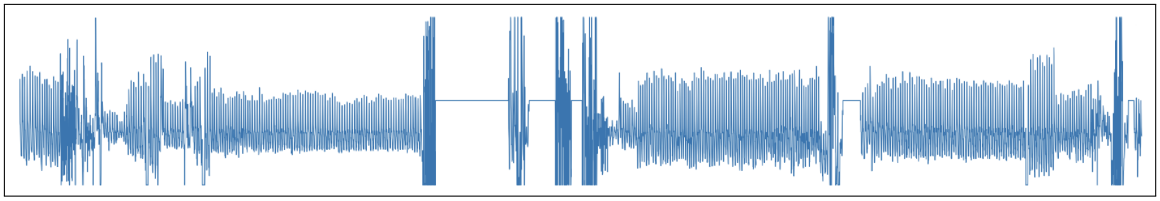
3.4.3 Excluded Cases

Certain cases of the MR-NIRP dataset were omitted from the study due to factors that could potentially affect the benchmark (Table 3.3).

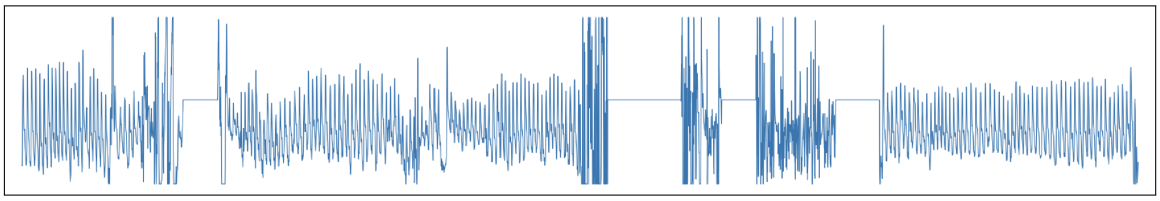
Table 3.3: Table of excluded MR-NIRP recordings from this study. The entry subject12* indicates that all recording from this subject were omitted.

Justification	Excluded Cases
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
PPG Sampling Error	subject7_driving_small_motion_975 subject7_driving_still_975 subject12*

Specifically, five videos were too dark to detect the subject's face while one case contained corrupted video frames. Additionally, in two cases the recorded ground-truth PPG signals had large spans of zeroes indicating some kind of error in the sampling process (Figure 3.9).



(a) subject7_driving_still_975



(b) subject7_driving_small_motion_975

Figure 3.9: Ground-truth PPG signals with large spans of zeros that indicate sampling error

Furthermore, all samples from Subject 12 were excluded from the study as PSD analysis of the ground-truth signal showed excessive levels of noise, indicating potential sampling error (Figure 3.10). In addition to that, the extracted HR from these PPG signals was consistently lower than 50 Beats/min which is not indicative of a healthy adult male like Subject 12. This phenomenon was observed in most recordings of this subject, making them unsuitable for accurate ground-truth vital sign extraction [66].

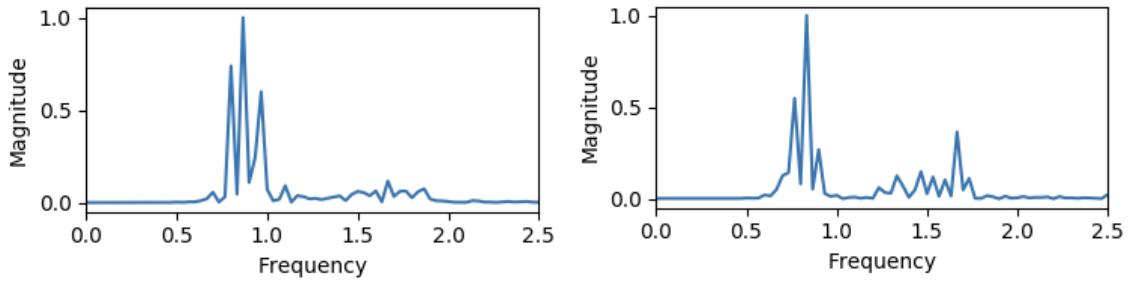


Figure 3.10: Example PSD plots from Subject 12. The depicted samples were extracted from the recordings "subject12_garage_small_motion_975" and "subject12_driving_still_940" respectively.

3.5 Training Setup

To achieve consistent and reproducible results across our tests, this benchmark was conducted using uniform hardware configuration, featuring a single Nvidia Tesla

V100 GPU with 32GB of memory. However, as highlighted in [61], the pretrained models included in the rPPG-Toolbox were not trained with fixed hardware specifications. Despite that, all test results were obtained using our predefined GPU setup.

3.5.1 Dataset Folds

In order to evaluate the models' performance in vehicular environments effectively, it was necessary to use the MR-NIRP dataset for both training and testing. Therefore, to avoid overfitting and ensure unbiased predictions, distinct train-validation-test splits were required. Due to the absence of a set of folds for training and testing on the MR-NIRP dataset, we followed the procedure described by Gideon et al. [64]. Specifically, the dataset was split into 5 folds by subject ID with a different held-out test set each time (Table 3.4). All models were trained and tested on every fold with the average of the results across all folds being reported in this benchmark.

Table 3.4: The subject IDs in every train-validation-test set for each fold.

	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

For intra-dataset testing, there was no need for a distinct test set from the MR-NIRP dataset since the evaluation was performed on different datasets. In that case, the entire MR-NIRP dataset was utilized for training, with a simple 80-20 split for the training and validation sets and no folds were used.

3.5.2 Training Parameters

To facilitate a fair comparison between the pretrained models provided by the rPPG-Toolbox and our own, similar training parameters were adopted. This included the use of the AdamW optimizer (Adam optimizer for PhysNet) and the negative Pearson loss function. Additionally, we employed a one-cycle learning rate scheduler with a peak learning rate of 0.009 for all models. Training was carried out over 30 epochs, selecting the model with the lowest validation loss at the end of each epoch. Moreover, a batch size of 4 was used for all experiments and 20% dropout was applied to avoid overfitting.

The training configuration for PhysFormer diverged from the conventional setup due to its distinct architecture, and it adhered to the experimental details laid out by the authors of the model. Specifically, for PhysFormer, the Adam optimizer was

employed with an initial learning rate of 0.0001 and weight decay of 0.00005 with no learning rate scheduler. Instead of negative Pearson loss, a dynamic loss implementation was utilized that incorporates negative Pearson loss, frequency cross-entropy loss, and label distribution loss. Additionally, the model's parameters were set as follows: $N = 12$, $h = 4$, $D = 96$, $D' = 144$, matching the configuration used on the pretrained models provided in the rPPG-Toolbox [48][61].

3.6 Evaluation Metrics

Evaluating the effectiveness of rPPG algorithms in realistic automotive environments requires an assessment of both their accuracy and efficiency. The following evaluation metrics were chosen to ensure that the tested algorithms accurately predict vital signs under dynamic vehicular conditions and are viable for real-time applications. These metrics offer a holistic view of the models' performance in practical scenarios and their suitability for deployment in realistic automotive settings.

3.6.1 Performance Metrics

The selection of performance metrics in this study was guided by the goal of providing a comprehensive evaluation of the adaptability and accuracy of the tested algorithms in estimating HR and RR from rPPG signals within a dynamic vehicular environment. Therefore, five metrics were selected, each addressing distinct aspects of algorithm performance, from error rates to signal quality.

Mean Absolute Error (MAE): Calculates the average of the absolute differences between the predicted (R_{Pred}) and ground-truth (R_{GT}) signal rate (HR or RR) over all observation windows (T). This simple metric provides an intuitive understanding of the prediction accuracy without considering the error direction. It is particularly suited to health monitoring applications where small prediction deviations can have significant implications.

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

Root Mean Square Error (RMSE): Measures the magnitude of the prediction error between the predicted (R_{Pred}) and ground-truth (R_{GT}) signal rates over all observation windows (T). By penalizing larger errors more severely than smaller ones, RMSE offers insight into the reliability and consistency of the models. This characteristic is crucial for rPPG technologies in dynamic environments like vehicles, where outliers can drastically affect performance reliability.

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

Mean Absolute Percentage Error (MAPE): Calculates the average of the absolute differences between the predicted (R_{Pred}) and ground-truth (R_{GT}) signal rates, expressed as a percentage of the ground-truth signal rate over all observation windows (T). MAPE quantifies the error as a percentage, making it a scale-independent measure of the prediction accuracy.

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

Pearson Correlation Coefficient (ρ): A statistical measure that calculates the linear correlation between the predicted (R_{Pred}) and ground-truth (R_{GT}) signal rates across all windows (T). This measure quantifies the algorithm's precision and reliability in capturing the trends of the vital signs. A high Pearson correlation indicates that a model can consistently predict the direction and magnitude of changes in HR and RR.

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

Signal-to-Noise Ratio (SNR): Defined according to the method proposed by de Haan et al. as the ratio between the area under the curve of the power spectrum around the first and second harmonic of the ground-truth signal rate frequency and the area under the curve of the rest of the power spectrum. This metric is critical for determining a model's effectiveness in distinguishing physiological signals from external disturbances, such as vehicle vibrations and lighting variations. High SNR values indicate superior performance in isolating the vital signs from noise [52].

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

$\hat{S}$ represents the power spectrum of the predicted signal S, f denotes the frequency, and $U_t(f)$ is a binary template that is equal to 1 around the first and second harmonics of the ground-truth signal and 0 everywhere else. In this implementation, we consider only the power spectrum between the frequencies (lf - hf) of 0.75 - 2.5 Hz for HR and 0.08 - 0.5 Hz for RR.

3.6.2 Efficiency Metric

In real-world automotive settings, the responsiveness of health monitoring systems is critical, requiring immediate action based on instantaneous, real-time data processing. Additionally, efficient algorithms are particularly valuable in these environments since computational and power resources are often limited. Therefore, assessing the

efficiency of such systems is crucial to understanding their practicality and effectiveness in real-world applications. To evaluate this aspect of the tested algorithms, a comparative measure was employed that enables the assessment of the NN models' efficiency relative to each other.

Throughput: Measures the average number of video frames that the NN models can process within one second and is quantified as frames per second (FPS). This metric provides an indicator of the algorithm's performance in terms of speed and real-time processing capability. Higher throughput rates indicate that the algorithm can handle live video streams efficiently, making it suitable for real-time applications. The throughput was calculated by dividing the total number of frames processed (N) by the total time taken by the GPU (t_{GPU}) to process these frames.

$$\textbf{Throughput} = \frac{N}{t_{GPU}}$$

As opposed to the performance metrics introduced in section 3.6.1, which evaluate the accuracy of the model on specific input data, throughput is completely independent of the benchmark datasets. Instead, it relies on hardware configuration and the NN model's architecture. By maintaining a consistent GPU setup and training parameters, as described in section 3.5, this measure was used to compare the efficiency of different NN architectures for rPPG. It should be noted that this measure only takes into consideration the processing time of the algorithms, meaning that preprocessing and postprocessing are not accounted for. Furthermore, throughput is applicable only to the NN models as the unsupervised methods perform all the processing on the CPU.

4
Results

In this thesis we conducted a benchmark study focusing on the effectiveness of ML models, particularly supervised NN models, as well as unsupervised methods, for rPPG signal detection in the dynamic environment of autonomous vehicles. This study utilized the MR-NIRP car dataset to evaluate the performance of various algorithms in detecting physiological signals, such as HR and RR. The NN algorithms tested included DeepPhys, TS-CAN, EfficientPhys-C, PhysNet and PhysFormer, which were trained and tested on the benchmark dataset using 5-Fold cross-validation. The detailed performance metrics of the models for each fold are presented in Appendix A.2.

To assess the NN algorithms' ability to generalize across different environments the same model architectures, pretrained on three general datasets, PURE, UBFC-rPPG and SCAMPS, were tested on MR-NIRP. Additionally, to determine the applicability of features learned in a vehicular setting to more general environments, the NN models trained on MR-NIRP were tested on the PURE and UBFC-rPPG datasets. SCAMPS was excluded from this testing phase as it contains only synthetic data that would not reflect the algorithm's performance in the real world.

Furthermore, several conventional unsupervised methods were tested including ICA, POS, CHROM, GREEN, LGI, and PBV to act as a baseline for assessing the effectiveness of the NN rPPG models.

4.1 Heart Rate estimation

In this section, the test results for HR estimation using both unsupervised methods and supervised NN models are presented.

4.1.1 Unsupervised methods

For HR estimation, the unsupervised methods showed inconsistent performance across different scenarios.

Table 4.1: Overall HR performance metrics of each unsupervised method evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets.

Unsupervised Methods - HR															
	MR-NIRP					PURE					UBFC-rPPG				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
ICA	9.08	14.34	13.14	0.32	-6.60	3.06	12.27	3.57	0.84	8.41	4.15	12.95	3.91	0.77	3.26
POS	**7.74**	14.26	12.30	0.39	**-3.71**	4.23	12.81	8.41	0.85	8.12	**0.54**	**1.71**	**0.55**	**0.99**	**6.90**
CHROM	8.94	15.60	14.05	0.31	-5.41	5.38	14.94	10.58	0.79	5.11	0.70	2.58	0.72	0.99	4.23
GREEN	12.44	17.19	17.71	0.16	-11.59	11.17	23.71	12.89	0.29	-4.25	16.15	27.83	15.18	0.31	-9.88
LGI	7.87	**13.33**	**11.21**	**0.44**	-4.81	**1.09**	**5.34**	**1.46**	**0.97**	**9.38**	1.57	7.13	1.60	0.92	4.94
PBV	9.72	15.45	13.95	0.28	-7.13	7.31	17.92	8.74	0.64	0.62	3.73	11.56	3.68	0.82	0.82

On the MR-NIRP dataset, POS exhibited the lowest MAE, indicating its superior accuracy in HR estimation among the unsupervised methods (Table 4.1). LGI followed closely with regard to MAE but outperformed all methods in terms of RMSE. This suggests that LGI, while slightly less accurate on average than POS, has a better performance in minimizing large errors in HR estimation. The MAPE metric further supports the effectiveness of LGI, presenting the lowest value among all methods, indicating its higher overall precision across varying HR ranges. In terms of correlation with the true HR (ρ), LGI again achieved the best performance. However, when comparing the SNR, POS outperforms all, suggesting that it is better at distinguishing the HR signal from background noise, a critical factor in dynamic environments like those encountered in autonomous vehicles.

Further evaluation of the unsupervised methods on the PURE and UBFC-rPPG datasets revealed that the performance differences among the algorithms were highly inconsistent and heavily dependent on the context and conditions of the data they were tested on. This evaluation also highlighted that the MR-NIRP dataset posed substantially greater challenges compared to the general datasets tested for the unsupervised methods.

Table 4.2: Case specific HR performance metrics of each unsupervised method evaluated on the MR-NIRP subsets.

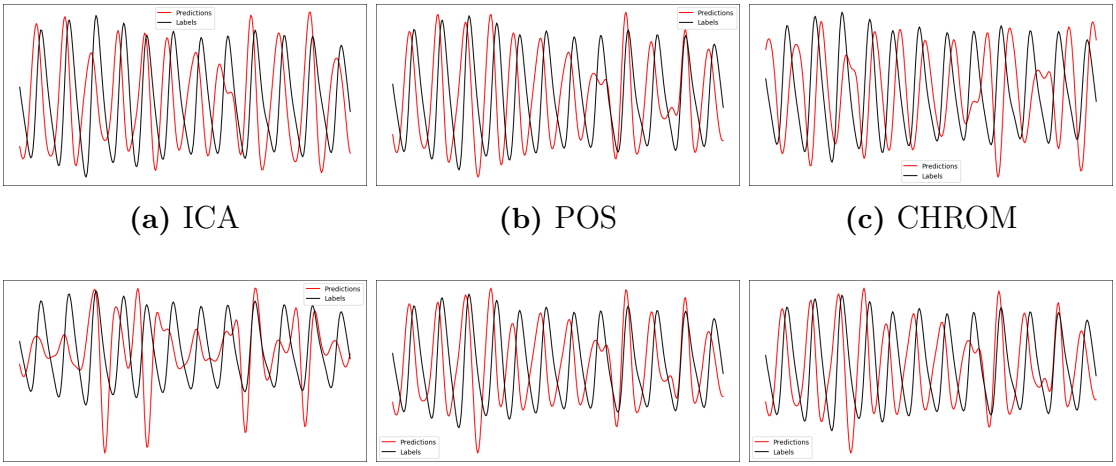
Unsupervised Methods - HR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
ICA	11.14	15.42	14.77	0.23	-10.96	10.58	14.92	14.43	0.22	-10.11	9.46	14.07	13.34	0.30	-7.60
POS	**9.16**	**14.22**	**13.01**	**0.34**	**-8.12**	**8.54**	**13.87**	**12.26**	**0.33**	**-6.96**	**8.14**	**13.97**	**12.37**	**0.40**	**-3.92**
CHROM	10.90	16.19	15.36	0.21	-9.75	9.78	15.22	14.01	0.27	-8.07	8.38	14.68	12.68	0.35	-4.72
GREEN	14.71	18.67	19.87	0.01	-14.16	13.27	17.31	17.81	0.11	-13.74	11.91	16.61	16.80	0.17	-10.86
LGI	12.23	17.16	16.32	0.09	-10.37	10.68	15.41	14.36	0.26	-8.84	8.74	14.08	12.51	0.34	-5.32
PBV	11.74	16.44	15.49	0.11	-10.66	12.01	16.88	16.19	0.13	-10.18	9.86	15.56	14.06	0.23	-6.55
	Garage Large Motion					Garage Small Motion					Garage Still				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
ICA	10.75	16.18	13.82	0.13	-9.22	6.19	11.93	8.57	0.44	-3.64	1.64	5.23	2.50	0.89	5.54
POS	**5.97**	12.56	9.00	0.53	**-4.23**	2.97	7.46	4.43	0.75	**-0.24**	1.73	7.29	3.04	0.80	6.11
CHROM	7.82	13.46	11.06	0.48	-6.98	5.53	11.50	8.22	0.51	-3.67	1.99	8.07	3.53	0.76	3.50
GREEN	13.66	18.10	18.16	0.10	-13.69	12.85	17.43	17.73	0.24	-12.16	6.71	12.83	9.99	0.42	-5.72
LGI	6.37	**11.14**	**8.93**	**0.61**	-5.83	**2.90**	**6.44**	**3.98**	**0.82**	-0.95	**0.91**	**2.50**	**1.30**	**0.98**	**6.19**
PBV	8.04	12.89	10.89	0.52	-6.50	5.69	10.17	7.75	0.66	-5.37	5.41	12.22	7.87	0.51	-1.07

To assess the algorithms' performance in specific vehicular conditions we are evaluating them on subsets of the MR-NIRP dataset. As detailed in Section 3.3.2, MR-NIRP is divided into two distinct subsets based on the vehicle's state. The first contains recordings captured in a garage setting with a stationary vehicle where illumination is stable. The second subset consists of recordings taken during driving, where background motion, window reflections, and fluctuating lighting conditions present more challenging conditions. Each subset is further split into three categories based on the subject's movement: still, small motion, large motion.

In realistic driving conditions, the majority of unsupervised methods proved ineffective in predicting accurate HR measurements (Table 4.2). This is evident as all methods had RMSE close to, or greater than 15 beats/min, which indicates that the HR was incorrectly estimated and the rPPG signal was not recovered well. Across all three driving conditions, POS consistently showed the most robust performance in all evaluation measures. LGI, while slightly less effective in higher motion scenarios, offered competitive accuracy, especially in driving still conditions.

In the controlled garage environment, the performance differences among the algorithms were more pronounced (Table 4.2). LGI demonstrated the best performance, achieving the lowest overall error and highest ρ in all garage scenarios. In large and small motion scenarios POS had comparable performance to LGI while also outperforming it in certain metrics.

Generally, GREEN and PBV under-performed in comparison to the other tested methods. Notably, ICA, POS and CHROM exhibited strong performance under optimal conditions, such as the garage still environment but did not adapt well to the other cases. Furthermore, all methods experienced a drastic decrease in accuracy when subjected to conditions involving motion or changes in illumination.



(a) ICA **(b)** POS **(c)** CHROM

(d) GREEN **(e)** LGI **(f)** PBV

Figure 4.1: Example plot of the predicted HR signal from each unsupervised method (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset.

The prediction examples depicted in Figure 4.1, while circumstantial, effectively illustrate the common issues encountered in unsupervised rPPG methods. These methods often fail to accurately capture the shape and temporal dimension of the BVP, as shown by the clear misalignment with the ground-truth signal in all cases, which can cause inaccuracies in extracting vital signs like HR.

4.1.2 Neural Network models

The comparative analysis of the supervised NN models for HR estimation reveals clear variations in performance across different model architectures. Tables 4.3 and 4.4 compare the performance of all the evaluated NN models when trained on the MR-NIRP dataset.

Table 4.3: Overall HR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset and evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets.

NN Models - HR															
	MR-NIRP					PURE					UBFC-rPPG				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	7.45	12.69	9.95	0.31	-4.37	1.91	9.64	2.07	0.91	8.25	1.98	7.11	2.02	0.92	**3.16**
TS-CAN	6.78	11.76	9.17	0.40	-4.20	1.16	6.00	**1.60**	0.97	8.85	**1.38**	**5.02**	**1.44**	**0.96**	2.82
EfficientPhys-C	7.01	12.20	9.46	0.36	-4.25	1.36	6.04	1.99	0.96	7.06	3.95	12.88	3.73	0.78	1.37
Physnet	**4.28**	**9.02**	**5.67**	**0.55**	**2.02**	1.32	6.59	2.00	0.96	**9.01**	12.91	24.34	11.50	0.33	-3.40
PhysFormer	6.42	11.01	8.33	0.38	-0.39	**1.14**	**4.77**	1.82	**0.98**	8.87	24.39	33.11	22.02	-0.05	-12.45

On the MR-NIRP dataset, PhysNet demonstrated the strongest performance in all evaluation metrics, followed by PhysFormer and TS-CAN. DeepPhys and EfficientPhys-C, while not performing as well as the top models, still outperformed many unsupervised methods.

Table 4.4: Case specific HR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset using 5-Fold cross-validation and evaluated on the MR-NIRP subsets.

NN Models - HR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	11.79	16.49	15.71	0.08	-10.18	9.95	14.56	13.00	0.21	-8.02	7.00	12.06	9.74	0.32	-3.47
TS-CAN	11.73	15.98	15.58	0.18	-10.45	8.83	13.48	11.81	0.31	-8.09	7.09	11.73	9.91	0.36	-4.20
EfficientPhys-C	11.07	15.48	14.78	0.20	-10.50	9.64	14.41	12.67	0.24	-8.28	7.44	12.52	10.42	0.31	-3.98
PhysNet	**7.04**	**11.55**	**9.20**	**0.35**	**-4.21**	**5.80**	**10.47**	**7.56**	**0.44**	**-1.27**	**4.37**	**9.00**	**6.11**	**0.51**	**2.02**
PhysFormer	9.17	12.97	12.08	0.19	-7.22	7.88	12.28	10.16	0.26	-3.39	6.35	10.81	8.57	0.37	-0.10
	Garage Large Motion					Garage Small Motion					Garage Still				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	10.47	14.65	13.31	0.24	-8.72	4.99	9.46	6.65	0.51	-3.36	1.14	3.80	1.65	0.87	6.60
TS-CAN	8.45	12.90	10.87	0.31	-7.13	3.32	7.66	4.58	0.63	-1.97	0.81	2.32	1.13	0.88	7.38
EfficientPhys-C	8.25	12.70	10.62	0.31	-7.64	3.66	8.27	4.98	0.61	-1.86	0.96	3.08	1.41	0.87	7.12
PhysNet	**5.43**	**10.01**	**6.53**	**0.52**	**-0.17**	**1.96**	**5.37**	**2.46**	**0.78**	**5.54**	**0.35**	**1.32**	**0.52**	**0.98**	**11.78**
PhysFormer	7.07	11.28	9.10	0.35	-2.45	4.11	7.96	5.00	0.54	2.92	3.19	6.72	3.75	0.73	9.05

In intra-dataset testing, the performance differences between the NN models were more inconsistent. While models trained on the MR-NIRP dataset showed minimal performance differences when tested on the PURE dataset, disparities were more pronounced on UBFC-rPPG. All models performed well on PURE, reinforcing the observation that PURE is generally a less challenging dataset. Notably, TS-CAN outperformed all models on the UBFC-rPPG test set, while PhysNet and PhysFormer struggled to adapt their features effectively. Similar observations were made when models trained on the general-purpose datasets were tested on MR-NIRP, details of which will be presented in subsequent sections.

A comprehensive analysis of each model's performance is detailed in subsequent sections.

4.1.2.1 DeepPhys

When trained and tested on the MR-NIRP dataset, DeepPhys demonstrated promising results (Table 4.3), achieving performance comparable to the best unsupervised methods tested (Table 4.1).

Table 4.5: Overall HR performance metrics of the DeepPhys model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

DeepPhys - HR - MR-NIRP					
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**7.45**	**12.69**	**9.95**	**0.31**	**-4.37**
PURE	10.10	15.31	13.72	0.24	-8.09
UBFC-rPPG	10.14	15.35	13.75	0.25	-8.12
SCAMPS	11.17	16.04	15.22	0.19	-9.95

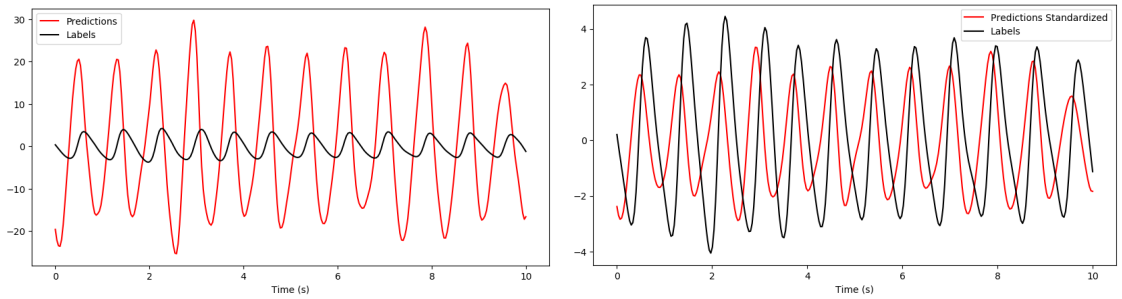
Specifically, DeepPhys outperformed all unsupervised methods in every error metric, MAE, RMSE and MAPE. However, despite its superior accuracy, DeepPhys exhibited slightly lower Pearson correlation coefficient (ρ) and SNR values compared to some unsupervised approaches. This indicates that while DeepPhys is adept at minimizing absolute errors, its correlation with the actual HR signal and its ability to filter out noise may not always surpass those of the top-performing unsupervised algorithms.

Examining case specific test results, it is evident that the model experienced a similar decrease in accuracy as the unsupervised methods when subjected to non-optimal conditions (Table 4.4). While DeepPhys performs adequately in garage still scenarios it seems to not be able to adapt well when faced with motion artifacts and changes in illumination.

Table 4.6: Case specific HR performance metrics of the DeepPhys trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

DeepPhys - HR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**11.79**	**16.49**	**15.71**	0.08	**-10.18**	**9.95**	**14.56**	**13.00**	**0.21**	**-8.02**	**7.00**	**12.06**	**9.74**	**0.32**	**-3.47**
PURE	12.89	16.88	17.29	**0.14**	-12.92	12.68	17.05	17.10	0.15	-11.68	10.12	15.26	14.21	0.30	-7.56
UBFC-rPPG	13.67	17.90	18.17	0.09	-13.26	13.06	17.28	17.68	0.16	-12.44	10.61	15.67	14.94	0.25	-8.24
SCAMPS	13.62	17.37	18.04	0.07	-12.48	12.40	16.59	16.71	0.21	-11.66	9.97	15.39	14.00	0.29	-7.95
	Garage Large Motion					Garage Small Motion					Garage Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**10.47**	**14.65**	**13.31**	**0.24**	**-8.72**	**4.99**	**9.46**	**6.65**	**0.51**	**-3.36**	**1.14**	**3.80**	**1.65**	**0.87**	**6.60**
PURE	13.00	18.29	16.60	0.05	-11.65	9.46	14.86	12.46	0.13	-8.23	2.31	7.56	3.64	0.77	3.50
UBFC-rPPG	12.96	18.29	16.61	0.04	-10.96	8.30	13.70	11.07	0.27	-7.21	1.53	5.55	2.10	0.87	4.69
SCAMPS	14.74	19.03	19.17	0.02	-13.77	11.66	16.89	15.93	0.00	-10.93	6.62	11.84	9.53	0.46	-5.32

The generalization capability of DeepPhys was further explored through its performance when trained on the PURE, UBFC-rPPG, and SCAMPS datasets (Tables 4.5, 4.6). While these models showed an expected increase in error metrics as compared to the one trained on MR-NIRP, the differences are not significant. It should also be noted that the one trained on the synthetic dataset, SCAMPS, had the most drastic drop in accuracy.



(a) Raw Predicted Signal **(b)** Standardized Predicted Signal

Figure 4.2: Example plot of the predicted HR signal from DeepPhys (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the DeepPhys model trained using fold 1 (3.4). The raw predicted signal is depicted in plot (a) and the standardized predicted signal in (b).

It is a common phenomenon that NN models for rPPG struggle to retain the scale of the labels they are trained on. This is evidently the case for DeepPhys as shown by the example in Figure 4.2 and also for all other CNN-based NN models tested in this benchmark (Figures 4.3, 4.4, 4.5). Even though the magnitude of the predicted signal does not match the ground-truth, the model seems to be able to predict the frequency contents well, which is the only factor necessary for HR estimation. This is an interesting observation, but the discrepancy in the scale of the two signals should not affect the accuracy of the model in extracting most vital signs.

4.1.2.2 TS-CAN

The TS-CAN model managed to outperform DeepPhys in all metrics when trained and tested on the MR-NIRP dataset (Table 4.3).

Table 4.7: Overall HR performance metrics of the TS-CAN model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

TS-CAN - HR - MR-NIRP					
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**6.78**	**11.76**	**9.17**	**0.40**	**-4.20**
PURE	10.75	16.51	15.00	0.24	-8.59
UBFC-rPPG	9.93	15.28	13.81	0.28	-7.71
SCAMPS	12.88	17.89	18.20	0.13	-11.41

Overall, TS-CAN exhibited a significantly higher Pearson's correlation coefficient (ρ) with the ground-truth HR signal than DeepPhys. This was a point of contention for DeepPhys since it was not able to compete with unsupervised methods such as POS and LGI. Although TS-CAN did not outperform LGI in this metric, it effectively narrowed the gap, showcasing a competitive edge in accurately correlating with the true HR signals.

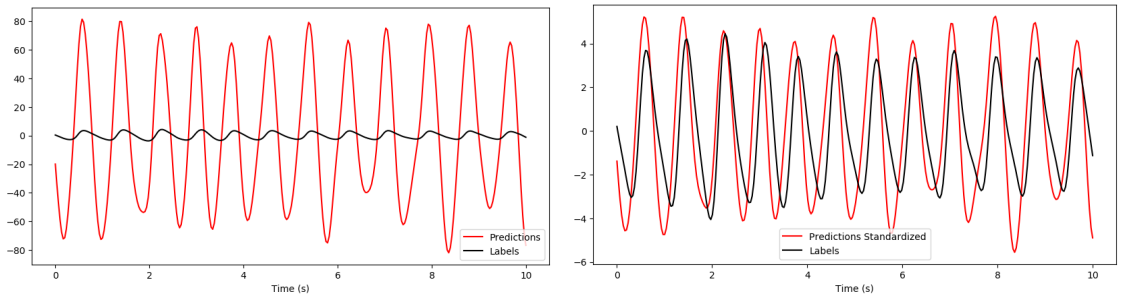
Table 4.8: Case specific HR performance metrics of the TS-CAN trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

TS-CAN - HR - MR-NIRP															
	Driving Large Motion					**Driving Small Motion**					**Driving Still**				
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**11.73**	**15.98**	**15.58**	**0.18**	**-10.45**	**8.83**	**13.48**	**11.81**	**0.31**	**-8.09**	**7.09**	**11.73**	**9.91**	**0.36**	**-4.20**
PURE	12.97	17.49	17.47	0.16	-12.20	12.82	17.87	17.57	0.16	-11.29	11.17	16.69	16.25	0.29	-7.86
UBFC-rPPG	14.33	18.44	19.66	0.07	-12.92	13.26	17.93	18.43	0.15	-11.54	9.75	14.89	14.01	0.32	-7.47
SCAMPS	15.03	18.95	20.77	0.06	-13.83	14.25	18.68	20.00	0.13	-13.26	12.73	18.08	18.70	0.19	-10.58
	Garage Large Motion					**Garage Small Motion**					**Garage Still**				
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**8.45**	**12.90**	**10.87**	**0.31**	**-7.13**	**3.32**	**7.66**	**4.58**	**0.63**	**-1.97**	**0.81**	**2.32**	**1.13**	**0.88**	**7.38**
PURE	11.55	16.62	15.17	0.19	-11.26	10.45	16.82	14.36	0.14	-8.44	4.34	11.84	6.70	0.57	-0.71
UBFC-rPPG	11.11	16.26	14.25	0.14	-9.76	7.82	13.34	10.74	0.38	-6.30	1.99	6.64	2.97	0.82	2.84
SCAMPS	15.16	19.74	20.20	-0.01	-12.72	12.95	18.26	18.27	0.07	-11.43	7.61	13.23	10.81	0.30	-6.76

In case-specific results, TS-CAN demonstrated commendable robustness when subjected to motion, acquiring significant performance improvements over DeepPhys in the garage tests (Table 4.4). However, similar to DeepPhys, it encountered difficulties under varying illumination conditions, failing to provide accurate HR predictions in the more challenging driving scenarios.

Regarding its generalization capabilities, TS-CAN performed similarly to DeepPhys, with very accurate results in the least challenging garage still test case but struggled to adapt to varying motion and illumination conditions (Tables 4.7, 4.8). Notably, a

significant disparity in performance was observed when TS-CAN was trained on the PURE versus the UBFC-rPPG dataset. Furthermore, the model did not produce accurate results when trained on the synthetic SCAMPS dataset.



(a) Raw Predicted Signal **(b)** Standardized Predicted Signal

Figure 4.3: Example plot of the predicted HR signal from TS-CAN (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the TS-CAN model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels.

4.1.2.3 EfficientPhys-C

EfficientPhys-C is designed to be efficient and lightweight while achieving comparable performance to other NN models for rPPG.

Table 4.9: Overall HR performance metrics of the EfficientPhys-C model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

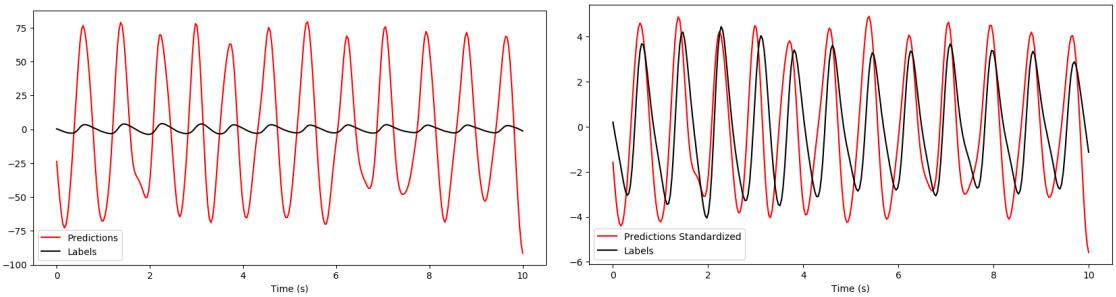
EfficientPhys-C - HR - MR-NIRP					
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**7.01**	**12.20**	**9.46**	**0.36**	**-4.25**
PURE	9.22	14.52	12.59	0.30	-7.63
UBFC-rPPG	10.02	15.53	13.82	0.26	-7.60
SCAMPS	12.86	17.45	17.99	0.14	-11.69

When trained and evaluated on the MR-NIRP dataset using 5-Fold cross-validation, EfficientPhys-C performed similarly to the DeepPhys and TS-CAN. Specifically, it managed to outperform the DeepPhys model in all measures while falling slightly behind TS-CAN in all performance metrics. It should be noted that EfficientPhys-C is considerably faster than TS-CAN (Table 4.19) while achieving comparable accuracy, which highlights a potential advantage over other NN models.

Table 4.10: Case specific HR performance metrics of the EfficientPhys-C trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

EfficientPhys-C - HR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**11.07**	**15.48**	**14.78**	**0.20**	**-10.50**	**9.64**	**14.41**	**12.67**	**0.24**	**-8.28**	**7.44**	**12.52**	**10.42**	**0.31**	**-3.98**
PURE	12.23	16.63	16.60	**0.20**	-12.17	11.46	16.43	15.47	0.21	-10.07	8.41	13.55	11.94	0.41	-5.91
UBFC-rPPG	13.23	17.35	17.99	0.18	-12.42	12.77	17.77	17.50	0.15	-10.98	10.05	15.63	14.31	0.26	-6.55
SCAMPS	13.76	17.43	19.16	0.15	-13.42	13.95	18.46	19.61	0.11	-12.45	11.63	17.14	16.77	0.19	-9.72
	Garage Large Motion					Garage Small Motion					Garage Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**8.25**	**12.70**	**10.62**	**0.31**	**-7.64**	**3.66**	**8.27**	**4.98**	**0.61**	**-1.86**	**0.96**	**3.08**	**1.41**	**0.87**	**7.12**
PURE	11.59	16.96	15.01	0.08	-10.37	8.14	13.44	11.01	0.26	-7.68	3.82	9.39	5.36	0.61	-0.78
UBFC-rPPG	12.71	17.98	16.27	-0.03	-11.23	8.60	14.18	11.67	0.30	-7.35	2.48	7.75	3.87	0.75	2.39
SCAMPS	13.86	18.05	18.39	0.00	-13.30	13.68	17.79	18.71	0.08	-12.57	11.16	15.76	15.72	0.19	-10.52

In the context of case-specific and intra-dataset evaluations (Tables 4.9, 4.10), the model follows the general trend observed in its overall performance metrics. While its results are consistently close to those achieved by TS-CAN, EfficientPhys-C falls slightly behind in nearly all performance metrics.



(a) Raw Predicted Signal **(b)** Standardized Predicted Signal

Figure 4.4: Example plot of the predicted HR signal from EfficientPhys-C (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the EfficientPhys-C model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels.

4.1.2.4 PhysNet

The PhysNet model, when trained and tested on the MR-NIRP dataset for HR estimation, significantly outperformed both the supervised and unsupervised methods previously discussed (Table 4.3).

Table 4.11: Overall HR performance metrics of the PhysNet model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

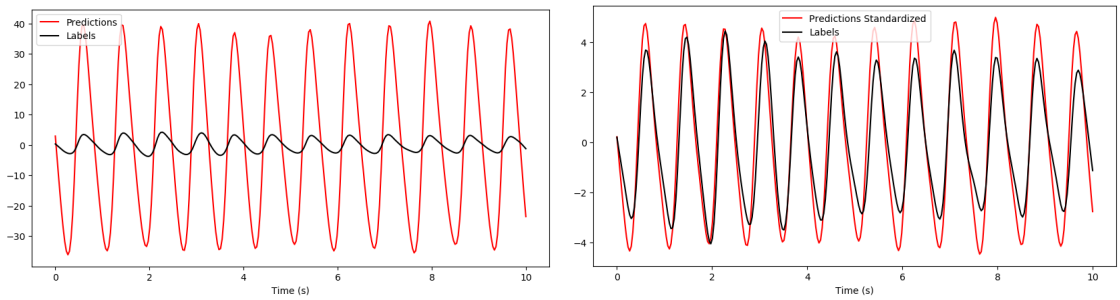
PhysNet - HR - MR-NIRP					
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**4.28**	**9.02**	**5.67**	**0.55**	**2.02**
PURE	14.87	23.40	21.39	0.14	-8.91
UBFC-rPPG	13.10	19.52	19.48	0.22	-9.32
SCAMPS	21.99	27.79	32.80	0.04	-13.92

When trained directly on the MR-NIRP dataset, PhysNet managed to achieve outstanding results in all test cases, with a significant reduction in all error metrics compared to the other evaluated methods. Notably, the Pearson correlation coefficient (ρ) and SNR achieved with PhysNet were higher than those obtained with unsupervised methods, indicating a strong agreement with the true HR signals and effectiveness in canceling out noise. The main distinguishing characteristic of this model to the other CNN based architectures is the integration of a temporal dimension which seems to offer a distinct advantage in accurately capturing the rPPG signals in environments with increased motion and background noise.

Table 4.12: Case specific HR performance metrics of the PhysNet trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

PhysNet - HR - MR-NIRP															
	Driving Large Motion					**Driving Small Motion**					**Driving Still**				
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**7.04**	**11.55**	**9.20**	**0.35**	**-4.21**	**5.80**	**10.47**	**7.56**	**0.44**	**-1.27**	**4.37**	**9.00**	**6.11**	**0.51**	**2.02**
PURE	10.94	15.77	14.99	0.26	-10.06	10.85	16.70	14.82	0.30	-8.95	9.12	16.16	13.16	0.31	-5.80
UBFC-rPPG	15.85	20.86	22.96	0.03	-12.41	14.87	20.64	21.73	0.19	-11.46	14.99	22.08	22.90	0.18	-8.55
SCAMPS	20.26	25.69	29.36	-0.04	-14.41	20.41	26.28	30.04	0.07	-13.99	19.26	25.32	29.21	0.12	-12.47
	Garage Large Motion					**Garage Small Motion**					**Garage Still**				
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**5.43**	**10.01**	**6.53**	**0.52**	**-0.17**	**1.96**	**5.37**	**2.46**	**0.78**	**5.54**	**0.35**	**1.32**	**0.52**	**0.98**	**11.78**
PURE	20.75	28.09	29.65	-0.09	-12.46	24.48	33.68	35.85	-0.04	-11.17	24.19	33.96	36.32	0.10	-9.43
UBFC-rPPG	12.78	18.12	18.11	0.20	-10.99	11.07	16.87	16.26	0.26	-9.04	5.88	13.07	9.57	0.49	-3.27
SCAMPS	22.85	27.98	33.32	0.00	-15.16	25.64	31.34	38.35	-0.04	-14.61	27.51	32.51	41.99	0.07	-14.69

When examining the model's performance across different training datasets not specifically tailored to the dynamic conditions of autonomous vehicles, a sharp decline in accuracy is observed (Tables 4.11, 4.12). Specifically, the model did not retrieve an accurate HR signal in all intra-dataset tests as indicated by the extremely high error metrics. This drop in performance underscores the challenge of generalizing across datasets with varying characteristics and conditions.



(a) Raw Predicted Signal **(b)** Standardized Predicted Signal

Figure 4.5: Example plot of the predicted HR signal from PhysNet (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the PhysNet model trained using fold 1 (3.4). The predicted signal in (a) does not retain the magnitude of the ground-truth. In (b) the predicted signal is standardized to have a more accurate comparison with the labels.

4.1.2.5 PhysFormer

As a Transformer-based model, PhysFormer is designed to leverage temporal dynamics and long-range dependencies within rPPG signals, which could be particularly advantageous for complex and dynamic scenarios such as those presented in autonomous vehicles.

Table 4.13: Overall HR performance metrics of the PhysFormer model trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

PhysFormer - HR - MR-NIRP					
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
MR-NIRP	**6.42**	**11.01**	**8.33**	**0.38**	**-0.39**
PURE	10.79	17.27	15.31	0.32	-8.23
UBFC-rPPG	14.34	20.78	21.74	0.15	-9.86
SCAMPS	28.61	34.05	42.87	-0.01	-14.92

When trained and tested on the MR-NIRP dataset, PhysFormer showcased performance that surpassed the majority of the evaluated models across most metrics (Table 4.3). Despite not reaching the high accuracy level set by PhysNet, PhysFormer achieved lower error rates than all the 2D-CNN models included in the benchmark. However, it is important to note that while PhysFormer had lower errors compared to TS-CAN, which is the top-performing 2D-CNN model, it fell short in terms of the Pearson correlation coefficient (ρ) and SNR. The lower performance in these areas suggests that while PhysFormer is adept at reducing average estimation errors, its predictions may not align as closely with the ground-truth across the entire range

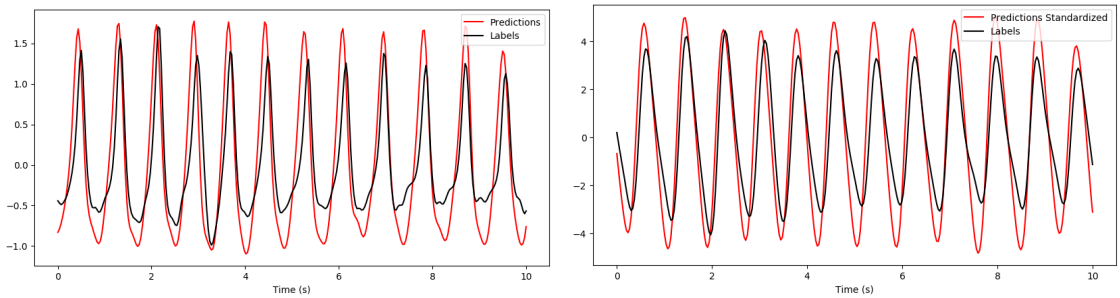
of HRs, and it may struggle more with noise in the signal compared to TS-CAN.

Table 4.14: Case specific HR performance metrics of the PhysFormer trained on MR-NIRP, PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

PhysFormer - HR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**9.17**	**12.97**	**12.08**	**0.19**	**-7.22**	**7.88**	**12.28**	**10.16**	**0.26**	**-3.39**	**6.35**	**10.81**	**8.57**	**0.37**	**-0.10**
PURE	13.26	18.16	18.38	**0.19**	-11.86	12.66	18.71	17.63	0.26	-10.31	11.30	18.03	16.71	0.29	-7.34
UBFC-rPPG	16.77	21.80	24.73	0.07	-13.68	17.09	22.50	25.47	0.08	-12.72	15.89	22.05	24.62	0.06	-9.86
SCAMPS	25.86	30.62	37.72	0.01	-15.88	27.84	33.34	41.24	-0.03	-14.90	29.06	35.02	44.27	0.00	-14.49
	Garage Large Motion					Garage Small Motion					Garage Still				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
MR-NIRP	**7.07**	**11.28**	**9.10**	**0.35**	**-2.45**	**4.11**	**7.96**	**5.00**	**0.54**	**2.92**	**3.19**	**6.72**	**3.75**	**0.73**	**9.05**
PURE	10.09	16.13	13.72	0.32	-9.30	9.09	16.37	12.81	0.42	-6.86	6.11	13.18	8.89	0.56	-3.42
UBFC-rPPG	12.51	17.25	17.82	0.22	-11.11	12.24	19.69	18.58	0.18	-8.35	7.19	16.29	11.95	0.35	-1.85
SCAMPS	30.10	34.79	44.16	-0.09	-15.77	29.49	34.78	43.74	0.00	-15.06	30.24	35.56	46.77	0.00	-14.28

When examining the performance of PhysFormer across different scenarios, it demonstrated great adaptability to situations involving high motion and varying illumination conditions (Table 4.4). Interestingly, the model had the weakest performance out of all tested algorithms in the garage still scenario, which presented optimal conditions. In contrast, PhysFormer exhibited remarkable consistency in its performance across varying degrees of motion, as its accuracy was less impacted by increased motion and illumination artifacts.

In terms of generalization performance, PhysFormer struggled to produce viable results during intra-dataset testing (Tables 4.13, 4.14). Mirroring the trend of PhysNet, the model experienced a sharp decline in performance when trained on datasets with different conditions from its training set.



(a) Raw Predicted Signal **(b)** Standardized Predicted Signal

Figure 4.6: Example plot of the predicted HR signal from PhysFormer (red) against the ground-truth signal (black). The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the PhysFormer model trained using fold 1 (3.4). The raw predicted signal is depicted in plot (a) and the standardized predicted signal in (b).

Noteworthy is the ability of PhysFormer to capture the actual scale of the ground-truth data. As mentioned before, NN models are not usually able to retain the magnitude of the signal they are trained on. As illustrated in Figure 4.6, PhysFormer not only aligns with the temporal pattern of the ground-truth signal but also its amplitude.

4.2 Respiration Rate estimation

In this section, the test results for RR estimation using both unsupervised methods and supervised NN models will be presented.

Many of the rPPG algorithms tested in this benchmark did not manage to predict a signal that contained adequate respiratory information. For those methods there is no value in evaluating and comparing their performance as they did not accurately capture the respiratory signal.

4.2.1 Unsupervised methods

For RR detection, the effectiveness of unsupervised rPPG methods demonstrated mixed results.

Table 4.15: Overall RR performance metrics of the unsupervised methods evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. The unsupervised methods GREEN, LGI, and PBV did not yield accurate RR predictions.

Unsupervised Methods - RR															
	MR-NIRP					**PURE**					**UBFC-rPPG**				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
ICA	**2.45**	**4.86**	**26.45**	**0.04**	**21.69**	**2.44**	**4.96**	**26.06**	0.04	**19.39**	**2.66**	**5.36**	**24.09**	0.10	19.38
POS	4.03	7.88	53.16	0.02	18.02	2.50	5.31	26.40	0.01	18.47	3.04	6.07	29.29	-0.09	**19.71**
CHROM	5.50	9.60	77.65	-0.01	14.90	3.11	6.12	35.08	0.05	16.48	4.55	8.38	53.66	0.04	17.34
GREEN	11.81	14.62	185.25	-0.04	4.26	8.40	11.79	119.82	-0.04	8.25	8.38	11.78	126.91	0.01	9.49
LGI	10.87	13.80	168.96	-0.01	4.82	4.10	7.38	52.95	**0.18**	14.66	5.95	9.91	90.02	**0.34**	9.84
PBV	11.26	14.15	176.63	-0.03	4.97	5.23	8.79	70.84	0.17	12.31	7.98	11.52	115.03	0.13	7.53

Overall, unsupervised methods did not produce accurate RR predictions as indicated by the large errors (Table 4.15). It should be noted that for RR detection, all results with RMSE greater than 5 would be considered inaccurate. The main exception is ICA, which showed some promise with the lowest errors across all tested algorithms. However, the low Pearson correlation coefficient (ρ) suggests a lack of alignment between the predicted respiration signals and the actual ground-truth. This discrepancy indicates that while this method can estimate RR with relatively minimal error in terms of absolute values, it still struggles to capture the precise

timing and pattern of respiration, pointing to a significant weakness in the predictive capabilities of current unsupervised methods.

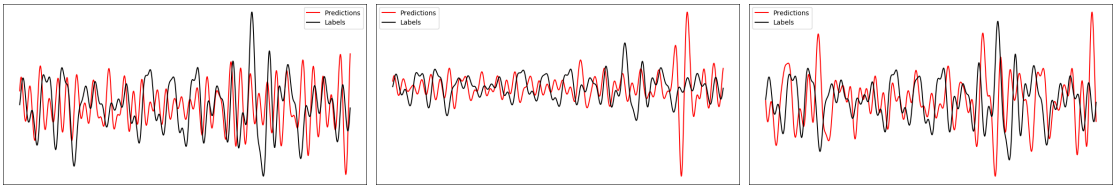
Table 4.16: Case specific RR performance metrics of the unsupervised methods evaluated on the MR-NIRP dataset. The unsupervised methods GREEN, LGI, and PBV did not yield accurate RR predictions in any test case.

Unsupervised Methods - RR - MR-NIRP

	Driving Large Motion					Driving Small Motion					Driving Still				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
ICA	**3.20**	**6.29**	**31.19**	-0.08	**20.13**	**2.59**	**5.05**	**27.50**	**0.16**	**21.16**	**2.08**	**4.20**	**22.90**	**0.07**	**22.24**
POS	4.24	8.15	50.07	**0.05**	16.34	4.33	8.21	59.57	0.12	16.81	4.79	9.04	65.51	-0.01	16.60
CHROM	6.45	10.80	85.41	-0.03	12.67	5.86	10.15	86.46	0.07	13.54	7.25	11.39	109.32	-0.12	12.78
GREEN	13.58	15.85	215.38	-0.09	1.40	13.72	15.81	218.05	-0.04	0.43	13.20	15.60	209.30	-0.01	2.84
LGI	13.37	15.55	210.18	-0.05	0.51	13.05	15.30	203.44	-0.04	0.07	12.92	15.10	204.57	0.00	1.96
PBV	13.13	15.30	209.13	-0.04	1.13	14.08	16.07	223.91	-0.05	0.31	12.13	14.73	191.03	0.06	2.57

	Garage Large Motion					Garage Small Motion					Garage Still				
Method:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
ICA	2.94	**5.50**	**28.47**	-0.19	**21.63**	**2.54**	**4.96**	**27.22**	**0.01**	**21.41**	**1.89**	**3.43**	**25.41**	-0.05	**23.33**
POS	**3.05**	5.99	30.04	-0.03	19.19	2.79	5.92	30.90	-0.04	20.99	2.22	4.28	30.58	**-0.02**	22.37
CHROM	5.44	9.30	70.10	**0.11**	15.97	3.39	6.96	39.45	-0.01	18.15	3.05	5.72	42.93	-0.03	19.42
GREEN	14.06	16.00	219.71	-0.05	1.59	9.37	12.80	141.86	-0.13	8.30	6.04	9.95	92.18	-0.05	13.26
LGI	12.55	14.88	191.88	-0.05	3.71	7.57	11.28	111.93	0.01	9.95	3.98	7.48	61.15	-0.05	17.35
PBV	13.60	15.72	206.93	-0.23	2.04	8.65	12.07	132.74	-0.07	10.69	5.01	8.92	77.84	-0.09	16.49

Case-specific assessments of the unsupervised methods revealed considerable resilience to motion variations, with ICA and POS displaying minimal deviation in their performance across different motion cases (Table 4.16). However, POS achieved RMSE slightly lower than 5 only in the optimal garage still tests. Furthermore, ICA exhibited good adaptability in varying illumination, showing consistent results when comparing controlled garage environments to dynamic driving scenarios.



(a) ICA **(b)** POS **(c)** CHROM

Figure 4.7: Example plot of the predicted RR signal against the ground-truth signal (black) for ICA, POS, CHROM unsupervised methods. The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset.

The example comparison of signals predicted by the three unsupervised methods against the ground-truth, shown in Figure 4.7, underscores the observations made regarding the low ρ values. It is evident from these visualizations that the predicted signals bear little to no resemblance to the actual respiration signal. This reinforces the conclusion that there is a significant discrepancy in capturing the true pattern of respiration.

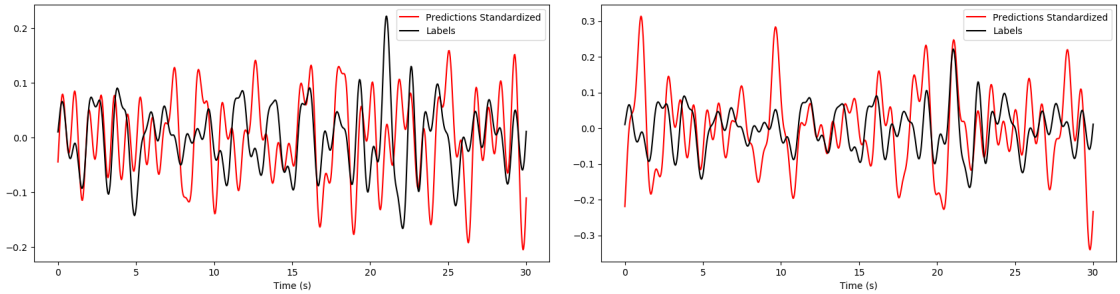
4.2.2 Neural Network models

Similarly to the unsupervised methods, most supervised NN models did not produce accurate RR estimates. PhysNet and PhysFormer, however, demonstrated promising results, surpassing the top performing unsupervised methods in certain metrics.

Table 4.17: Overall RR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset and evaluated on the MR-NIRP, PURE, UBFC-rPPG and SCAMPS datasets. DeepPhys, TS-CAN and EfficientPhys-C did not yield accurate RR predictions.

NN Models - RR															
	MR-NIRP					PURE					UBFC-rPPG				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	10.95	13.70	175.18	0.07	5.23	4.35	7.95	64.88	0.21	15.32	10.05	13.22	156.47	**0.23**	5.86
TS-CAN	10.37	13.40	164.64	0.04	6.11	3.12	6.50	41.84	0.23	17.51	7.84	11.48	114.38	-0.04	9.46
EfficientPhys-C	10.51	13.52	166.78	0.03	5.81	3.96	7.36	55.81	0.27	14.54	7.87	11.51	118.22	0.21	9.16
Physnet	**3.12**	**6.36**	**39.39**	**0.15**	**16.84**	**1.18**	**3.47**	**16.19**	**0.48**	**19.38**	6.55	10.50	96.60	-0.01	10.91
PhysFormer	4.10	7.49	54.09	-0.01	15.95	3.86	7.46	49.79	0.07	16.21	**4.81**	**8.11**	**59.43**	0.09	**13.74**

When trained and evaluated on the MR-NIRP dataset, PhysNet emerged as the superior model for RR estimation among the supervised NN models, showcasing the lowest MAE, RMSE, and MAPE, as well as the highest Pearson correlation coefficient (ρ) and SNR (Table 4.17). Notably, PhysNet achieved significantly higher correlation with the ground-truth signal than any other supervised or unsupervised method tested, indicating its superior ability to accurately capture the shape of the respiration signal. The example plot presented in Figure 4.8 confirms this observation. The visual comparison clearly shows that the predicted signals from these models capture the shape of the ground-truth respiratory signal better than the unsupervised methods in Figure 4.7.



(a) PhysNet **(b)** PhysFormer

Figure 4.8: Example plot of the predicted RR signal against the ground-truth signal (black) for PhysNet, PhysFormer NN models. The specific example corresponds to "subject18_garage_still_940" recording of the MR-NIRP dataset. The predicted signal was extracted with the models trained using fold 1 (3.4) and standardized.

In intra-dataset testing, the NN models were mostly unable to predict RR accurately. An exception to this occurred when the NN models were trained on the MR-NIRP

dataset and tested on PURE, where they performed adequately, with PhysNet even surpassing ICA, the best-performing unsupervised method on this dataset, across all performance metrics (Table 4.17). When trained on general-purpose datasets, no NN model was able to predict accurate RR measurements on the MR-NIRP dataset. Therefore, these test results are not reported in the main section of the study and are included in the Appendix A.1, instead. This outcome suggests a notable limitation in the models' ability to generalize RR estimation across datasets not specifically tailored for RR detection. It also underscores the difficulty these methods face in generating rPPG signals that accurately capture the respiration patterns present in the BVP.

Table 4.18: Case specific RR performance metrics of all the supervised NN models when trained on the the MR-NIRP dataset using 5-Fold cross-validation and evaluated on the MR-NIRP subsets. DeepPhys, TS-CAN and EfficientPhys-C did not yield accurate RR predictions in any test case.

NN Models - RR - MR-NIRP															
	Driving Large Motion					Driving Small Motion					Driving Still				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	10.48	13.56	166.99	0.14	4.42	12.30	14.58	191.16	0.05	3.72	11.42	13.96	186.00	0.04	4.80
TS-CAN	12.16	14.18	198.33	-0.01	5.87	12.94	15.01	202.08	-0.17	3.31	11.39	14.31	181.08	-0.13	7.58
EfficientPhys-C	13.18	15.12	211.18	0.08	4.64	11.85	14.29	187.24	0.01	5.05	10.78	13.66	171.87	-0.16	6.43
PhysNet	**4.09**	**7.71**	**44.99**	**0.03**	**14.08**	**4.42**	**8.11**	**54.83**	**-0.03**	**13.93**	**2.95**	**5.90**	**36.39**	**0.40**	**16.80**
PhysFormer	5.36	8.77	73.63	-0.01	12.99	4.93	8.28	60.39	-0.02	13.41	4.22	7.59	55.40	-0.04	15.92
	Garage Large Motion					Garage Small Motion					Garage Still				
NN Model:	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	ρ↑	**SNR↑**
DeepPhys	13.77	16.35	211.11	-0.49	6.06	11.50	14.09	194.79	-0.09	5.57	5.20	8.30	79.03	0.23	12.50
TS-CAN	8.06	10.61	115.00	0.00	9.27	12.30	15.02	202.99	0.04	5.86	3.66	7.05	59.31	0.04	15.74
EfficientPhys-C	10.99	14.22	167.22	-0.33	4.79	9.89	13.13	163.82	0.05	5.89	4.61	9.35	77.15	-0.05	14.49
PhysNet	**1.79**	**3.45**	**24.04**	**0.07**	**17.45**	**2.35**	**4.03**	**34.58**	**0.19**	**19.25**	**1.25**	**2.88**	**18.77**	**0.40**	**22.50**
PhysFormer	4.81	8.10	68.49	-0.08	15.27	3.43	6.39	45.92	-0.11	18.43	1.44	2.92	20.72	0.39	21.43

Case-specific assessments revealed that PhysNet achieved superior performance in all garage scenarios, but experienced a significant increase in errors during the driving test cases (Table 4.18). This indicates that the model can be resilient to motion in controlled environments, but it fails to adapt well in realistic driving conditions with varying illumination. As for Physformer, although it performed comparably to PhysNet in the optimal garage still tests, it did not achieve accurate results in scenarios involving motion or changes in lighting.

4.3 Efficiency of Neural Network models

In this section we present the efficiency evaluation of the NN models. The throughput for all NN models was calculated across the tests conducted in this benchmark and is presented in Table 4.19.

Table 4.19: Average Throughput of NN models.

NN Model:	**Throughput**↑ (FPS)
DeepPhys	15863
TS-CAN	15412
EfficientPhys-C	**25776**
Physnet	1812
PhysFormer	8076

EfficientPhys-C demonstrated the highest throughput, significantly outperforming the other models. This model's speed is largely attributed to its design optimizations for computational efficiency, making it particularly suitable for real-time applications in resource-constrained environments [47].

Both DeepPhys and TS-CAN showed comparable throughput levels. The incorporation of the TSM in TS-CAN appears to enhance its accuracy capabilities without substantially impacting its processing speed, thus maintaining a competitive throughput. Conversely, Physnet and PhysFormer, while delivering superior performance in terms of accuracy (Section 4.1), recorded the lowest throughput out of all NN models. Apparently, their clear performance advantage over the other tested NN models comes at the cost of efficiency. Their slower processing rates may limit their utility in scenarios where immediate data processing is crucial.

It is important to note that all models achieved throughput rates sufficiently high to handle real-time processing demands under our experimental setup and hardware configuration (Section 3.5). Given that the video data were recorded at 30Hz, each model is capable of processing at least 60 seconds of input frames every second (Throughput $>$ 1800 FPS), which is greater than the evaluation window used in this benchmark for both HR and RR, confirming their feasibility for real-time application in automotive systems.

5
Discussion

In this chapter, we present a summary of the results that answers the two main research questions (Section 1.3) of this thesis. This overview highlights key insights and conclusions drawn from the benchmark analysis of NN models and unsupervised methods in the context of HR and RR estimation within dynamic vehicle environments. Following this summary, we discuss various factors that may have influenced the results of this benchmark.

5.1 Discussion of Results

This section aims to summarize the results and provide our conclusions regarding the two research questions of this thesis. RQ1 aims to explore the feasibility and effectiveness of rPPG NN models in accurately detecting HR and RR in a vehicular environment. RQ2 aims to assess the effectiveness of NN models trained for generalized environments in accurately estimating HR and RR in a vehicle environment.

5.1.1 Heart Rate estimation in vehicles

To answer RQ1 for the HR estimation task, we evaluated and compared the performance of various NN models and unsupervised methods for rPPG in a vehicular setting. According to the results in section 4.1, the NN models, specifically trained on the MR-NIRP dataset, demonstrated commendable performance in the dynamic environment of a vehicle. In general, the majority of NN models were able to outperform conventional unsupervised methods in HR estimation, under both optimal and challenging conditions (Figure 5.1).

Particularly notable is the performance of PhysNet, which displayed superior accuracy across all test cases. This model exhibited remarkable resilience to motion artifacts, showing only minimal decreases in accuracy during high-motion tests. Additionally, while all methods experienced some performance degradation under varied illumination conditions that are present on the realistic driving scenarios, PhysNet proved to be the most robust, consistently outperforming others.

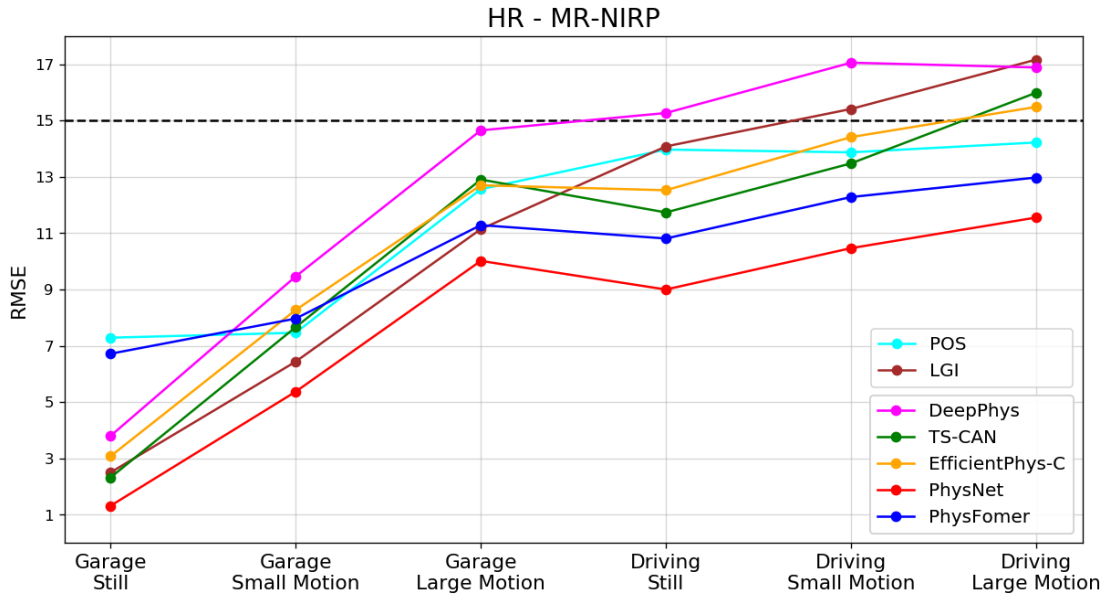


Figure 5.1: Visualization of case specific HR estimation RMSE of POS, LGI unsupervised methods and all the supervised NN models when trained and evaluated on the the MR-NIRP dataset (Tables 4.2, 4.4).

Another point of discussion could be the ability of PhysFormer to capture the actual scale of the ground-truth data. As mentioned in section 4.1.2.1, CNN-based models fail to retain the magnitude of the ground-truth signal they are trained on. This seemingly unique feature of Physformer could be a result of its distinct transformer-based architecture or the use of a specialized loss function for training, which might have enabled the model to accurately learn the magnitude of the data it is trained on. Even though this characteristic does not offer any benefits when it comes to HR estimation, it could be a considerable advantage of this model in other rPPG signal processing applications.

5.1.2 Respiration Rate estimation in vehicles

For the RR estimation task, the answer to RQ1 is somewhat ambiguous, as many of the tested algorithms failed to produce meaningful results. The best performing NN models, PhysNet and PhysFormer, were not able to outperform any of the unsupervised methods in the error metrics. However, these models did generate respiration signals that correlated more closely with the ground-truth compared to unsupervised methods. Additionally, PhysNet once again showed the best motion robustness in the garage test cases. Despite this, like all NN models, it was not as effective under varying illumination when compared to the best unsupervised methods.

Overall, the performance of the best NN models and unsupervised methods was largely constrained by challenging conditions, with these algorithms only managing to produce accurate predictions in optimal, controlled environments, such as the garage still scenarios. In more dynamic and less predictable settings, such as dur-

ing driving with large or small motion and varying illumination, their effectiveness significantly diminished (Figure 5.2).

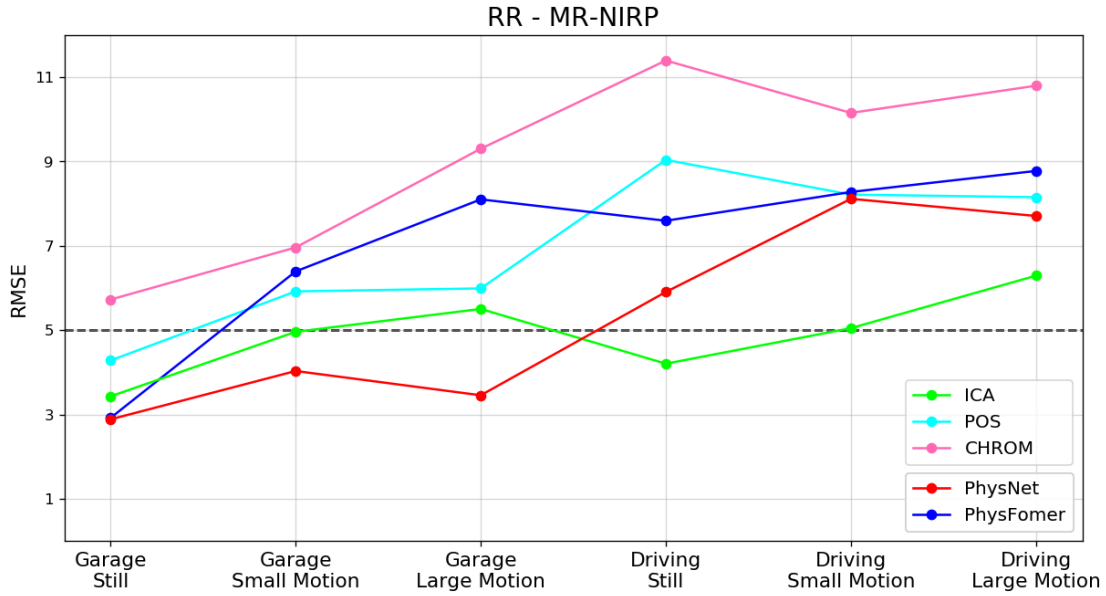


Figure 5.2: Visualization of case specific RR estimation RMSE of ICA, POS, CHROM unsupervised methods and PhysNet, PhysFormer supervised NN models when trained and evaluated on the the MR-NIRP dataset (Tables 4.16, 4.18).

The generally low ρ observed across all RR tests does raise some doubts on the reliability of these results for evaluating the effectiveness of the algorithms. The challenge of RR detection is compounded by the nature of the respiration signal, which occupies a very narrow and low frequency range ($\sim$0.08-0.5 Hz), in contrast to HR signals ($\sim$0.75-2.5 Hz) which typically has much greater amplitude and range. This results in the ground-truth respiration signal in the recorded BVP to have very low SNR which makes it inherently difficult to distinguish the respiration signal from potential noise.

5.1.3 Generalization Performance

To address RQ2 we tested the performance of the NN models when trained with general purpose datasets and evaluated on the MR-NIRP. For HR estimation, the models trained on PURE, UBFC-rPPG or SCAMPS and tested on MR-NIRP generally exhibited a decline in performance compared to when trained and tested on MR-NIRP directly (Figure 5.3).

Overall, the ability of the NN models to generalize across settings is variable and independent from their performance in RQ1. Models like DeepPhys, TS-CAN and EfficientPhys demonstrated moderate generalization abilities, managing adequate performance under optimal, garage still, vehicular conditions. However, in more challenging cases with dynamic lighting and motion changes their performance de-

teriorated greatly. This aligns with the limitations of the general datasets as they do not contain the range of motion, lighting, and other environmental factors specific to vehicular settings, thereby limiting the models' ability to adapt to these conditions.

Furthermore, the best performing models, PhysNet and PhysFormer, exhibited notably poor generalization performance with large errors even in optimal conditions. This discrepancy indicates that the models might overfit to the specific environmental characteristics of the training datasets. As a result these models are able to isolate the necessary ROIs from background noise making them robust and effective in the setting they are trained on, at the cost of generalizability across different environments.

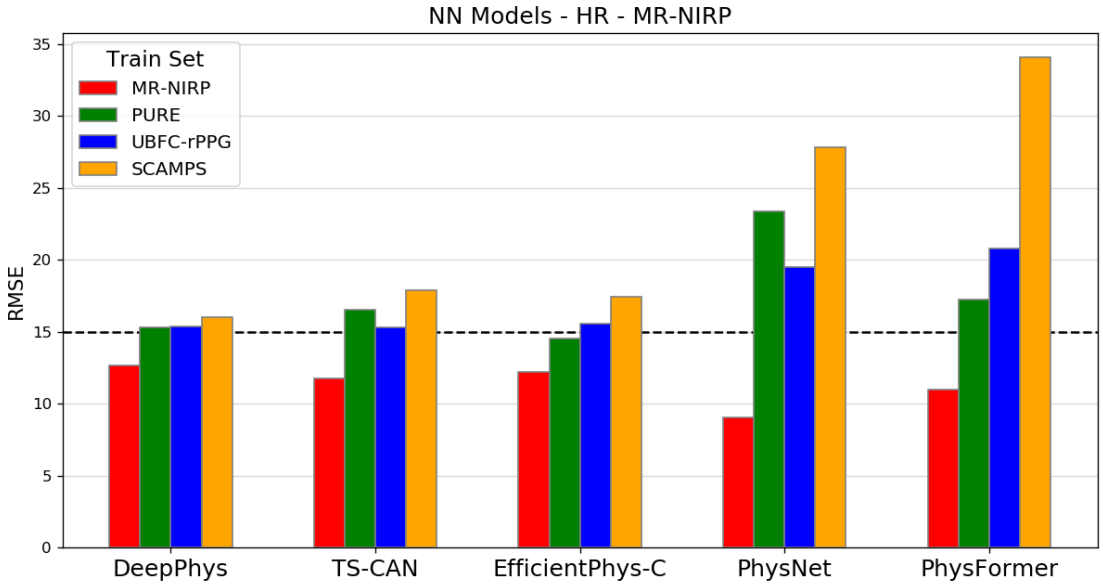


Figure 5.3: Visualization of the overall HR estimation RMSE of all the supervised NN models when trained on MR-NIRP, PURE, UBFC-rPPG, or SCAMPS and evaluated on the the MR-NIRP dataset (Table 4.3).

The generalization performance of NN models for RR estimation was particularly poor, as most of the evaluated models did not produce viable RR estimates in intra-dataset testing. As discussed previously (Section 5.1.2), extracting RR using rPPG poses significant challenges, where detecting subtle respiratory signals becomes particularly difficult since these signals are often obscured by noise. The task of generalizing features from stationary or less dynamic settings to highly variable environments, such as inside a moving vehicle, complicates this issue.

5.2 Influencing Factors

There are several factors that may influence the validity of the results. Understanding these factors is crucial for interpreting the findings accurately and for considering

the broader implications and potential limitations of this benchmark study.

5.2.1 Vehicular Dataset limitations

This benchmark study relies solely on the MR-NIRP dataset for testing in a vehicular environment, presenting a notable limitation due to the lack of dataset diversity. The videos in the dataset were recorded from a single type of vehicle cabin, restricting the variety of vehicle environments represented. This constraint raises concerns about the ability of the models to generalize across different vehicular settings. Additionally, employing the same dataset for training, validation, and testing could potentially introduce bias, which we have attempted to mitigate by implementing subject exclusive 5-Fold cross-validation. By adopting this approach, we ensure that all subjects are included in the test set, while also minimizing the risk of the models overfitting to any individual subject's facial features.

Furthermore, as mentioned in section 3.3.2, MR-NIRP consists mainly of male participants. As a result, it lacks representation of various aspects of women's appearances, such as cultural nuances or makeup styles. This dataset imbalance may introduce bias and result in poor performance when tested on more diverse data.

Additionally, by the dataset's design, all subjects were placed in the passenger seat. This could potentially compromise the validity of the study with regard to drivers' vital sign monitoring applications. This aspect however, does not pose a significant cause for concern as the exclusive focus of the tested algorithms is the face, meaning that the subject and camera placement should not impact the accuracy of the predictions as long as there is a clear view of their face.

5.2.2 ROI extraction for Unsupervised methods

ROI extraction is a common preprocessing step when evaluating the performance of unsupervised rPPG methods. This process enables algorithms to concentrate on the color fluctuations of skin pixels rather than the background noise in the frames. While ROI extraction may be unnecessary under optimal conditions with a uniform background and consistent lighting, it became essential in the dynamic setting of the MR-NIRP dataset, where external illumination and motion introduced significant noise [56][40].

Additionally, the selection of specific facial landmarks can significantly impact the effectiveness of rPPG algorithms. Extensive research has explored how different landmark regions influence the accuracy of rPPG measurements. For the purposes of this benchmark study, however, we chose to utilize the entire face as the ROI. This approach prioritizes simplicity and ensures that all facial regions, including those not typically ideal, are considered, as they may still hold rPPG information that the algorithms can leverage [67][68].

5.2.3 Excluding Cases vs. Pseudo-Labeling

In this benchmark, we decided to exclude certain cases that displayed evidence of incorrect sampling in the ground-truth PPG data (Section 3.4.3). An alternative approach would be pseudo-labeling, which involves substituting faulty signals with labels generated through an unsupervised rPPG method applied to the video data. While this technique has been employed in other studies to preserve the utility of cases with sampling errors for training and testing neural network models, we opted against it [64].

Our primary concern was that pseudo-labeling could introduce significant bias into the neural network models, potentially compromising the accuracy of the results. Moreover, utilizing labels generated by the same unsupervised methods that the study aims to evaluate would raise substantial questions about the benchmark's validity, as it could obscure the true comparative performance of the tested algorithms. Thus, excluding problematic cases was deemed a more sensible choice to ensure the integrity and reliability of our findings.

5.2.4 Selection of Loss function

For the training configuration (Section 3.5.2), we maintained most parameters consistent with those utilized in the pretrained models of the rPPG-Toolbox. However, we diverged in our choice of loss function, opting for negative Pearson correlation loss for all models except PhysFormer, in contrast to the Mean Squared Error (MSE) loss employed by the original authors for training DeepPhys, TS-CAN, and EfficientPhys [61].

This deliberate choice was influenced by the unique challenges posed by the MR-NIRP dataset, which exhibits greater variations in lighting and motion compared to the other datasets. Given these characteristics, training with MSE loss proved to be ineffective due to its high sensitivity to scale. Adjusting the loss function helped accommodate the specific needs of MR-NIRP for additional normalization, enhancing the models' robustness against these variabilities.

5.2.5 Accuracy of extraction methods

A critical factor that can influence rPPG-based estimation is the accuracy of the vital sign extraction methods. This process must effectively isolate the subtle physiological signals from background noise and artifacts typical in dynamic environments, such as those encountered in vehicle settings. In this benchmark, we utilized a FFT-based approach to isolate the dominant frequency that corresponds to the appropriate HR or RR. This is a simple and accurate method commonly used in many rPPG applications. However, it is possible that a different combination of filtering and extraction techniques could yield more accurate readings.

Regarding the extraction of RR, as mentioned in section 2.3, the process of respiration has multiple measurable effects on the BVP. Using a FFT-based method, that only takes into consideration the frequency domain of the signal, means it relies solely on the effects of RSA for calculating the RR. Therefore other approaches that leverage a combination of the other respiration induced variations could be more effective in isolating the respiration signal from potential noise.

5.3 Ethical Considerations

While conducting this study, we have prioritized the ethical sourcing of all data and the protection of privacy concerns. All data used in this research are permitted for academic purposes. The datasets utilized have explicit permissions for research applications, ensuring compliance with legal and ethical standards. We have ensured that our research aligns with the guidelines and expectations set by the data providers and the academic community. Furthermore, all published data, including images, adhere to permissible use guidelines. The visual data presented in this thesis are sourced from subjects that permit academic publication, and we have ensured that no sensitive information is disclosed.

6
Conclusion

In this thesis, we explore the feasibility and effectiveness of utilizing camera-based, deep learning algorithms for the detection of vital signs in autonomous vehicles. The objective was to benchmark various rPPG techniques in the context of dynamic vehicular environments, characterized by challenges such as motion artifacts and fluctuating lighting conditions.

Our findings indicate that the application of ML models, particularly NN-based approaches, demonstrated significant potential in accurately estimating HR and RR from video data in these settings. The results validated our hypotheses, showing that supervised NN models can perform robustly under the conditions tested, despite the inherent challenges. By addressing the specific requirements and obstacles of in-vehicle rPPG measurement, this study contributes valuable insights towards the integration of non-invasive vital sign monitoring systems in autonomous vehicles. The success of our work underscores the promise of deep learning methodologies in enhancing driver safety and health monitoring in real-world automotive applications.

6.1 Future Work

Despite the accomplishments of this study, there are areas for enhancement in future research. As mentioned in section 5.2.1, the vehicular dataset used for this benchmark is imbalanced, posing a risk of generating biased results. To address this potential risk, future work could involve incorporating a larger and more balanced in-vehicle dataset for training and testing.

This benchmark study focuses on comparing methods solely tested on RGB recordings. To expand the scope of the benchmark in future efforts, NIR recordings could also be integrated. As described in the related work, section 2.2.3, both NIR and RGB-based methods offer distinct advantages and disadvantages in vital sign detection. By evaluating both methodologies, a more comprehensive benchmark can be established that accounts for the diverse strengths and limitations of each approach.

To expand the benchmark's scope further, future research could encompass testing on additional vital signs like HRV and SPO2. While our current focus has been only on HR and RR detection due to time constraints, it's important to recognize

that there are other vital signs that should be considered for holistic assessment of a driver's health status. Exploring these additional parameters could potentially provide a more comprehensive understanding of different physiological dynamics.

Considering the continuous evolution of this field, it's essential to regularly update the benchmark with the most recent advancements in NN-based methodologies. This proactive approach is important for the future work in preserving the benchmark's reliability and utility, ensuring its continued value as a trusted resource for researchers and software engineers.

The results obtained from RR testing, as discussed in section 5.1.2, did not yield satisfactory outcomes. Future research could focus on enhancing this aspect by exploring alternative methods better suited for RR detection. Although extracting the RR signal from rPPG is feasible, it presents challenges and the potential for inaccuracies. Therefore, in future work, it may be advantageous to explore optical-flow based methods, as mentioned in section 2.2.2, that could potentially offer improved accuracy in RR estimation.

Bibliography

[1] Guo, K., Zhai, T., Purushothama, M. H., Dobre, A., Meah, S., Pashollari, E., Vaish, A., DeWilde, C., & Islam, M. N. (2022, apr). Contactless Vital Sign Monitoring system for in-vehicle driver monitoring using a near-infrared time-of-Flight camera. *Appl. Sci. (Basel), 12*(9), 4416. https://doi.org/10.3390/app12094416

[2] Kumar, M., Veeraraghavan, A., & Sabharwal, A. (2015). DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomed. Opt. Express, 6*(5), 1565-1588. 10.1364/BOE.6.001565

[3] Peruzzini, M., Foco, E. & Reboa, A. (2018). Toward the Definition of a Technological Set-up for Drivers' Health Status Monitoring. *Studies in Health Technology and Informatics 251*, 221. 10.3233/978-1-61499-898-3-221.

[4] Singh, S. (2015). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. February 2015.

[5] Nowara, E. M., Marks, T. K., Mansour, H., & Veeraraghavan, A. (2018). *SparsePPG: Towards Driver Monitoring Using Camera-Based Vital Signs Estimation in Near-Infrared.* IEEE. 10.1109/CVPRW.2018.00174

[6] Othman, W., Kashevnik, A., Ryabchikov, I., & Shilov, N. (2023). Contactless Camera-Based Approach for Driver Respiratory Rate Estimation in Vehicle Cabin. *Intelligent Systems and Applications, 2*, 429 - 442. https://doi.org/10.1007/978-3-031-16078-3_29

[7] Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express, 18*(10), 10762-74. 10.1364/OE.18.010762

[8] Daimiwal, N., Sundhararajan, M. & Shriram, R. (2014). Respiratory rate, heart rate and continuous measurement of BP using PPG. *International Conference on Communication and Signal Processing*, 999-1002. 10.1109/ICCSP.2014.6949996

[9] Reiss, A., Indlekofer, I., Schmidt, P. & Van Laerhoven, K. (2019). Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors 19*(14), 3079. https://doi.org/10.3390/s19143079

[10] Peper, E., Havrey, R., I-Mei, L., Tylova, H. & Moss, D. (2007). Is There More to Blood Volume Pulse Than Heart Rate Variability, Respiratory Sinus Arrhythmia, and Cardiorespiratory Synchrony?. *Biofeedback 35*(2), 54-61. https://api.semanticscholar.org/CorpusID:15486681

[11] Nardelli, M., Vanello, N., Galperti, G., Greco, A. & Scilingo, E.P. (2020). Assessing the Quality of Heart Rate Variability Estimated from Wrist and Finger

PPG: A Novel Approach Based on Cross-Mapping Method. *Sensors 20*(11), 3156. https://doi.org/10.3390/s20113156

[12] Tasli, H. E., Gudi, A. & den Uyl, M. (2014). Remote PPG based vital sign measurement using adaptive facial regions. *IEEE International Conference on Image Processing (ICIP)*, 1410-1414. 10.1109/ICIP.2014.7025282

[13] Karlen, W., Raman, S., Ansermino, J. M., & Dumont, G. A. (2013). Multiparameter Respiratory Rate Estimation From the Photoplethysmogram. *IEEE Transactions on Biomedical Engineering 60*(7), 1946-1953. https://doi.org/10.1109/TBME.2013.2246160

[14] Sun, Y., & Thakor, N. (2016). Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. *IEEE transactions on bio-medical engineering, 63*(3), 463–477. https://doi.org/10.1109/TBME.2015.2476337

[15] Mohan, P. M., Nisha, A. A., Nagarajan, V. & Jothi, E. S. J. (2016) Measurement of arterial oxygen saturation (SpO2) using PPG optical sensor. *2016 International Conference on Communication and Signal Processing (ICCSP)*, 1136-1140. 10.1109/ICCSP.2016.7754330

[16] Qiao, D., Zulkernine, F., Masroor, R., Rasool R. & Jaffar N. (2021) Measuring Heart Rate and Heart Rate Variability with Smartphone Camera. *22nd IEEE International Conference on Mobile Data Management (MDM)*, 248-249. 10.1109/MDM52706.2021.00049

[17] Kanva, A. K., Sharma, C. J., & Deb, S. (2014). Determination of SpO2 and heart-rate using smartphone camera. *Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, 237-241. 10.1109/CIEC.2014.6959086

[18] Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable photoplethysmographic sensors—Past and present. *Electronics, 3*(2), 282-302. https://doi.org/10.3390/electronics3020282

[19] Leonhardt, S., Leicht, L., & Teichmann, D. (2018). Unobtrusive Vital Sign Monitoring in Automotive Environments—A Review. *Sensors (Basel), 18*(9), 3080. https://doi.org/10.3390/s18093080

[20] Jeong, I. C., Lee, D. H., Park, S. W., Ko, J. I. & Yoon, H. R. (2007) Automobile driver's stress index provision system that utilizes electrocardiogram. *IEEE Intelligent Vehicles Symposium*, 652-656. 10.1109/IVS.2007.4290190

[21] Lee, H. B. et al. (2007). Nonintrusive Biosignal Measurement System in a Vehicle. *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2303-2306. 10.1109/IEMBS.2007.4352786

[22] Shin, H. S., Jung, S. -J., Kim, J. -J. & Chung, W. -Y. (2010). Real time car driver's condition monitoring system. *Sensors, IEEE*, 951-954. 10.1109/ICSENS.2010.5690904

[23] Schneider, J., Koellner, C. & Heuer, S.,(2012) An approach to automotive ECG measurement validation using a car-integrated test framework. *IEEE Intelligent Vehicles Symposium*, 950-955. 10.1109/IVS.2012.6232289

[24] Jung, S. J., Shin, H. S., Yoo, J. H. & Chung, W. Y. (2012) Highly sensitive driver condition monitoring system using nonintrusive active electrodes. *IEEE International Conference on Consumer Electronics (ICCE)*, 305-306. 10.1109/ICCE.2012.6161880

[25] Jung, S.-J., Shin, H.-S. & Chung, W.-Y. (2014). Driver fatigue and drowsiness monitoring system with embedded electrocardiogram sensor on steering wheel. *IET Intell. Transp. Syst., 8*, 43-50. https://doi.org/10.1049/iet-its.2012.0032

[26] Sidikova, M., Martinek, R., Kawala-Sterniuk, A., Ladrova, M., Jaros, R., Danys, L., & Simonik, P. (2020). Vital Sign Monitoring in Car Seats Based on Electrocardiography, Ballistocardiography and Seismocardiography: A Review. *Sensors, 20*(19), 5699. 10.3390/s20195699

[27] Leem, S.K., Khan, F. & Cho, S.H. (2017). Vital Sign Monitoring and Mobile Phone Usage Detection Using IR-UWB Radar for Intended Use in Car Crash Prevention. *Sensors, 17*(6), 1240. https://doi.org/10.3390/s17061240

[28] Goldberger, A. L., Goldberger, Z. D., & Shvilkin, A. (2023). Goldberger's Clinical Electrocardiography - E-Book: A Simplified Approach. *Elsevier Health Sciences*. https://books.google.se/booksid=KQWtEAAAQBAJ

[29] D'Angelo, L. T., Parlow, J., Spiessl, W., Hoch, S. & Lüth, T. C. (2010). A system for unobtrusive in-car vital parameter acquisition and processing. *4th International Conference on Pervasive Computing Technologies for Healthcare, Munich*, 1-7. 10.4108/ICST.PERVASIVEHEALTH2010.8877

[30] Walter, M., Eilebrecht, B., Wartzek, T. et al. (2011). The smart car seat: personalized monitoring of vital signs in automotive applications. *Pers Ubiquit Comput 15*, 707–715. https://doi.org/10.1007/s00779-010-0350-4

[31] Wusk, G. & Gabler, H. (2018). Non-Invasive Detection of Respiration and Heart Rate with a Vehicle Seat Sensor. *Sensors 185*, 1463. https://doi.org/10.3390/s18051463

[32] Wu, B. F., Chu, Y. W., Huang, P. W., Chung, M. L. & Lin, T. M. (2017). A Motion Robust Remote-PPG Approach to Driver's Health State Monitoring. *Computer Vision – ACCV 2016 Workshops 10116, Springer*. https://doi.org/10.1007/978-3-319-54407-6_31

[33] Wu, B. F., Chu, Y. W., Huang P. W. & Chung, M. L. (2019). Neural Network Based Luminance Variation Resistant Remote-Photoplethysmography for Driver's Heart Rate Monitoring. *IEEE Access, 7*, 57210-57225. 10.1109/ACCESS.2019.2913664

[34] Hernandez-Ortega, J., Nagae, S., Fierrez, J. & Morales, A. (2019). Quality-Based Pulse Estimation from NIR Face Video with Application to Driver Monitoring. *Pattern Recognition and Image Analysis, 11868*. https://doi.org/10.1007/978-3-030-31321-0_10

[35] Huang, P. -W., Wu, B. -J. & Wu, B. -F. (2021). A Heart Rate Monitoring Framework for Real-World Drivers Using Remote Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics, 25* 5, 1397-1408. 10.1109/JBHI.2020.3026481

[36] Xu, M., Zeng, G., Song, Y., Cao, Y., Liu, Z. & He, X. (2023). Ivrr-PPG: An Illumination Variation Robust Remote-PPG Algorithm for Monitoring Heart Rate of Drivers. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-10. 10.1109/TIM.2023.3271760

[37] Gong Z. et al. (2024). Heart Rate Estimation in Driver Monitoring System Using Quality-Guided Spectrum Peak Screening. *IEEE Transactions on In-*

strumentation and Measurement, 73, 1-14. 10.1109/TIM.2024.3352710

[38] Chiu, L. -W., Chou, Y. -R., Wu, Y. -C. & Wu, B. -F. (2023). Deep-Learning-Based Remote Photoplethysmography Measurement in Driving Scenarios With Color and Near-Infrared Images. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-12. 10.1109/TIM.2023.3328703

[39] Guo, T., Lin, Q. & Allebach, J. (2021). Remote estimation of respiration rate by optical flow using convolutional neural networks. *Proc. IS&T Int'l. Symp. on Electronic Imaging: Imaging and Multimedia Analytics in a Web and Mobile World*, 267-1 - 267-11. https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-267

[40] Wang, Z., Yang, X., Lu, H., Shan, C. & Wang W. (2023). Benchmark of Physiological Model Based and Deep Learning Based Remote Photoplethysmography in Automotive Applications. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. 10.1109/ICASSP49357.2023.10095078

[41] Vinci, G., Lenhard, T., Will, C. & Koelpin, A. (2015). Microwave interferometer radar-based vital sign detection for driver monitoring syst. *IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, 1-4. 10.1109/ICMIM.2015.7117940

[42] Lee, K. J., Park, C. & Lee, B. (2016). Tracking driver's heart rate by continuous-wave Doppler radar. *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5417-5420. 10.1109/EMBC.2016.7591952

[43] Schires, E., Georgiou, P. & Lande, T. S. (2018). Vital Sign Monitoring Through the Back Using an UWB Impulse Radar With Body Coupled Antennas. *IEEE Transactions on Biomedical Circuits and Systems, 12*2, 292-302. 10.1109/TBCAS.2018.2799322

[44] Chen, W., & McDuff, D. (2018). DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. *ArXiv, abs/1805.07888*. https://doi.org/10.48550/arXiv.1805.07888

[45] Liu, X., Fromm, J., Patel, S., & McDuff, D. (2021). Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. *arXiv:2006.03790v2*. https://doi.org/10.48550/arXiv.2006.03790Focustolearnmore

[46] Yu, Z., Li X. & Zhao, G. (2019). Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. *arXiv:1905.02419*. https://doi.org/10.48550/arXiv.1905.02419

[47] Liu, X. Hill B., Jiang Z., Patel S. & McDuff D. (2023). EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Cardiac Measurement. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4997-5006. 10.1109/WACV56688.2023.00498

[48] Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., & Zhao, G. (2022). PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. *arXiv preprint arXiv:2111.12082*. https://arxiv.org/abs/2111.12082

[49] Comon, P. (1994). Independent component analysis, A new concept?. *Signal Processing 36*(3), 287-314. https://doi.org/10.1016/0165-1684(94)90029-9

[50] Poh, M.-Z., McDuff, D. J. & Picard, R. W. (2011). Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering 58*(1), 7-11. 10.1109/TBME.2010.2086456

[51] Wang, W., den Brinker, A. C., Stuijk, S. & de Haan, G. (2017). Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering 64*(7), 1479-1491. 10.1109/TBME.2016.2609282

[52] de Haan, G. & Jeanne, V. (2013). Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering 60*(10), 2878-2886. 10.1109/TBME.2013.2266196

[53] Verkruysse, W., Svaasand, L.O., & Nelson, J.S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express 16*(26), 21434-21445. https://doi.org/10.1364/OE.16.021434

[54] Pilz, C., Zaunseder, S., Krajewski, J., & Blazek, V. (2018). Local Group Invariance for Heart Rate Estimation from Face Videos in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, https://doi.org/10.1109/CVPRW.2018.00172

[55] de Haan, G. & Leest, A. (2014). Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological Measurement 35*, 1913-1926. https://doi.org/10.1088/0967-3334/35/9/1913

[56] Nowara, E. M., Marks, T. K., Mansour, H., & Veeraraghavan, A. (2022). Near-Infrared Imaging Photoplethysmography During Driving. *IEEE Transactions on Intelligent Transportation Systems, 23*(4), 3589 - 3600. 10.1109/TITS.2020.3038317

[57] Stricker, R., Müller, S. & Gross, H.-M. (2014). Non-contact Video-based Pulse Rate Measurement on a Mobile Service Robot. *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 1056-1062. 10.1109/ROMAN.2014.6926392

[58] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A. & Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters 124*, 82-90. https://doi.org/10.1016/j.patrec.2017.10.017

[59] McDuff, D., Wander, M., Liu, X., Hill, B. L., Hernandez, J., Lester, J. & Baltrusaitis, T. (2022). SCAMPS: Synthetics for Camera Measurement of Physiological Signals. *arXiv:2206.04197*. https://doi.org/10.48550/arXiv.2206.04197

[60] Sim, S. E., Easterbrook, S. & Holt, R. C. (2003). Using benchmarking to advance research: a challenge to software engineering. *25th International Conference on Software Engineering*, 74-83. 10.1109/ICSE.2003.1201189

[61] Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Wang, Y., Sengupta, S., Patel, S., & McDuff, D. (2022). rPPG-Toolbox: Deep Remote PPG Toolbox. *arXiv preprint arXiv:2210.00716*. https://arxiv.org/abs/2210.00716

[62] Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5203-5212.

[63] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172*. https://arxiv.org/abs/1906.08172

[64] Gideon, J. & Stent, S. (2021). The Way to my Heart is through Contrastive Learning: Remote Photoplethysmography from Unlabelled Video. *arXiv preprint arXiv:2111.09748*. https://arxiv.org/abs/2111.09748

[65] Shannon, C. E. (1949). Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1), 10–21. https://doi.org/10.1109/jrproc.1949.232969

[66] Hafeez, Y., & Grossman, S. A. (2023). Sinus Bradycardia. In *StatPearls [Internet]*. Treasure Island, FL: StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK493201/

[67] Kwon, S., Kim, J., Lee, D., & Park, K. (2015). ROI analysis for remote photoplethysmography on facial video. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4938-4941. https://doi.org/10.1109/EMBC.2015.7319499

[68] Wang, G. (2021). Influence of ROI Selection for Remote Photoplethysmography with Singular Spectrum Analysis. *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, Guangzhou, China, 416-420. https://doi.org/10.1109/AIID51893.2021.9456548

A
Appendix

A.1 Additional Respiration Rate results

Table A.1: Overall RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset.

NN Models - RR - MR-NIRP

	DeepPhys					TS-CAN					EfficientPhys-C				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	12.56	14.79	202.15	0.07	2.90	12.10	14.48	194.46	0.02	3.82	12.69	14.95	203.81	0.02	3.45
UBFC-rPPG	12.54	14.85	201.82	0.05	2.94	12.53	14.85	200.35	0.02	3.56	12.28	14.70	196.74	0.03	3.86
SCAMPS	12.95	15.16	207.81	0.01	3.21	11.69	14.32	185.70	0.06	4.89	8.57	11.86	134.97	0.04	9.46

	PhysNet					PhysFormer				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	7.92	11.75	120.28	-0.01	9.54	10.87	13.80	168.96	-0.01	4.82
UBFC-rPPG	7.27	10.74	108.34	0.03	10.56	11.26	14.15	176.63	-0.03	4.97
SCAMPS	4.80	8.67	65.12	0.02	12.97	3.93	7.41	55.42	-0.10	17.36

Table A.2: RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset over the Driving test cases.

NN Models - RR - MR-NIRP

	Driving Large Motion					Driving Small Motion					Driving Still				
DeepPhys															
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	12.54	15.06	205.80	0.11	1.93	13.08	15.05	208.68	0.04	1.19	13.22	15.28	213.46	0.12	2.10
UBFC-rPPG	13.68	15.85	223.84	-0.05	0.49	13.21	15.25	212.14	0.05	1.40	12.66	14.99	205.38	0.12	2.51
SCAMPS	12.86	15.38	206.97	-0.17	2.40	13.33	15.45	212.36	0.08	2.60	12.95	15.14	208.00	0.03	3.52
TS-CAN															
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	10.61	14.10	178.05	0.17	6.74	9.33	12.46	144.87	0.08	7.14	8.24	11.62	133.88	0.03	8.24
UBFC-rPPG	10.48	13.69	167.60	0.32	3.89	10.94	13.54	171.34	0.02	3.25	12.03	14.51	200.08	-0.09	3.98
SCAMPS	12.97	15.35	203.62	-0.20	1.24	13.01	15.58	192.47	0.05	2.97	9.33	12.91	138.52	-0.05	6.06
EfficientPhys-C															
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	10.79	14.10	169.66	0.33	4.77	9.19	12.39	142.41	0.11	6.49	8.46	12.09	137.66	0.03	8.23
UBFC-rPPG	11.85	14.74	185.37	-0.01	3.82	11.56	14.51	175.92	0.08	3.08	10.71	13.48	180.20	-0.09	6.34
SCAMPS	14.15	16.31	216.42	-0.20	1.69	11.50	13.95	171.96	0.13	2.15	9.28	12.71	139.17	-0.06	5.20
PhysNet															
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	2.70	6.42	24.72	0.29	17.17	3.73	7.41	55.01	-0.17	17.55	2.21	5.15	33.10	0.28	19.14
UBFC-rPPG	5.04	8.52	46.50	-0.04	13.22	5.85	9.46	56.11	0.19	9.92	3.69	7.03	48.29	0.31	14.55
SCAMPS	4.31	8.33	46.74	-0.12	12.32	4.63	8.86	60.53	-0.09	14.45	4.15	8.13	41.54	0.17	14.35
PhysFormer															
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
PURE	2.89	6.21	35.65	-0.02	15.64	3.59	6.62	45.36	0.01	16.06	2.77	5.80	42.00	0.22	18.34
UBFC-rPPG	3.16	5.94	32.23	0.38	14.21	5.43	8.89	47.07	-0.10	12.09	2.83	6.18	27.65	-0.13	17.89
SCAMPS	8.07	12.24	107.87	-0.05	10.36	5.06	8.36	54.00	-0.04	12.18	6.90	10.71	90.21	-0.26	10.61

Table A.3: RR performance metrics of all the NN methods trained on PURE, UBFC-rPPG or SCAMPS and evaluated on the MR-NIRP dataset over the Garage test cases.

NN Models - RR - MR-NIRP															
	Garage Large Motion					Garage Small Motion					Garage Still				
DeepPhys															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
PURE	11.43	13.98	182.45	0.17	3.66	9.29	12.61	148.45	0.12	6.42	9.56	12.80	169.17	-0.25	7.07
UBFC-rPPG	11.87	14.65	208.33	0.08	2.29	9.17	12.25	151.85	0.17	4.18	8.09	11.20	130.40	-0.09	9.94
SCAMPS	12.30	15.00	186.67	-0.21	0.64	10.99	14.09	176.49	0.00	6.23	10.99	13.24	167.50	0.29	7.05
TS-CAN															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
PURE	12.52	15.29	217.47	0.10	3.80	8.04	12.05	132.26	0.05	8.81	12.52	15.48	216.67	-0.05	5.92
UBFC-rPPG	13.18	14.85	212.64	0.09	1.43	11.24	13.92	191.55	-0.08	4.46	4.64	7.77	65.07	0.39	13.86
SCAMPS	12.45	15.21	213.89	0.28	5.82	11.57	14.78	176.25	-0.22	2.50	5.42	8.13	80.90	0.18	11.97
EfficientPhys-C															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
PURE	10.88	13.90	181.25	0.21	5.06	7.60	11.18	126.55	-0.07	8.88	9.94	12.94	164.38	0.19	5.51
UBFC-rPPG	10.55	13.52	168.61	0.21	4.04	11.36	14.45	185.89	0.30	5.35	6.82	10.65	103.20	0.18	10.61
SCAMPS	9.67	13.33	158.89	0.39	5.55	10.91	13.83	161.54	-0.25	3.33	6.01	9.52	91.25	0.16	11.43
PhysNet															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
PURE	1.21	1.92	18.54	0.28	18.97	0.69	1.37	11.25	0.57	20.80	0.77	1.58	12.08	0.56	23.32
UBFC-rPPG	0.88	1.60	12.22	-0.20	17.17	3.01	5.10	46.90	0.08	19.12	1.62	3.27	17.82	0.76	21.16
SCAMPS	3.96	7.78	43.89	-0.15	10.59	5.35	9.59	74.10	-0.14	14.32	1.83	3.95	30.00	-0.05	20.19
PhysFormer															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
PURE	4.72	8.73	60.08	-0.19	14.19	2.45	5.57	41.55	-0.11	19.15	1.26	2.93	19.79	0.52	19.60
UBFC-rPPG	5.86	9.60	71.94	-0.12	12.14	1.82	4.14	29.94	-0.06	20.82	1.20	1.96	18.40	0.39	22.15
SCAMPS	8.50	12.70	134.44	-0.21	10.15	5.42	9.05	49.16	-0.04	14.88	2.34	5.76	27.85	0.50	19.85

A.2 Results per Fold

A.2.1 Heart Rate results

Table A.4: Overall HR performance metrics of all the NN methods for each Fold.

NN Models - HR															
	DeepPhys					TS-CAN					EfficientPhys-C				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	7.71	12.84	10.80	0.18	-4.99	7.66	12.64	10.82	0.28	-5.16	8.15	13.10	11.48	0.20	-5.25
FOLD2	6.48	11.19	9.50	0.48	-3.71	5.71	10.52	8.47	0.55	-2.83	5.79	11.09	8.65	0.50	-2.43
FOLD3	5.55	9.94	8.03	0.28	-3.06	5.55	9.81	8.09	0.30	-3.29	5.81	10.11	8.49	0.29	-3.57
FOLD4	9.22	16.22	10.64	0.52	-5.01	7.66	14.23	8.97	0.59	-4.42	8.13	15.09	9.36	0.57	-4.77
FOLD5	8.31	13.25	10.77	0.11	-5.10	7.31	11.57	9.47	0.27	-5.28	7.18	11.59	9.30	0.24	-5.24

	PhysNet					PhysFormer				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	4.73	8.76	6.81	0.46	-0.09	5.44	9.44	7.87	0.38	-0.20
FOLD2	3.64	8.46	5.66	0.68	2.28	4.75	9.32	7.46	0.59	1.32
FOLD3	2.72	5.81	3.96	0.60	3.18	4.45	8.02	6.64	0.41	2.70
FOLD4	5.78	13.16	6.17	0.63	1.60	12.46	19.14	13.37	0.17	-7.44
FOLD5	4.54	8.89	5.72	0.40	3.15	5.00	9.12	6.32	0.33	1.69

Table A.5: HR performance metrics of all the NN methods for each Fold over the Driving test cases.

NN Models - HR															
	Driving Large Motion					Driving Small Motion					Driving Still				
	DeepPhys														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	4.73	8.76	6.81	0.46	-0.09	4.73	8.76	6.81	0.46	-0.09	4.73	8.76	6.81	0.46	-0.09
FOLD2	3.64	8.46	5.66	0.68	2.28	3.64	8.46	5.66	0.68	2.28	3.64	8.46	5.66	0.68	2.28
FOLD3	2.72	5.81	3.96	0.60	3.18	2.72	5.81	3.96	0.60	3.18	2.72	5.81	3.96	0.60	3.18
FOLD4	5.78	13.16	6.17	0.63	1.60	5.78	13.16	6.17	0.63	1.60	5.78	13.16	6.17	0.63	1.60
FOLD5	4.54	8.89	5.72	0.40	3.15	4.54	8.89	5.72	0.40	3.15	4.54	8.89	5.72	0.40	3.15
	TS-CAN														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20
FOLD2	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32
FOLD3	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70
FOLD4	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44
FOLD5	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69
	EfficientPhys-C														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	4.73	8.76	6.81	0.46	-0.09	4.73	8.76	6.81	0.46	-0.09	4.73	8.76	6.81	0.46	-0.09
FOLD2	3.64	8.46	5.66	0.68	2.28	3.64	8.46	5.66	0.68	2.28	3.64	8.46	5.66	0.68	2.28
FOLD3	2.72	5.81	3.96	0.60	3.18	2.72	5.81	3.96	0.60	3.18	2.72	5.81	3.96	0.60	3.18
FOLD4	5.78	13.16	6.17	0.63	1.60	5.78	13.16	6.17	0.63	1.60	5.78	13.16	6.17	0.63	1.60
FOLD5	4.54	8.89	5.72	0.40	3.15	4.54	8.89	5.72	0.40	3.15	4.54	8.89	5.72	0.40	3.15
	PhysNet														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20
FOLD2	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32
FOLD3	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70
FOLD4	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44
FOLD5	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69
	PhysFormer														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20	5.44	9.44	7.87	0.38	-0.20
FOLD2	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32	4.75	9.32	7.46	0.59	1.32
FOLD3	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70	4.45	8.02	6.64	0.41	2.70
FOLD4	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44	12.46	19.14	13.37	0.17	-7.44
FOLD5	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69	5.00	9.12	6.32	0.33	1.69

Table A.6: HR performance metrics of all the NN methods for each Fold over the Garage test cases.

NN Models - HR															
	Garage Large Motion					Garage Small Motion					Garage Still				
DeepPhys															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
FOLD1	3.52	5.86	5.11	0.60	-3.36	2.64	5.34	3.71	0.75	-1.99	0.24	0.93	0.37	0.99	9.86
FOLD2	9.59	14.68	12.36	0.46	-8.60	6.70	11.42	10.20	0.57	-4.00	2.12	7.16	3.51	0.81	3.75
FOLD3	10.84	15.04	15.72	0.01	-9.89	3.39	7.10	4.68	0.45	-2.15	0.70	2.11	1.06	0.94	7.16
FOLD4	20.80	26.21	23.43	-0.08	-12.68	6.74	13.46	7.64	0.51	-6.08	0.93	3.95	1.05	0.97	5.50
FOLD5	7.62	11.45	9.91	0.21	-9.06	5.47	9.98	7.02	0.27	-2.59	1.71	4.85	2.24	0.66	6.75
TS-CAN															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
FOLD1	3.91	7.86	5.71	0.52	-1.74	2.88	9.05	4.33	0.49	-0.86	0.24	0.93	0.35	0.99	11.02
FOLD2	9.52	14.35	12.88	0.40	-7.65	3.89	8.34	5.85	0.75	-1.66	0.48	1.39	0.76	0.99	5.09
FOLD3	5.86	8.85	8.63	0.19	-6.76	2.47	5.30	3.52	0.69	-1.24	0.61	1.46	0.94	0.97	8.53
FOLD4	12.89	19.82	14.08	0.20	-10.19	3.08	7.12	3.69	0.86	-4.78	0.59	1.44	0.83	1.00	6.73
FOLD5	10.06	13.63	13.06	0.23	-9.29	4.30	8.49	5.53	0.36	-1.33	2.15	6.39	2.75	0.45	5.51
EfficientPhys-C															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
FOLD1	2.93	5.56	4.24	0.69	-1.52	2.39	5.67	3.45	0.72	-0.30	0.24	0.93	0.36	0.99	10.47
FOLD2	9.52	14.69	12.79	0.34	-7.91	4.85	10.73	7.47	0.60	-1.92	1.25	5.32	2.20	0.88	5.01
FOLD3	6.25	10.38	9.46	0.10	-8.02	2.76	6.71	3.97	0.62	-1.14	0.66	1.52	1.00	0.97	8.40
FOLD4	13.87	20.62	15.40	0.25	-11.79	3.86	9.69	4.41	0.72	-3.68	0.39	1.17	0.56	1.00	6.46
FOLD5	8.69	12.25	11.23	0.17	-8.98	4.44	8.56	5.63	0.39	-2.25	2.25	6.45	2.95	0.49	5.25
PhysNet															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
FOLD1	1.37	3.42	2.07	0.84	3.91	0.24	0.93	0.35	0.99	7.69	0.10	0.59	0.14	1.00	13.83
FOLD2	5.79	11.76	7.95	0.56	-0.46	1.72	5.60	2.85	0.86	4.59	0.59	1.68	0.96	0.99	8.60
FOLD3	2.44	4.27	3.46	0.75	0.88	0.96	2.57	1.36	0.91	6.53	0.28	1.71	0.44	0.96	12.72
FOLD4	12.30	21.00	12.57	0.12	-5.85	4.59	11.65	4.80	0.59	1.80	0.39	1.17	0.51	1.00	10.61
FOLD5	5.27	9.63	6.63	0.32	0.65	2.29	6.10	2.91	0.57	7.08	0.39	1.44	0.55	0.97	13.15
PhysFormer															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	**ρ↑**	**SNR↑**
FOLD1	3.03	5.77	4.48	0.66	4.53	1.07	2.75	1.59	0.92	6.46	0.24	0.93	0.36	0.99	14.21
FOLD2	7.40	12.19	10.78	0.50	-2.29	2.85	6.46	4.44	0.80	4.34	2.34	7.42	4.27	0.77	7.57
FOLD3	5.76	9.25	8.58	0.07	-0.25	2.18	4.91	3.02	0.64	7.10	0.66	2.07	1.01	0.93	12.80
FOLD4	13.48	18.93	14.54	0.25	-13.83	11.33	18.74	12.02	-0.12	-7.89	11.67	20.39	11.73	0.06	-0.38
FOLD5	5.66	10.28	7.14	0.26	-0.41	3.13	6.96	3.93	0.46	4.61	1.03	2.78	1.40	0.88	11.06

A.2.2 Respiration Rate results

Table A.7: Overall RR performance metrics of all the NN methods for each Fold.

NN Models - RR															
	DeepPhys					TS-CAN					EfficientPhys-C				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	10.28	13.37	165.59	0.09	6.43	10.79	13.63	171.49	-0.06	7.14	10.56	13.58	169.27	-0.01	6.72
FOLD2	12.00	14.41	194.79	-0.01	4.83	9.69	13.08	159.44	0.08	7.20	9.57	12.83	153.69	0.14	6.73
FOLD3	10.23	13.05	164.78	0.12	5.28	10.45	13.32	168.89	0.10	5.10	10.51	13.60	166.89	0.08	5.46
FOLD4	11.41	14.16	179.96	0.11	4.26	10.66	13.82	162.75	-0.02	5.04	10.22	13.38	155.33	0.03	4.74
FOLD5	10.85	13.53	170.80	0.04	5.34	10.23	13.15	160.66	0.12	6.05	11.71	14.20	188.72	-0.09	5.38

	PhysNet					PhysFormer				
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	3.37	6.73	47.11	0.18	16.24	4.53	7.72	68.28	0.07	15.52
FOLD2	2.26	5.35	31.86	0.09	19.19	2.91	6.02	40.34	0.03	17.36
FOLD3	3.87	7.15	43.28	0.23	14.96	3.32	6.56	34.87	0.03	16.58
FOLD4	4.11	8.10	50.34	-0.01	14.68	5.82	9.74	71.52	-0.08	12.93
FOLD5	1.98	4.46	24.34	0.28	19.15	3.93	7.41	55.42	-0.10	17.36

Table A.8: RR performance metrics of all the NN methods for each Fold over the Driving test cases.

NN Models - RR															
	Driving Large Motion					Driving Small Motion					Driving Still				
	DeepPhys														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	8.35	11.77	127.71	0.26	8.46	9.81	12.93	164.60	0.23	4.71	12.92	15.38	204.48	-0.12	4.60
FOLD2	13.58	16.02	214.89	-0.21	2.97	13.33	15.26	208.76	0.07	4.03	12.55	14.40	208.33	-0.16	4.93
FOLD3	10.42	13.35	154.47	0.07	4.72	10.55	13.06	162.58	0.15	5.28	11.27	13.82	188.99	0.07	4.49
FOLD4	10.30	13.17	175.40	0.28	2.37	14.26	16.21	217.85	0.00	1.93	9.57	12.86	154.17	0.30	5.64
FOLD5	9.76	13.46	162.50	0.31	3.57	13.55	15.43	202.00	-0.22	2.65	10.81	13.32	174.02	0.10	4.36
	TS-CAN														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	12.16	14.18	198.33	-0.01	5.87	12.94	15.01	202.08	-0.17	3.31	11.39	14.31	181.08	-0.13	7.58
FOLD2	10.61	14.10	178.05	0.17	6.74	9.33	12.46	144.87	0.08	7.14	8.24	11.62	133.88	0.03	8.24
FOLD3	10.48	13.69	167.60	0.32	3.89	10.94	13.54	171.34	0.02	3.25	12.03	14.51	200.08	-0.09	3.98
FOLD4	12.97	15.35	203.62	-0.20	1.24	13.01	15.58	192.47	0.05	2.97	9.33	12.91	138.52	-0.05	6.06
FOLD5	9.23	13.05	162.83	0.23	4.39	10.55	13.15	156.32	0.13	3.83	11.01	13.82	169.71	0.21	6.53
	EfficientPhys-C														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	13.18	15.12	211.18	0.08	4.64	11.85	14.29	187.24	0.01	5.05	10.78	13.66	171.87	-0.16	6.43
FOLD2	10.79	14.10	169.66	0.33	4.77	9.19	12.39	142.41	0.11	6.49	8.46	12.09	137.66	0.03	8.23
FOLD3	11.85	14.74	185.37	-0.01	3.82	11.56	14.51	175.92	0.08	3.08	10.71	13.48	180.20	-0.09	6.34
FOLD4	14.15	16.31	216.42	-0.20	1.69	11.50	13.95	171.96	0.13	2.15	9.28	12.71	139.17	-0.06	5.20
FOLD5	12.48	15.09	218.50	0.20	3.29	12.25	14.21	176.53	-0.09	3.99	10.55	13.52	172.60	-0.09	6.36
	PhysNet														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	5.92	9.62	78.13	0.16	10.64	5.18	9.06	75.26	-0.11	13.18	2.62	5.14	34.04	0.54	15.81
FOLD2	2.70	6.42	24.72	0.29	17.17	3.73	7.41	55.01	-0.17	17.55	2.21	5.15	33.10	0.28	19.14
FOLD3	5.04	8.52	46.50	-0.04	13.22	5.85	9.46	56.11	0.19	9.92	3.69	7.03	48.29	0.31	14.55
FOLD4	4.31	8.33	46.74	-0.12	12.32	4.63	8.86	60.53	-0.09	14.45	4.15	8.13	41.54	0.17	14.35
FOLD5	2.48	5.65	28.86	-0.12	17.05	2.71	5.77	27.21	0.06	14.56	2.08	4.08	25.00	0.70	20.14
	PhysFormer														
Train Set:	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑	MAE↓	RMSE↓	MAPE↓	ρ↑	SNR↑
FOLD1	5.08	7.83	73.83	-0.14	13.79	5.75	9.26	90.80	0.23	12.13	5.34	8.29	78.98	-0.10	14.67
FOLD2	2.89	6.21	35.65	-0.02	15.64	3.59	6.62	45.36	0.01	16.06	2.77	5.80	42.00	0.22	18.34
FOLD3	3.16	5.94	32.23	0.38	14.21	5.43	8.89	47.07	-0.10	12.09	2.83	6.18	27.65	-0.13	17.89
FOLD4	8.07	12.24	107.87	-0.05	10.36	5.06	8.36	54.00	-0.04	12.18	6.90	10.71	90.21	-0.26	10.61
FOLD5	7.59	11.65	118.56	-0.21	10.94	4.80	8.25	64.74	-0.22	14.58	3.25	7.00	38.16	0.08	18.09

Table A.9: RR performance metrics of all the NN methods for each Fold over the Garage test cases.

NN Models - RR															
	Garage Large Motion					Garage Small Motion					Garage Still				
DeepPhys															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
FOLD1	13.77	16.35	211.11	-0.49	6.06	11.50	14.09	194.79	-0.09	5.57	5.20	8.30	79.03	0.23	12.50
FOLD2	11.43	13.98	182.45	0.17	3.66	9.29	12.61	148.45	0.12	6.42	9.56	12.80	169.17	-0.25	7.07
FOLD3	11.87	14.65	208.33	0.08	2.29	9.17	12.25	151.85	0.17	4.18	8.09	11.20	130.40	-0.09	9.94
FOLD4	12.30	15.00	186.67	-0.21	0.64	10.99	14.09	176.49	0.00	6.23	10.99	13.24	167.50	0.29	7.05
FOLD5	13.92	15.42	221.67	-0.14	4.93	9.01	11.81	137.99	-0.30	8.52	8.64	11.74	140.21	0.18	8.71
TS-CAN															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
FOLD1	8.06	10.61	115.00	0.00	9.27	12.30	15.02	202.99	0.04	5.86	3.66	7.05	59.31	0.04	15.74
FOLD2	12.52	15.29	217.47	0.10	3.80	8.04	12.05	132.26	0.05	8.81	12.52	15.48	216.67	-0.05	5.92
FOLD3	13.18	14.85	212.64	0.09	1.43	11.24	13.92	191.55	-0.08	4.46	4.64	7.77	65.07	0.39	13.86
FOLD4	12.45	15.21	213.89	0.28	5.82	11.57	14.78	176.25	-0.22	2.50	5.42	8.13	80.90	0.18	11.97
FOLD5	8.35	11.29	114.72	0.52	7.28	10.40	13.03	157.57	-0.38	7.03	10.33	13.23	177.71	-0.12	8.04
EfficientPhys-C															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
FOLD1	10.99	14.22	167.22	-0.33	4.79	9.89	13.13	163.82	0.05	5.89	4.61	9.35	77.15	-0.05	14.49
FOLD2	10.88	13.90	181.25	0.21	5.06	7.60	11.18	126.55	-0.07	8.88	9.94	12.94	164.38	0.19	5.51
FOLD3	10.55	13.52	168.61	0.21	4.04	11.36	14.45	185.89	0.30	5.35	6.82	10.65	103.20	0.18	10.61
FOLD4	9.67	13.33	158.89	0.39	5.55	10.91	13.83	161.54	-0.25	3.33	6.01	9.52	91.25	0.16	11.43
FOLD5	15.82	17.19	254.44	-0.26	3.27	8.13	10.75	117.29	-0.16	7.04	11.35	14.41	198.26	-0.20	7.58
PhysNet															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
FOLD1	2.20	4.45	36.11	-0.10	21.42	1.25	1.93	18.13	0.26	20.90	0.66	1.48	9.93	0.62	23.16
FOLD2	1.21	1.92	18.54	0.28	18.97	0.69	1.37	11.25	0.57	20.80	0.77	1.58	12.08	0.56	23.32
FOLD3	0.88	1.60	12.22	-0.20	17.17	3.01	5.10	46.90	0.08	19.12	1.62	3.27	17.82	0.76	21.16
FOLD4	3.96	7.78	43.89	-0.15	10.59	5.35	9.59	74.10	-0.14	14.32	1.83	3.95	30.00	-0.05	20.19
FOLD5	0.73	1.52	9.44	0.51	19.08	1.46	2.15	22.50	0.19	21.13	1.39	4.14	24.03	0.09	24.65
PhysFormer															
Train Set:	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**	**MAE↓**	**RMSE↓**	**MAPE↓**	$\rho\uparrow$	**SNR↑**
FOLD1	2.20	4.03	26.11	0.34	19.14	4.03	7.50	61.11	-0.21	18.40	1.03	1.76	15.97	0.43	22.37
FOLD2	4.72	8.73	60.08	-0.19	14.19	2.45	5.57	41.55	-0.11	19.15	1.26	2.93	19.79	0.52	19.60
FOLD3	5.86	9.60	71.94	-0.12	12.14	1.82	4.14	29.94	-0.06	20.82	1.20	1.96	18.40	0.39	22.15
FOLD4	8.50	12.70	134.44	-0.21	10.15	5.42	9.05	49.16	-0.04	14.88	2.34	5.76	27.85	0.50	19.85
FOLD5	2.78	5.44	49.86	-0.21	20.70	3.44	5.71	47.85	-0.11	18.90	1.39	2.18	21.60	0.10	23.16