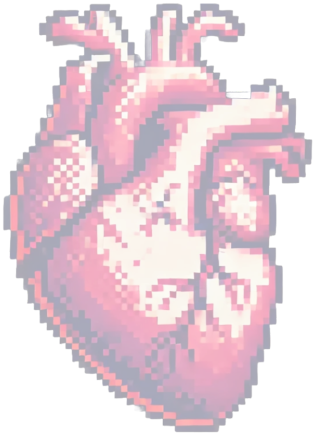
From Pixels to Pulse

Enhancing Trust, Quality and Robustness in Remote Video-Based Pulse Measurement



Eirini Kateri

UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

From Pixels to Pulse: Enhancing Trust, Quality and Robustness in Remote Video-Based Pulse Measurement

by

Eirini Kateri

ORCiD: 0000-0003-4777-0814

A thesis for the degree of
Doctor of Philosophy

November 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

From Pixels to Pulse: Enhancing Trust, Quality and Robustness in Remote Video-Based Pulse Measurement

by Eirini Kateri

Remote photoplethysmography (rPPG) enables non-contact heart rate measurement using everyday cameras, offering a promising alternative to traditional contact-based methods like electrocardiography and photoplethysmography. By leveraging subtle changes in skin color and micro-movements induced by blood flow, rPPG has the potential to revolutionize health monitoring. However, despite its potential, challenges such as motion artifacts, variations in lighting conditions and dataset biases challenge its robustness and reliability. This thesis investigates the foundations of rPPG, presenting a comprehensive study across signal processing, machine learning, video quality assessment and uncertainty quantification. We explore traditional signal processing techniques for rPPG, establishing a baseline for pulse estimation while highlighting their sensitivity to motion artifacts. To address these limitations, we propose a novel spatiotemporal two-stage learning framework (ST2S-rPPG), which integrates video stabilization, machine learning and adaptive region of interest selection to enhance pulse estimation accuracy. Recognizing the influence of video quality on rPPG performance, we systematically analyze the impact of motion, resolution, illumination and occlusions among other video quality factors, introducing video quality metrics tailored to rPPG. These metric provide a structured approach to assess video suitability for pulse extraction. Finally, we explore the application of conformal predictions to rPPG, establishing a framework for uncertainty quantification and compare MAE-based and quality-aware nonconformity measures. The findings of this thesis contribute toward making rPPG more practical for real-world deployment, with applications ranging from remote patient monitoring and telehealth to mental health assessments and human-computer interaction. While challenges remain in generalization across diverse populations and environmental conditions, these advancements lay a foundation for future research in making rPPG a reliable and scalable tool for healthcare and beyond.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: 1. Kateri Eirini, and Katayoun Farrahi. "Video-Based Pulse Estimation through Spatiotemporal Meta-Learning." Proc. MobiUK (2024), 2. Kateri Eirini, and Katayoun Farrahi. "ST2S-rPPG: A Spatiotemporal Two-Stage Learning Approach for Pulse Estimation Using Video." Machine Learning for Health. PMLR, 259:550–562, 2024.
8. Code for the submissions can be found in: *https://github.com/eirkateri/ST2S-RPPG*

Signed:..... Date:.....

Acknowledgements

Completing this PhD has been a journey filled with challenges and growth. I was fortunate to have had the opportunity to research something I remain deeply passionate about. I was incredibly lucky to be surrounded by brilliant, kind people who supported me through this turbulent (to say the least) time. I will never forget your kindness. I would like to take this space to express my gratitude to all those who made this possible and helped me achieve what I have.

First and foremost, to my supervisor, Kate Farrahi - thank you for offering me the opportunity to pursue this PhD and for your support. To Adam Prügel-Bennett - thank you for always having my back and for encouraging me to challenge my own boundaries. To Jon Hare - thank you for always making time to listen to my concerns despite your busy schedule. To everyone in the Vision, Learning and Control group - you were my family these past few years. Thank you for listening to my crises, for pulling me back to reality when I was spiraling, for taking me out for drinks and for all the fun times we shared. Daniela, Harry, Millie, Jay, Peter, Alex, Rachel, Frixos, Ioan, Mona and everyone else — spending my days with you was truly a gift.

To my housemates and chosen family Matt, Grace, Alex, Kai and Paolo - from late-night work rants to summer barbecues and trips to Greece, you made my life outside the lab feel like a dream, thank you. To my family - thank you for supporting me in every way possible since the beginning of my life. I would like to especially mention my uncle Dimitris and my grandma Rena, who we lost during my PhD - thank you for looking over me. You are missed beyond words. Finally, and most importantly, to my partner, Devon - you were my rock. I would not be where I am without you. Through every failure and success, and then back to the same cycle, I knew I had someone by my side to take long evening walks to put everything into perspective and dream about a future filled with sausage dogs.

Finally, I wish to highlight some of the things I accomplished during the past four years. When things seem hard in the future (as they surely will!), I can look back at this and remember that even through the hardest times, I always made it through. I am grateful for the opportunities to present my research, travel the world, explore entrepreneurial journeys, be recognized by my research group, write a fairly successful newsletter, organize talks, events, a conference, parties, trips and activities, feature in Forbes, meet friends for life and live some of the most creative, productive and fun years of my life.

Thank you to everyone who was part of this wild ride - and to myself (as Snoop Dog would say), for never quitting!

Abbreviations

Specialized Terms	
$2D$	Two Dimensional
$3D$	Three Dimensional
AI	Artificial Intelligence
$AROI$	Adaptive ROI
BPM	Beats per Minute
BP	Blood Pressure
BVP	Blood Volume Pulse
$DROI$	Dynamic ROI
ECG	Electrocardiogram
FPS	Frames per Second
GPU	Graphics Processing Unit
HR	Heart Rate
HRV	Heart Rate Variability
ICA	Independent Component Analysis
ICU	Intensive Care Unit
MAE	Mean Absolute Error
$MAPE$	Mean Absolute Percentage Error
ME	Mean Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
$PIPs$	Persistent Independent Particles
PPG	Photoplethysmography
RGB	Red, Green, Blue Channels
$RMSE$	Root Mean Squared Error
ROI	Region of Interest
SD	Standard Deviation
SPO_2	Arterial Oxygen Saturation
TCM	Traditional Chinese Medicine
Models & Algorithms	
BSS	Blind Source Separation

$BS-rPPG$	Blind-Signal rPPG
$CHROM$	Chrominance-Based Remote Photoplethysmography
CNN	Convolutional Neural Network
CP	Conformal Prediction
DTW	Dynamic Time Warping
GAN	Generative Adversarial Network
ICP	Inductive Conformal Predictions
KNN	K-Nearest Neighbor
$MDTW$	Moving Dynamic Time Warping
$ML-QM$	ML Quality Metric
$MP-rPPG$	Motion-Pulse rPPG
$MTTS-CAN$	Multi-task Temporal-Shift Convolutional Attention Network
NLP	Natural Language Processing
PCA	Principal Component Analysis
POS	Plane-Orthogonal-to-Skin
$PSNR$	Peak Signal-to-Noise Ratio
RNN	Recurrent Neural Network
SCP	Split Conformal Predictions
SNR	Signal-to-Noise Ratio
SSA	Singular Spectrum Analysis
$SSIM$	Structural Similarity Index Measure
$STVEN$	Spatiotemporal Video Enhancement Network
TCP	Transductive Conformal Prediction
TDT	Temporal Difference Transformer
$TS-CAN$	Temporal-Shift Convolutional Attention Network
VIF	Variance Inflation Factor
$WS-QM$	Weighted Sum Quality Metric
$XGBoost$	Extreme Gradient Boosting
$ST2S$	Spatiotemporal Two-Stage Model
Datasets	
$MMSE-HR$	Multimodal Spontaneous Expression - Heart Rate dataset
$O-HR$	Ostankovich et. al dataset for video-based pulse detection
$UBFC-rPPG$	Universite Bourgogne Franche-Comte dataset for rPPG
$COHFACE$	Facial video dataset collected by the Idiap Research Institute

Prologue

"The pulse of a man is the truest index of his emotions." – Jane Austen

Computer vision has rapidly evolved over the past decade, progressing from object recognition to advanced applications in healthcare, autonomous systems and beyond. One of its most promising frontiers? The measurement of human health using only a video. A standard webcam or a phone camera has the potential to extract vital physiological information like pulse, without the need for physical contact, wearable sensors or specialized medical equipment.

This technology has the potential to transform how we monitor vital signs, offering a seamless, non-contact alternative to conventional sensors. In neonatal intensive care units it could reduce the need for adhesive sensors that may cause discomfort or skin irritation in fragile newborns. For elderly individuals with limited mobility, it allows early detection of cardiovascular issues while preserving their independence.

Beyond individual care, its impact can extend to broader healthcare systems. The Covid-19 pandemic underscored the urgent need for remote health monitoring, not just for convenience but as protection. Reducing direct contact between healthcare workers and patients minimized infection risks, and contactless technologies like this could further enhance patient safety in future outbreaks. In low-resource settings where access to medical equipment is limited, it can offer a cost-effective and scalable alternative for tracking vital signs using available cameras.

Despite its promise, remote Photoplethysmography (rPPG) presents significant challenges. It relies on detecting minor color variations or subtle displacements on the skin caused by blood flow, making it highly sensitive to external factors. Changes in lighting conditions, motion, skin tone variations and camera characteristics (such as resolution and frame rate among others) can all affect its accuracy. Unlike conventional medical devices, like the pulse oximeter, which operate in controlled environments, rPPG must function reliably in real-world settings where these variables are unpredictable. Developing a robust system that can generalize across diverse conditions remains one of the most pressing challenges in the field.

Since the pioneering work of Verkruysse et al. (2008), rPPG research has expanded rapidly. Early methods primarily relied on classical signal processing techniques to extract pulse signals, but these approaches struggled with robustness against environmental noise. More recent advancements have integrated deep learning, significantly improving accuracy and resilience to variations in lighting, motion and skin tone. Yet, despite significant progress, the technology is still far from reaching its full potential. The challenge is no longer proving that rPPG works - it is ensuring that it works consistently, accurately and in ways that are both scalable and accessible for widespread adoption in the wild.

At its core, rPPG represents more than a technical innovation; it is a step toward a future where health monitoring is seamless, non-invasive and widely available. A world where basic vital sign assessments do not require a hospital visit, where continuous patient monitoring is effortless and where healthcare becomes more accessible through the power of computer vision. While significant challenges remain, the impact of rPPG could be transformative.

This is why rPPG is so exciting. And this is why it's worth pursuing.

Contributions

This research focuses on pulse estimation, driven by the recognition that cardiovascular health is both a critical and challenging area of study. Traditional cardiovascular measurements often require contact-based sensors, limiting their accessibility and usability in continuous or remote settings. The complexity of extracting meaningful cardiovascular signals from video due to factors such as motion, lighting variations and skin tone differences, presents both a technical challenge and an opportunity for innovation. Given its fundamental role in assessing cardiovascular function and its feasibility for non-contact measurement, pulse emerged as an ideal starting point for exploring how rPPG can contribute to more accessible and scalable health monitoring solutions. Our research has made several contributions to the field of rPPG, listed below:

- A comprehensive analysis of signal processing techniques for remote pulse estimation, exploring how traditional methods can be optimized to improve heart rate estimation from video (Chapter 2).
- The introduction of a novel spatiotemporal two-stage learning approach that bridges traditional signal processing with machine learning, incorporating stabilization techniques and feature selection mechanisms (Chapter 3).
- An extensive analysis of the impact of video quality factors (e.g., blur, illumination, motion) on rPPG model performance, highlighting their effects on accuracy and reliability (Chapter 4).

- The development of novel video quality metrics tailored to rPPG, enabling better evaluation of algorithmic robustness across varying conditions (Chapter 4).
- The application of conformal prediction to rPPG, establishing a framework for uncertainty quantification in pulse estimation and improving confidence in predictions (Chapter 5).
- The introduction of our video quality metric as a nonconformity measure for conformal predictions, demonstrating its effectiveness in capturing data variability in rPPG models. (Chapter 5)

By addressing challenges in signal extraction, video quality effects and confidence estimation, this research contributes toward making rPPG a more robust and practical tool for real-world applications.

Publications

The contributions of this PhD research that have been published or are intended for submission are listed below.

Published

- Kateri Eirini, and Katayoun Farrahi. "Video-Based Pulse Estimation through Spatiotemporal Meta-Learning." Proc. MobiUK (2024)
- Kateri Eirini, and Katayoun Farrahi. "ST2S-rPPG: A Spatiotemporal Two-Stage Learning Approach for Pulse Estimation Using Video." Machine Learning for Health. PMLR, 259:550–562, 2024.

Planned Submission

Two journal papers expanding upon this work are currently being prepared for submission to the *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*.

- Kateri Eirini, Katayoun Farrahi, and Adam Prugel-Bennett. "Quantifying the effects of video quality on rPPG algorithms"
- Kateri Eirini, Katayoun Farrahi, and Adam Prugel-Bennett. "Uncertainty Quantification for Remote Photoplethysmography Using Conformal Predictions"

Thesis Structure

The title of this thesis, "From Pixels to Pulse: Enhancing Trust, Quality and Robustness in Remote Video-Based Pulse Measurement," captures the main themes explored in this research. Trust in this context refers to the degree to which rPPG systems can be relied upon to deliver accurate and transparent results across diverse scenarios. It is defined through measurable factors such as model accuracy, stability across conditions and the quantification of uncertainty using conformal prediction methods. By providing interpretable confidence estimates and identifying reliable inputs, the system becomes more dependable and transparent to end users. Quality captures the suitability of video data for reliable pulse extraction. This thesis introduces a video quality metric designed to assess how motion, illumination and blur affect rPPG signal integrity, allowing poor-quality data to be detected or filtered before analysis. This ensures that predictions are based on trustworthy visual input. Robustness describes the system's ability to maintain reliable performance under variations in lighting, motion and device conditions. By combining signal processing and deep learning approaches with video quality awareness, this work enhances robustness across datasets and experimental setups. Together, these three dimensions form the foundation for building video-based physiological measurement systems that are not only more accurate but also interpretable, reliable and ready for real-world use.

This thesis is structured to reflect the building blocks of a machine learning model, where each chapter plays a critical role in shaping the final outcome (Figure 1). Chapter 1 serves as the pre-processing step, filtering through history and existing literature to extract meaningful insights and setting the foundation for next steps. Chapter 2 reflects the input layer, where signal processing approaches are implemented to process raw video data and transform them into physiological signals. Chapter 3 acts as the hidden layers, with the proposed method capturing complex patterns and refining the extracted features. Chapter 4 serves as the output layer, where video quality metrics assess performance and reliability. Finally, chapter 5 functions as a calibration step, ensuring the models' confidence is well-calibrated and providing a measure of reliability using conformal predictions. This approach mirrors the way machine learning systems are designed, layer by layer, extracting insights at every step of the way to progressively refine knowledge.

Chapter 1: Pre-processing: History and Foundations for Pulse Measurement

This chapter introduces the fundamental concepts behind rPPG, providing historical context and physiological principles. It covers the evolution of pulse measurement, the physiological mechanisms underlying rPPG and the challenges of measuring heart rate using video. This chapter lays the groundwork for the subsequent technical contributions.

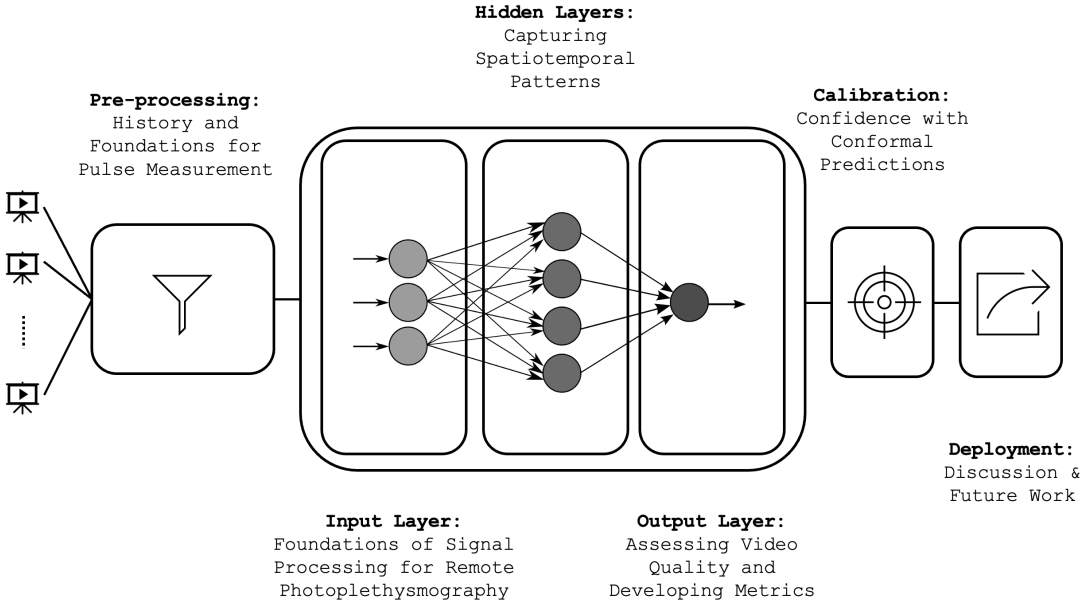


FIGURE 1: Visualization of the thesis structure, reflecting the building blocks of a Machine Learning model.

Chapter 2: Input Layer: Foundations of Signal Processing for Remote Photoplethysmography

We explore traditional signal processing approaches for pulse estimation, focusing on motion-based methods. This chapter details our adaptation and customization of existing methods for rPPG, incorporating adjustments to fit the specific characteristics of the data. We introduce enhancements, such as the integration of Persistent Independent Particles for feature tracking and K-means clustering for improved signal extraction.

Chapter 3: Hidden Layers: Capturing Spatiotemporal Patterns

This chapter introduces a novel spatiotemporal two-stage learning approach that integrates signal processing with machine learning. We present a framework that stabilizes video inputs, extracts meaningful temporal patterns and refines pulse estimation through machine learning techniques. Comparative evaluations highlight the strengths of this approach in improving accuracy and robustness.

Chapter 4: Output Layer: Assessing Video Quality and Developing Metrics

This chapter systematically examines how video quality factors such as blur, illumination changes, motion artifacts and resolution affect rPPG model performance. We propose and validate novel video quality metrics tailored to rPPG, providing a structured way to assess the reliability of pulse estimation in diverse real-world conditions.

Chapter 5: Calibration: Confidence with Conformal Predictions

This chapter introduces the application of conformal predictions to rPPG, establishing a framework for quantifying uncertainty in heart rate estimation. Given the inherent challenges of rPPG, such as motion artifacts, lighting variations and unpredictable video conditions, ensuring that predictions are both accurate and well-calibrated is critical for real-world applications. We explore different nonconformity measures, first applying conformal predictions with mean absolute error and then integrating a quality-aware metric developed in Chapter 4.

Chapter 6: Deployment: Discussion and Future Work

The final chapter summarizes the key contributions of this thesis and reflects on their implications for the future of rPPG. We discuss open challenges, potential applications and directions for future research, emphasizing how advancements in video-based pulse measurement can contribute to broader healthcare and human-computer interaction contexts.

Chapter 1

Pre-processing: History and Foundations for Pulse Measurement

Just as preprocessing sets the stage for effective learning in a model, understanding the historical and foundational context of pulse measurement illuminates the path for modern advancements in remote photoplethysmography.

Throughout history, humanity has strived to extract meaning from subtle body signals. While earlier civilizations practiced medicine in various forms, a marked shift can be seen in ancient Greece, where observation and diagnosis became central, moving away from the belief that illness was a punishment from the gods. This shift to rational medicine marked the beginning of a new way of understanding the body. Diagnosis started to rely on the body's rhythms, patterns and signals. Among these signals, the concept of vital signs began to emerge. Even though they were not formally introduced into clinical practice until the late 1800s, ancient physicians had already observed the correlation between illness, high pulse and variable breathing patterns. Pulse, in particular, has captivated physicians, philosophers, even poets throughout history, standing out as one of the earliest vital signs to be observed, studied and documented [Elsberg (1931)]. Its rhythmic beat was linked to the mysterious inner workings of the body, offering a window into health and disease long before diagnostic tools existed. Unlike other vital signs that required advanced instruments or indirect observation, pulse could be assessed with a simple touch, making it one of the most accessible indicators of physiological state. Ancient physicians relied on it to infer the heart's function, the

body's balance, even the patient's emotional state, cementing its role as a foundation of early medical practice.

Pulse has been celebrated in literature, poetry and art, becoming a metaphor for life and emotion, inspiring poets and writers across civilizations and time. It was often described as a reflection of passion or anxiety, with its rate tied to human experiences. While unique to each individual, pulse remains a universal biological rhythm. Ancient physicians were drawn to pulse because it was observable and changing in response to disease, physical strain or emotional states. This adaptability made it an excellent indicator of health long before scientific measurements and tools became available.

This chapter explores the rich history of pulse measurement, from its early interpretations to its evolution into a critical diagnostic tool. By understanding how pulse shaped early medical thought, we can appreciate its continuing role in modern medicine and its connection to life.

1.1 Tracing the Pulse Through Time

1.1.1 Pulse in Early Medicine

The ancient Greeks were among the first to study pulse as a diagnostic tool, associating its patterns to malady. Hippocrates recognized the importance of observing the body's rhythms, and though his understanding was rooted in the humoral theory, he noted that pulse variations could indicate illness or internal imbalances [Craik (2014)]. A strong, steady pulse was seen as a sign of health whereas a weak, irregular pulse was associated with disease or impending death. Galen built a more structured pulse diagnosis framework, based on Hippocrates' observations. He categorized pulse into four variables: magnitude, speed, frequency and regularity (or irregularity) [Wallis (2000)]. He even theorized that pulse could reflect emotional states, such as fear or anger, highlighting its connection to the nervous system. Galen also emphasized the role of the arteries and the heart in generating pulse, refining earlier anatomical knowledge.

While the Greeks viewed pulse through humoral imbalance, Traditional Chinese Medicine (TCM) developed a holistic approach to pulse. It became a central tool for assessing the body's internal state. Pulse was thought to reflect the balance of yin and yang and the flow of qi (vital energy). Disruptions in this balance were believed to be associated with illness, and pulse offered clues about the nature of the imbalance. Pulse diagnosis in TCM involved a complex system for assessment. Physicians examined pulse at three positions on each wrist: cun, guan and chi, by applying different levels of pressure (Figure 1.1). Each position corresponded to a specific organ, such as the heart, liver, kidneys, spleen or lungs [Velik (2015)]. Unlike Greek medicine, which categorized pulse mainly based on rhythm and strength, Chinese medicine identified 28

distinct pulse qualities including floating, sunken, slow, rapid, surging, fine and vacuous among others. For example, a floating pulse indicated external conditions like colds or fevers.

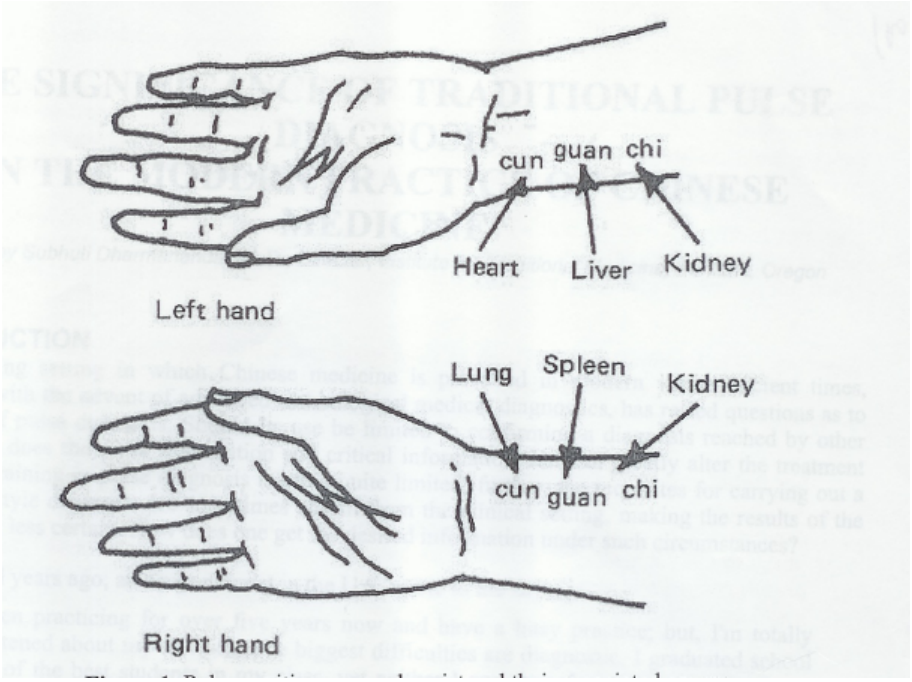


FIGURE 1.1: Distribution of organs in the six wrist positions for examining pulse and correlating it to different organs [Tang (2012)].

Unlike Greek humoral theories, which faded with time, Chinese pulse diagnosis has remained a vital practice in contemporary medicine, due to its adaptability, depth and holistic approach.

1.1.2 From Fingers to Instruments: Measuring Pulse Over Time

For centuries, pulse assessment relied solely on touch. While effective for detecting abnormalities, this method lacked precision, prompting efforts to develop objective measurement tools. This marked a significant shift from intuitive observation to quantitative analysis.

The first steps toward automation came in the 17th century with Santorio Santorio's pulsilogium [Bigotti and Taylor (2017)], a pendulum-based device with which he attempted to standardize pulse assessment. This invention represented an early effort to translate pulse into measurable data. However, the pulsilogium was impractical for clinical use. It relied on a pendulum, whose height had to be adjusted per patient to match their pulse rate. Due to its size and setup complexity, medical professionals preferred manual pulse palpation, which was simpler, faster and sufficiently reliable for most diagnoses. As a result, pulse assessment remained primarily physical for the next two centuries.

Building on William Harvey's 17th-century discovery of the circulatory system [Ribatti (2009)], physicians began to classify pulse irregularities and relate them to specific conditions. Jean-Baptiste Bouillaud, known as the "Father of Modern Cardiology," was one of the first to establish a link between pulse irregularities and rheumatic heart disease, contributing to early cardiovascular diagnostics [Silverman (1996)]. Advanced pulse understanding led to greater insights into cardiovascular health. Physicians began routinely assessing pulse rate, its rhythm and strength, correlating them with diseases like hypertension and heart failure. The invention of the stethoscope by René Laennec in 1816 added a new dimension to pulse assessment by enabling physicians to listen to heart sounds (Figure 1.2). Though primarily a tool for auscultation (listening to the internal sounds of the body), the stethoscope allowed physicians to deepen their understanding of cardiovascular function. It also played a key role in identifying conditions like valvular heart disease, where abnormal sounds correlated with an irregular heart rate (HR).

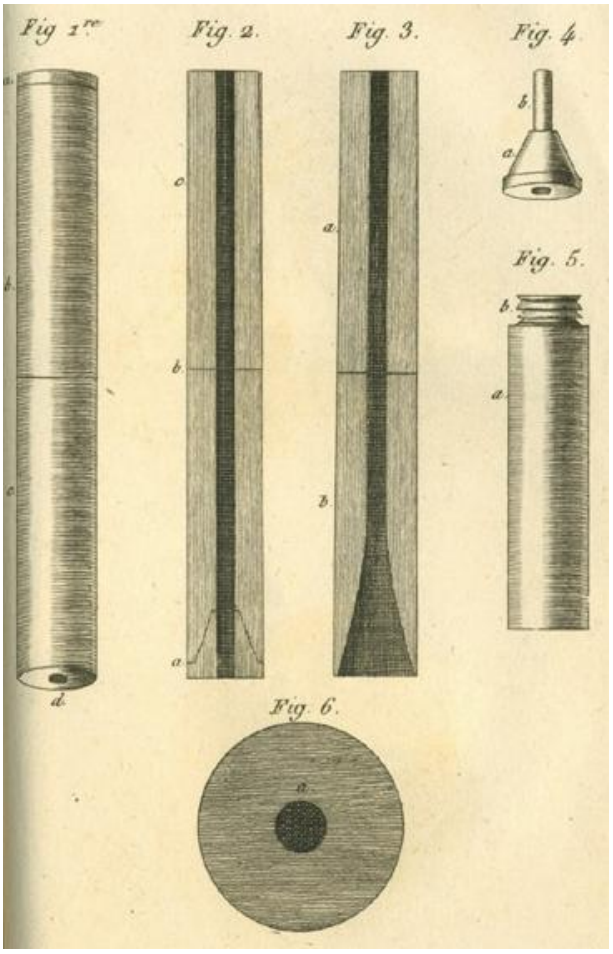


FIGURE 1.2: The first drawing of a stethoscope. 1) instrument assembled, 2) and 3) two portions of the instrument in longitudinal section, 4) detachable chest piece, 5) ear piece unscrewed, 6) transverse section [Roguin (2006)]. Original image courtesy of the US National Library of Medicine.

In the mid-19th century, the development of the sphygmograph by Karl von Vierordt

introduced the concept of pulse waveform analysis [Dudgeon (1882)]. This mechanical device recorded pulse waveforms as graphical tracings, offering a glimpse into arterial elasticity, blood flow and vascular health. While innovative, the sphygmograph was cumbersome and required significant operator expertise, something that limited its practical application and widespread adoption. Nevertheless, it bridged the gap between manual palpation and automated devices, introducing the concept of pulse waveform analysis as a diagnostic tool.

Building on Vierordt's work, Étienne-Jules Marey refined the sphygmograph in the late 19th century, making it more portable and clinically relevant [da Fonseca et al. (2014)] (Figure 1.3). Marey emphasized the diagnostic value of pulse waveforms, particularly in identifying arterial stiffness and cardiovascular disease. His advancements deepened the understanding of pulse propagation and its relationship to blood pressure and cardiac output, laying the foundation for future technologies.

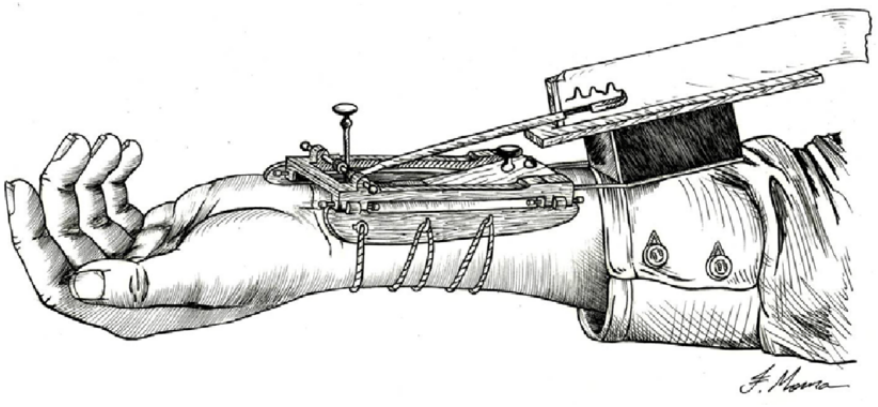


FIGURE 1.3: The Sphygmograph developed by Etienne Jules Marey in 1860 [da Fonseca et al. (2014)].

1.1.3 The 20th Century Pulse Revolution

The 20th century saw a technological revolution in medicine, driven by advances in physics, engineering and computing. These innovations introduced electrical devices that enhanced the ability to measure and monitor vital signs with high precision. Specifically, Willem Einthoven's invention of the electrocardiograph (ECG) in 1903 transformed cardiology [Cajavilca and Varon (2008)]. By measuring the heart's electrical activity, it enabled the detection of arrhythmia (irregular heartbeat), ischemia (reduced blood flow to tissues) and other cardiovascular abnormalities. His use of a string galvanometer to record electrical signals from the heart (Figure 1.4) earned him the Nobel Prize in Medicine in 1924. The introduction of portable ECG devices in the 1930s made this technology more accessible, and by the mid-20th century, ECGs had become an integral part of cardiovascular diagnostics.

As electrocardiography matured, advances in electrode placement led to the development of multi-lead configurations. Early systems expanded from single-lead to three and six-lead ECGs during the 1930s and 1940s, allowing clinicians to record the heart's electrical activity from multiple perspectives and detect abnormalities more accurately. These intermediate configurations paved the way for the standard twelve-lead ECG still used in clinical practice today.

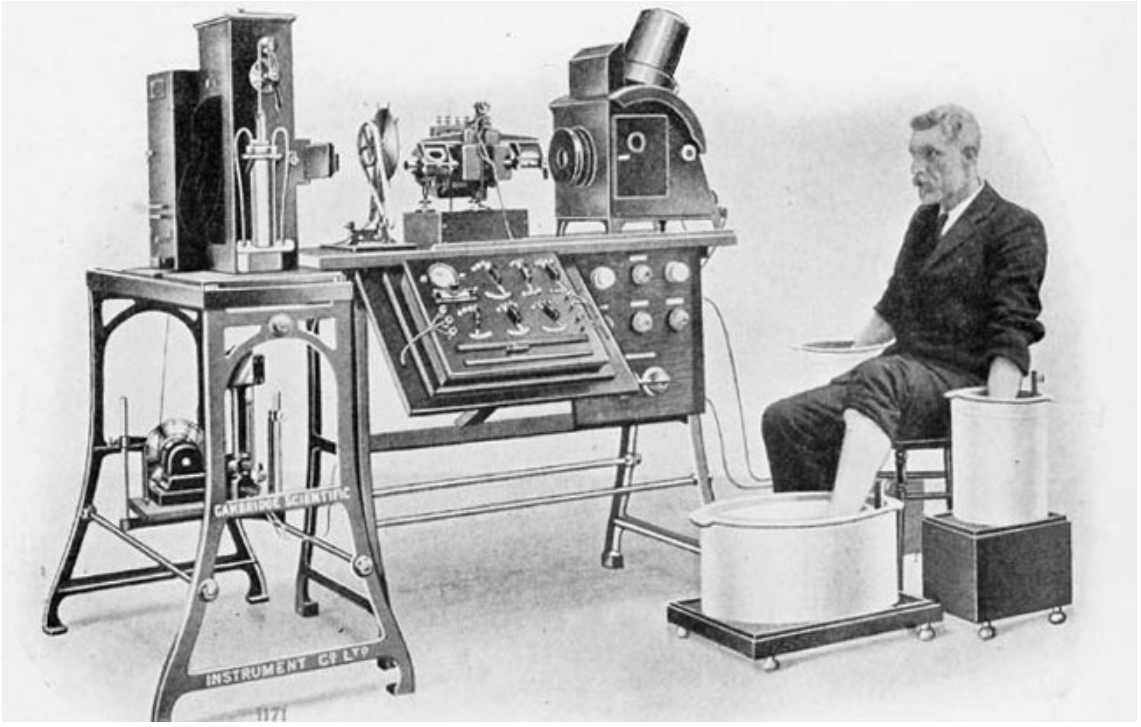


FIGURE 1.4: Illustration of early ECG recording using salt solution electrodes, a crucial breakthrough in non-invasive cardiac monitoring [Nanthakumar and Sivakumaran (2018)].

Norman Holter's 1949 invention of the Holter monitor (Figure 1.5) revolutionized ECG monitoring by enabling continuous 24–48-hour recordings, significantly improving the diagnosis of arrhythmia [DiMarco and Philbrick (1990)].

The invention of the pulse oximeter in the 1970s provided a revolutionary method for non-invasive measurement of arterial oxygen saturation (SpO_2) and pulse rate. Designed by Takuo Aoyagi [Severinghaus (2007)], the device relied on light absorption at different wavelengths to calculate SpO_2 based on the proportions of oxygenated and deoxygenated hemoglobin in the blood. Even though it was initially developed for surgery and critical care, it quickly became essential for monitoring oxygenation pulse. By the 1980s, portable pulse oximeters were widely adopted, making them a standard tool in operating rooms, intensive care units (ICUs) and emergency departments.

Eventually, technological advancements came together to create multi-parameter monitors, which could measure several vital signs at once in a single device. These tools integrated ECG, SpO_2, respiration rate and blood pressure monitoring, providing clinicians

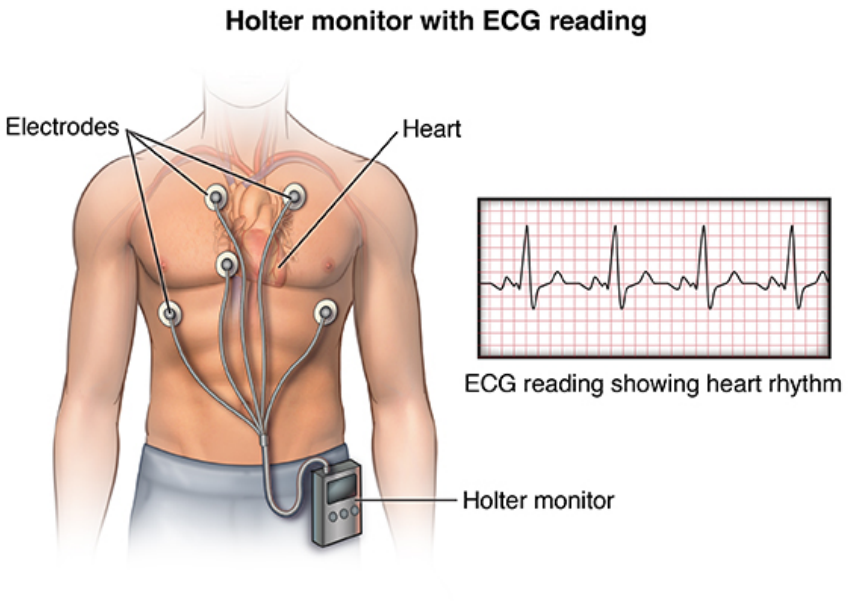


FIGURE 1.5: A Holter monitor and the output ECG reading illustrating heart rate. The number and position of electrodes varies by model, but most Holter monitors employ between three and eight. Image courtesy of the John Hopkins Medicine organization.

with a comprehensive real time overview of a patient's condition. Multi-parameter systems are the gold standard in modern medical settings where continuous monitoring is critical for managing critically or chronically ill patients.

1.1.4 Pulse Measurement in the Digital Age

In the 21st century, wearable technologies have brought pulse monitoring into everyday life. Devices like smartwatches [Phan et al. (2015); Reeder and David (2016); Sarhaddi et al. (2022)] and rings [Park et al. (2013); Cao et al. (2022); Kim et al. (2024)], use advanced optical sensors to continuously monitor HR, detect irregularities and even predict potential health issues. The Apple Watch, Fitbit and Garmin leverage photoplethysmography (PPG) to monitor pulse rate and rhythm. This technique measures blood volume changes in tissue using light absorption, enabling continuous tracking of HR during daily activities, exercise or sleep. Additionally, modern wearable devices incorporate features such as irregular rhythm detection, allowing users to monitor for signs of arrhythmia or atrial fibrillation. The combination of these physiological signals and the integration of Artificial Intelligence (AI) capabilities, has given the user the ability to track key health parameters, like activity, sleep, menstrual cycle, stress and predict energy levels and readiness scores. Some devices can even issue warnings for potentially dangerous conditions, such as tachycardia or bradycardia.

This shift from clinical tools to devices available to consumers underscores pulse's enduring importance as a vital sign, bridging the gap between medical and personal

health management. From the sphygmograph to AI-powered wearable devices, the history of pulse measurement reflects humanity's ongoing journey to understand the body's underlying rhythms.

1.2 I Can See Your Heartbeat: Video-Based Pulse Measurement

While these traditional methods advanced pulse monitoring in clinical settings, they all share a common limitation: they require direct contact with the patient. This, while effective in controlled environments, poses challenges in settings where physical contact or the use of traditional devices is impractical or unavailable, thus can lead to delays in diagnosis and sub-optimal management of health conditions. The rise of computer vision and machine learning has enabled remote photoplethysmography (rPPG), allowing pulse measurement from standard video recordings without physical sensors.

rPPG represents a groundbreaking shift in how vital signs can be monitored. Unlike conventional methods, rPPG utilizes standard cameras and advanced algorithms to detect subtle changes in skin color or micro motion caused by blood flow. These changes are captured by video and analyzed to extract pulse information. This innovation leverages accessible technology, eliminating the need for physical, often costly sensors while broadening the scope of which part of the body pulse measurements can be taken.

The increased risk of infection for healthcare workers and patients, the fragile skin of newborn infants or the elderly, the importance of continuous surveillance for chronic disease patients, individuals in inaccessible locations, lack of mobility, staff shortage or financial constraints are some of healthcare's current barriers that could benefit from remote monitoring solutions. Beyond medicine, applications can extend to wellbeing, with fitness, mental health and stress monitoring. Video-based pulse measurement aligns with the broader movement towards non-invasive, ubiquitous and user-friendly health monitoring solutions.

1.2.1 Uncovering the Science Behind rPPG

A critical step in remote HR estimation is the understanding of the underlying principles of pulse and its manifestation on the skin, particularly on the face. These principles have been well-established since the development of pulse oximetry in the 70s, which relies on light absorption by hemoglobin to measure blood oxygenation and pulse rate non-invasively. The following sections provide a detailed look at arterial pulse propagation, the role of hemoglobin in light absorption and reflection and the transition from traditional contact-based PPG systems to advanced video-based pulse estimation, which helps us understand how algorithms extract vital information from video data to enable real-time, non-invasive monitoring of pulse.

1.2.1.1 The Heartbeat's Journey

Pulse is defined as the rhythmic contraction and expansion of arteries caused by the ejection of blood by the heart's systolic phase (contraction phase, when blood is pumped out) and diastolic phase (relaxation phase, when the heart refills with blood) [Walker et al. (1990)]. The flow of blood in the arteries generates pressure waves that propagate through the arterial system, affecting surrounding tissues, including the skin. The two main reasons there are observable effects of pulse on the skin are the blood volume changes and the hemoglobin absorption and reflection.

Blood Volume Changes When the heart contracts during systole, it ejects blood into the aorta, creating a surge in blood pressure. This surge travels as a pressure wave away from the heart and through the arterial system, reaching the facial arteries. Since the vessels are located close to the surface, the skin experiences a subtle elevation during a cardiac cycle.

Hemoglobin Absorption and Reflection Hemoglobin is a protein found in red blood cells that plays a central role in the transport of oxygen from the lungs to tissues throughout the body. Hemoglobin also plays a role in transporting carbon dioxide, a waste product of metabolism, from the tissues back to the lungs for exhalation. The oxygenation state of hemoglobin can vary throughout the cardiac cycle. During systole, when blood is ejected into the arteries, there is higher concentration of oxyhemoglobin in the arterial blood. During diastole, when the heart is refilling with blood from the veins, the concentration of oxyhemoglobin in the arterial blood decreases (deoxyhemoglobin). These variations in hemoglobin oxygenation have a direct impact on how blood interacts with light, particularly in the visible and near-infrared spectrum.

1.2.1.2 Light, Blood and the Science of PPG

To quantify the pulse signal, researchers historically use PPG [Allen (2007)]. It is a non-invasive, contact-based technique that measures the light absorption or reflection caused by blood volume variations on the skin. A light source emits light in the visible or near-infrared spectrum, with specific wavelengths chosen to target the absorption characteristics of hemoglobin. It is placed on areas of the body where blood vessels are close to the surface. The emitted light from the source penetrates the skin and hits blood vessels beneath the skin's surface. Hemoglobin in the blood absorbs some of this light, while the remaining light is scattered and reflected back towards the photodetector. The PPG sensor detects these pulsatile changes in light absorption and converts them into an electrical signal. The frequency of the waveform corresponds to the HR, as each peak in the waveform represents a heartbeat. An illustration of the above procedure is presented in Figure 1.6.

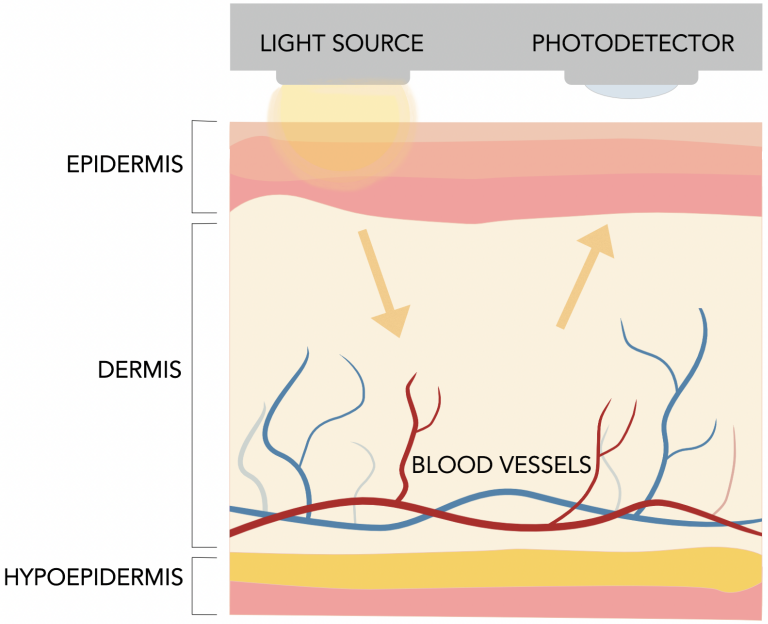


FIGURE 1.6: Illustration of pulse signal acquisition in PPG systems. The light source illuminates the skin, and variations in absorption due to blood flow are captured by the photodetector, forming the basis of pulse estimation.

Video-based pulse estimation harnesses these physiological responses to estimate pulse remotely and non-invasively. Since these subtle variations can manifest on the facial skin, by capturing them, video-based methods enable real-time monitoring of pulse without the need for physical contact or invasive sensors. RGB (Red, Green, Blue) cameras, which are commonly used in video-based pulse estimation due to their accessibility, capture images by measuring the intensity of light in three primary color channels: red, green and blue. The combination of these color channels allows RGB cameras to capture a wide range of colors and variations in skin appearance. By analyzing the changes in pixel values over time, algorithms can isolate the signal in the video, which corresponds to changes in blood volume and oxygenation due to the cardiac cycle.

1.3 Extracting Pulse from Pixels

The first study that demonstrated the feasibility of extracting pulse from video recordings of a person's face under ambient light was conducted in 2008 by Verkruysse et al. (2008). This groundbreaking work proved that consumer-grade cameras are capable of capturing subtle variations in light reflected from the skin, which correspond to blood volume changes produced by the cardiac cycle. Among the three primary color channels (red, green, blue), the green channel was identified as carrying the most robust pulse signal due to hemoglobin's peak light absorption in the green wavelength range. This finding was attributed to the fact that hemoglobin absorbs light more effectively

in the green wavelength range. The study also noted that while red and blue channels contain pulse signals, they are less distinct and more susceptible to noise.

Building on the findings of Verkruysse et al., Poh et al. (2010b,a) explored the application of webcam technology for pulse measurement in 2010. They introduced a systematic rPPG pipeline using Independent Component Analysis (ICA) to separate pulse signals from noise, including respiration and motion. They improved Region of Interest (ROI) tracking and extraction techniques to ensure reliable pulse estimation. While effective, the method struggled with significant motion artifacts and illumination variations. In their first study, [Poh et al. (2010b)] continuously selected the second independent component as the source signal for further analysis, assuming it represented the pulse signal. However, in their subsequent work Poh et al. (2010a), they refined their approach by choosing the independent component with the strongest frequency peak within the typical HR range. In 2011, Madej et al. (2011) introduced the use of Principal Component Analysis (PCA) as an alternative to ICA for separating the pulse signal. This method showed potential in scenarios with relatively low motion, but like ICA, it struggled in dynamic conditions with noise. CHROM, short for "chrominance-based remote photoplethysmography" by De Haan and Jeanne (2013), capitalized on the fact that the skin's chrominance (color) changes subtly during the cardiac cycle due to variations in blood volume and oxygenation. CHROM processes these color changes by linearly combining the red, green and blue channels, with adjustments to account for variations in skin tone reflectance.

While initial rPPG research focused mainly on analyzing changes in light reflection from the skin, a novel perspective emerged in 2013, introducing the concept of using facial micro-movements caused by blood flow as a pulse signal source. Balakrishnan et al. (2013) observed that the movement of blood through the facial arteries during the cardiac cycle causes subtle displacements of the skin. Blind Signal Separation (BSS), a statistical approach that separates a set of mixed signals into their independent sources, was used to filter out noise and extract the most periodic signal corresponding to the pulse. This approach demonstrated that pulse could be derived not only from light-based variations but also from the physical motion of the face.

These findings, regardless of wether the signal is extracted from light or motion changes, emphasized the importance of ROI selection for accurate rPPG, as not all facial regions contribute equally to the signal due to variations in blood vessel density. They highlighted that the distribution of blood vessels varies across facial regions, making some areas more suitable for pulse signal extraction. Kumar et al. (2015) evaluated different facial regions for pulse signal extraction and found the forehead and cheeks to be most reliable due to higher blood vessel density. Kwon et al. (2012) validated the effectiveness of smartphone cameras for pulse estimation. They systematically tested various facial regions, confirming the forehead and cheeks as optimal ROIs due to high Signal-to-Noise Ratio (SNR), whereas Lempe et al. (2013) favored the cheeks as they are rarely

occluded. Poh et al. (2010b) also briefly mentioned the importance of selecting stable regions of interest, such as the forehead, to improve signal quality. More recently, Kim et al. (2021), Wong et al. (2022) and Li et al. (2024) highlighted that the forehead and cheeks are conventionally regarded as preferred facial ROIs for rPPG measurements. Their research noted that regions with smaller angles of reflection, such as the forehead and cheeks, contained stronger rPPG signals and they are frequently used due to their favorable anatomical features and consistent signal quality.

However, static ROI selection can impact the robustness and accuracy of pulse estimation. A key issue is the assumption that specific facial regions consistently provide the optimal signal quality. In reality, the signal strength within these regions can vary significantly due to individual differences in skin tone, blood vessel distribution and physiological factors. Static ROIs are particularly susceptible to motion artifacts, as they do not account for shifts in the ROI caused by head movement. Uneven lighting or shadows can further degrade signal quality within a fixed ROI and occlusions like hair, glasses or facial covers, can obstruct the selected region, leading to signal loss. These challenges underscore the need for more adaptive or dynamic ROI selection methods, which can mitigate motion and illumination changes and individual variability.

To address this, Kiddle et al. (2023) introduced a tiling and aggregation algorithm that focuses on high-quality facial areas, particularly benefiting darker skin tones by dynamically identifying regions with stronger rPPG signals. They divided the face into small regions (tiles), evaluate their signal quality and combined the best-performing tiles to optimize rPPG signal extraction. Similarly, Feng et al. (2015) proposed a Dynamic ROI (DROI) method that uses K-means clustering to divide fixed ROIs into blocks, dynamically selecting the best-performing ones based on signal quality metrics like cross-correlation and signal-to-noise-ratio (SNR). This method showcased improved adaptability to physical and environmental variations. Building on these ideas, Po et al. (2018) developed an Adaptive ROI (AROI) approach that uses spatial-temporal blocks and mean-shift clustering (a method that groups data points by shifting them toward areas where similar points are most concentrated, automatically finding clusters without needing to specify how many there are) to create dynamic SNR maps, accounting for motion and illumination changes. Lastly, Wei et al. (2022) introduced a dynamic ROI tracking system that leverages facial landmarks and segmentation to optimize signal combination for robust pulse estimation, even with facial masks or varying video resolutions. Despite their contributions, these methods faced challenges in handling non-rigid facial motion, maintaining robustness under severe lighting variations and ensuring real-time performance. Furthermore, methods requiring predefined metrics or thresholds, like SNR or cross-correlation, can struggle with generalization across diverse datasets or applications. Despite these challenges, dynamic ROI selection remains an intriguing concept that aligns with our vision for improving rPPG and we incorporate similar principles into our work.

Even with the most advanced ROI selection algorithm, motion artifacts and illumination changes remain a significant challenge for rPPG. For the first, if the subject moves excessively, it becomes nearly impossible to extract a reliable pulse signal. There is a fundamental upper limit to the amount of motion that rPPG algorithms can tolerate. Much of the research in this field has focused on mitigating artifacts caused by common, smaller movements, such as talking, nodding or slight shifts in posture, to improve signal quality in more realistic scenarios. However, real-world settings, where individuals may move erratically or unpredictably, introduce a different level of complexity that remains an open challenge. For the latter, variations in lighting conditions can dramatically affect the intensity and visibility of skin color changes, which are crucial for rPPG signal extraction. Sudden shifts in brightness, shadows or different light sources can disrupt consistency, leading to signal degradation. While some methods attempt to normalize or compensate for these fluctuations, robust solutions that allow rPPG to function reliably across diverse lighting environments are needed.

To address motion and illumination artifacts, McDuff et al. (2014) demonstrated that using alternative color bands, such as cyan and orange, improves the quality of blood volume pulse (BVP) signals under varying lighting conditions. This method leverages ICA and filtering to mitigate noise but is limited by the dependency on predefined filtering parameters, which may not generalize across diverse scenarios. Similarly, Abdulrahaman (2024) introduced a two-stage motion artifact reduction algorithm, which partitions and recombines the green channel before applying wavelet denoising to extract HR. While effective, this approach depends heavily on the green channel, which may underperform in environments with uneven illumination or for individuals with darker skin tones. Xu et al. (2023) proposed a narrowband near-infrared (NIR) imaging system to address illumination changes, using facial landmarks to exclude segments with heavy motion noise. However, the reliance on specialized NIR equipment limits its applicability in general-purpose settings. To improve robustness to both motion and illumination, Li et al. (2014) used facial landmark tracking, adaptive filtering and segmentation to address rigid and non-rigid motion. Although it performed well under dynamic conditions, its computational complexity hindered real-time applications. Feng et al. (2014) leveraged a Lambertian model, where the surface was assumed to reflect light uniformly in all directions, and adaptive adjustments to color channels for motion compensation, but the method struggled with severe non-uniform illumination changes. By dynamically modifying the intensity of signals from the RGB channels in video recordings they aim to compensate for changes in lighting or motion. Since motion and illumination variations can distort the pulse signal captured in the RGB channels, adaptive adjustments aim to rebalance the signals by emphasizing the color channels most relevant to the pulse signal while minimizing the noise introduced by external factors. Lastly, Wang et al. (2014) introduced a pixel-based framework for motion compensation and spatial pruning, where each pixel in the video's ROIs is treated

as an independent sensor for extracting the rPPG signal. The method focuses on compensating for motion artifacts and removing noisy pixels (spatial pruning) to improve the quality of the extracted pulse signal. They significantly improved SNR but this method requires high-resolution video, which may not always be feasible. While all these methods demonstrate progress in addressing motion and illumination artifacts, challenges persist in generalizing across diverse datasets, ensuring real-time performance and maintaining robustness under extreme conditions. Nonetheless, we agree that every method is essential to address these artifacts one way or another, as they represent realistic conditions under which rPPG must perform accurately.

Signal processing techniques in rPPG often use frequency-domain analysis to extract pulse signals. For instance, Wang et al. (2014) applied Fast Fourier Transform (FFT) to identify the dominant frequency corresponding to the pulse rate, which works well for stationary signals with consistent periodicity. However, this approach assumes the signal is stationary and lacks the ability to track changes over time.

From our extensive review of the signal processing literature, we conclude that most approaches follow the structure presented in Figure 1.7. The framework begins with the pre-processing step, where the face is detected and specific ROIs, such as the forehead or cheeks, are identified. Feature extraction then extracts raw signals, such as pixel intensities or motion trajectories, from the ROIs. To isolate the frequency components associated with HR, filtering is applied - frequently using a bandpass filter tailored to the range of typical human HR. The filtered signals then undergo signal decomposition using methods like PCA or ICA, separating the pulse signal from noise caused by motion, lighting changes or respiration. Finally, frequency analysis identifies the dominant frequency for HR estimation.

While significant progress has been made in signal processing approaches for rPPG, several challenges remain. Static ROI selection assumes uniform signal quality across specific facial regions, ignoring individual variations in skin tone and physiological factors, which can lead to inconsistent performance. Motion artifacts, particularly during head movements or facial expressions, remain a persistent obstacle, with many methods failing to adapt to rapid motion despite advancements in motion compensation techniques. Illumination variability poses another major obstacle, with most approaches relying on assumptions of consistent lighting, making them less effective in real-world scenarios. The generalizability of these methods is limited by a lack of evaluation on diverse datasets, as many of these studies highlight, such as varying skin tones, ages and ethnicities. These challenges underscore the need for more robust, adaptive and inclusive approaches to improve the practicality and applicability of rPPG in real-world environments. Despite that, the significance of signal processing methods is highlighted to this day, despite the boom of Deep Learning (DL). These methods are straightforward, interpretable, computationally light and provide an excellent basis to understanding the limitations of rPPG in real-world settings.

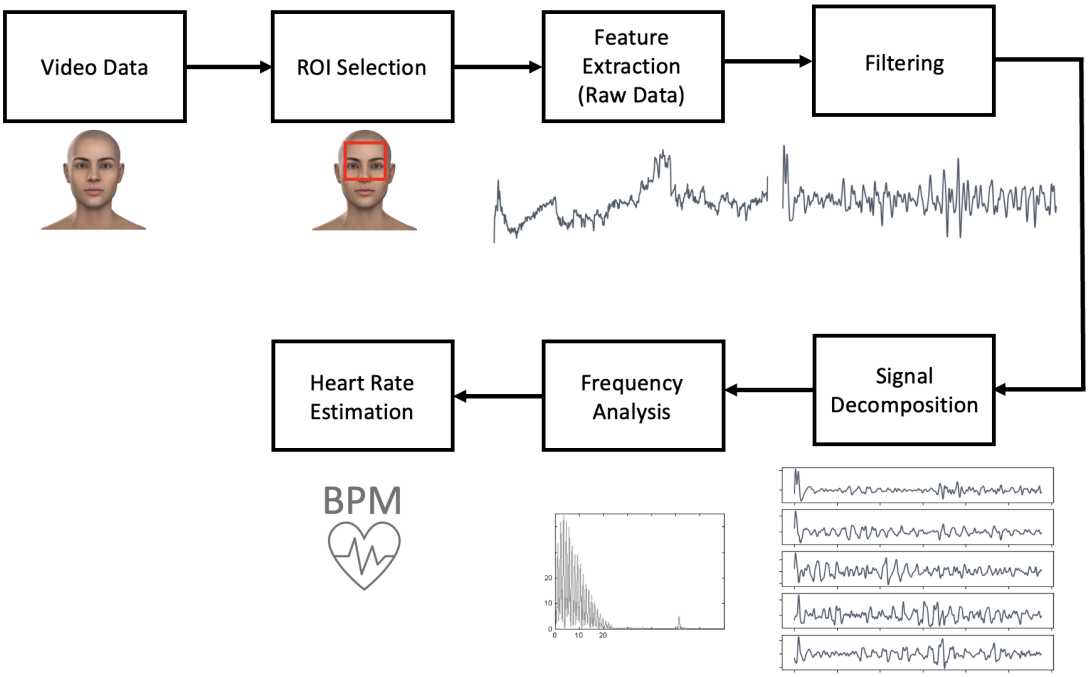


FIGURE 1.7: Flowchart of a typical signal processing algorithm for pulse estimation.

1.4 Teaching Machines to Read Pulse

As ML started becoming increasingly popular, researchers began to leverage its power to improve pulse estimation. ML methods offer several advantages over traditional signal processing approaches, including automatic feature extraction, scalability and adaptability. Unlike signal processing methods, which often rely on prior physiological knowledge and handcrafted features, ML models learn complex and subtle patterns in video data, enabling more robust and accurate pulse measurements even under challenging conditions.

The incorporation of ML in rPPG began with relatively simple algorithms aimed at augmenting existing signal processing frameworks. For example Monkaresi et al. (2013) extended Poh's signal decomposition approach by employing participant-specific K-nearest neighbors (KNN) and linear regression models. These models helped optimize the selection of independent components, improving pulse estimation accuracy in scenarios with realistic movement. Song et al. (2021) with PulseGAN, attempted to improve the quality of the extracted pulse waveforms using a Generative Adversarial Network (GAN)-based approach. A generator refines rough CHROM signals to closely match reference PPG signals, while a discriminator ensures realism.

While these methods are heavily reliant on signal processing pipelines, they marked the beginning of a shift towards end-to-end ML approaches. The field progressed rapidly as researchers began embracing DL for its ability to learn features directly from raw or minimally processed data, eliminating the need for handcrafted feature extraction.

Convolutional Neural Networks (CNNs) were among the first DL models explored for rPPG, leveraging their ability to detect spatial patterns in video frames for pulse signal extraction. Additionally, multi-modal models have emerged, enabling the simultaneous estimation of multiple vital signs, such as HR and respiratory rate, further expanding the capabilities and applications of rPPG systems.

Špetlík et al. introduced a two-step CNN architecture comprising of an Extractor component, which processes facial image sequences to extract signals optimized for SNR, and an HR Estimator, which predicts the HR from the extracted signal. DeepPhys [Chen and McDuff (2018)] utilized a deep attention CNN to calculate pulse and respiration rates. It combines motion and appearance-based attention mechanisms to identify and focus on ROIs, which enables robust measurement under challenging conditions. Its attention mechanism allows the spatio-temporal visualization of physiological signals, highlighting regions like the forehead and carotid arteries for pulse and nasal flaring for respiration. In 2021, Liu et al. (2020), continuing the DeepPhys work, developed a multitask temporal-shift convolutional attention network (MTTS-CAN) for real-time heart and respiration rate estimation. The temporal shift modules were used to remove noise, the attention mechanism to improve signal source separation and the multitask mechanism to estimate pulse and respiration rates jointly. It presented a two-branch structure, one for motion modeling and the other for extracting facial features. TS-CAN is another version of MTTS-CAN, however it can only assess pulse and respiration separately, not simultaneously.

In literature, three dimensional (3D) CNNs are often described as spatiotemporal networks due to their ability to model both spatial and temporal information in videos. However, this differs from methods using spatiotemporal maps, which explicitly encode physiological signals into structured two dimensional (2D) representations. For clarity we will thereafter refer to 3D CNNs as 3D approaches.

One such approach is PhysNet [Yu et al. (2019a)], designed to accurately estimate pulse for applications such as HR variability (HRV) analysis, atrial fibrillation and emotion recognition. PhysNet combined 3D CNNs and Recurrent Neural Networks (RNNs) to model both spatial and temporal features effectively (Figure 1.8). The network achieved accurate pulse peak detection using a negative Pearson correlation loss. Yu et al. (2019b) proposed a system with two components: Spatio-Temporal Video Enhancement Network (STVEN), which improved the quality of compressed videos and rPPGNet, a spatio-temporal network designed for accurate rPPG signal recovery. rPPGNet integrated attention mechanisms and partition constraints to enhance signal accuracy at both HR and HRV levels.

Some particularly interesting approaches that focus on developing lightweight models to address the computational challenges of rPPG deployment on resource-constrained devices are MobilePhys and EfficientPhys. MobilePhys [Liu et al. (2022)] leverages both

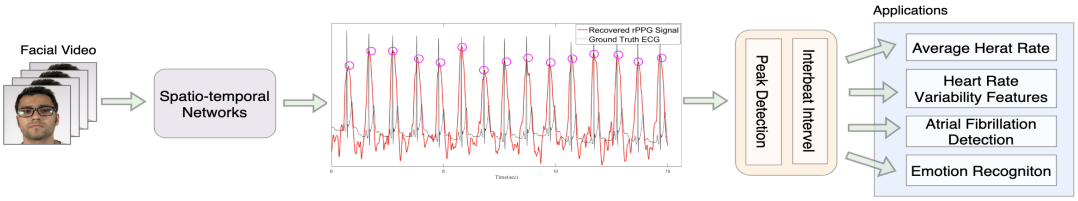


FIGURE 1.8: Framework of 3D network PhysNet [Yu et al. (2019a)].

front and rear cameras on smartphones to generate high-quality self-supervised labels, reducing reliance on large labeled datasets while maintaining robust performance. This approach enables real-time physiological monitoring without excessive computational overhead. Similarly, EfficientPhys [Liu et al. (2021a)] minimizes processing requirements by eliminating the need for face detection, segmentation, normalization, color space transformation, or any other preprocessing steps. Instead, it directly processes raw video frames using a lightweight convolutional architecture that implicitly learns spatial and temporal features relevant to physiological signal extraction. It introduces the first end-to-end on-device architecture specifically designed for mobile applications, balancing efficiency and accuracy while enabling real-time inference without traditional preprocessing pipelines.

Transformers, originally developed for natural language processing (NLP) tasks [Vaswani (2017)], have gained significant traction in rPPG and have shown promise by leveraging their powerful self-attention mechanism. Gupta et al. (2023) developed a novel signal embedding mechanism that captures rPPG-specific features and feeds them into a transformer model for temporal pattern extraction. The transformer architecture enables long-range dependency modeling across video frames, which enhances the detection of subtle blood volume changes. PhysFormer [Yu et al. (2022)] is a Temporal Difference Transformer (TDT) which enhances rPPG signal recovery by explicitly capturing subtle temporal differences in facial videos. It leverages a temporal difference attention mechanism, which emphasizes changes in blood volume over time rather than static facial features. PhysFormer++ [Yu et al. (2023)] builds upon the PhysFormer framework by introducing a SlowFast temporal difference transformer with two distinct pathways. Unlike PhysFormer, which utilizes only the slow pathway, PhysFormer++ combines both slow and fast pathways to better capture temporal context and periodic rPPG patterns from facial videos, thereby enhancing its ability to extract meaningful physiological signals. The slow pathway processes information at a lower temporal resolution, focusing on long-term trends and stable pulse variations, while the fast pathway operates at a higher temporal resolution, capturing rapid, subtle fluctuations in blood volume changes. TransPPG [Kang et al. (2024)] is a two-stream transformer that processes both foreground (facial) and background information, enabling feature-level subtraction to reduce noise caused by environmental factors like illumination and motion.

Impressive results have been reported in all these works, however, a major challenge

remains: the lack of publicly available, sufficiently large and diverse datasets essential for training DL models in rPPG. To mitigate this limitation, researchers have turned to synthetic data. For example, Niu et al. (2018) propose SynRhythm, a transfer learning strategy that pre-trains a deep HR estimator using synthetic heartbeat signals before fine-tuning on real video datasets. Their approach involves generating synthetic spatiotemporal maps of facial regions and synthetic rhythm signals to train a deep regression model, which is initially pre-trained on ImageNet. This pre-trained model is then adapted to real-world HR estimation, demonstrating improved performance on limited video data. Similarly, McDuff et al. (2021) introduce a method for generating synthetic facial videos with underlying blood flow patterns and breathing movements. Their synthetic avatars exhibit diverse characteristics such as varying skin tones, facial hair, facial expressions (e.g., smiling, blinking, mouth movements) and head motion. By testing these synthetic datasets on the MTTS-CAN model, they observed that rPPG performance significantly declines for individuals with darker skin tones when trained on conventional datasets. To address this, they trained separate models for dark and light skin tones and found that the dark-skin model performed better across all skin types, highlighting the importance of skin-tone-aware training in rPPG applications.

Unsupervised methods play a critical role in addressing the challenges caused by the lack of labeled datasets in rPPG. For example, Sun and Li (2022) avoided using ground truth signals during training by using a 3D CNN to extract spatiotemporal blocks from facial videos. It then applies contrastive learning to capture similarities within a video while ensuring differences across videos. Similarly, pseudo-labeling bridges the gap between supervised and unsupervised learning by generating approximate labels from the data itself, allowing for further fine-tuning of models [Li and Yin (2023), Savic and Zhao (2024), Liu et al. (2021b)]. These approaches not only mitigate the dependency on large annotated datasets but also pave the way for more scalable and generalizable solutions in rPPG research.

Despite their impressive advancements, ML and DL approaches face notable challenges. Most importantly, they rely on large labeled datasets, which remains a bottleneck. While synthetic data generation has alleviated some of this burden, the generated datasets may not fully capture the complexity of real-world scenarios. Additionally, many models demonstrate reduced performance under extreme motion or lighting conditions, as the training data often lacks sufficient diversity to generalize across such scenarios. Finally, the computational requirements of DL models can be a bottleneck for real-time applications on resource-constrained devices.

1.5 Mapping Pulse Through Space and Time

Hybrid or spatiotemporal approaches, as we will be referring to them in this thesis, are a combination of signal processing and ML methods, in the sense that they extract handcrafted features followed by a DL network for pulse estimation. These methods take advantage of spatiotemporal maps, which are subsequently fed through a machine or DL framework to extract the pulse rate. Spatiotemporal maps combine spatial information from ROIs with temporal information, providing a condensed yet rich representation of physiological signals over time.

For instance, Wu et al. (2012) employed the Eulerian Video Magnification approach to amplify subtle spatiotemporal features from facial videos, enhancing signal quality before further analysis. Similarly, Niu et al. (2018, 2019b,a) generated spatiotemporal maps by aggregating signals from multiple ROIs, improving the SNR for pulse estimation. These spatiotemporal maps effectively encode physiological cues by mapping pixel intensity variations across time and space into a more interpretable 2D representation. Jaiswal and Meenpal (2022) compressed spatial redundancy by generating a compact 2D spatiotemporal map of ROIs, preserving temporal dynamics while reducing computational load. Shao et al. (2023) took this a step further by designing a spatiotemporal transformer module, leveraging self-attention to focus on important regions and aggregate physiological cues from facial areas, further improving robustness and accuracy. Song et al. (2020) constructed spatiotemporal images by organizing pulse signals extracted from conventional rPPG methods into a structured 2D format which captures both the spatial and temporal characteristics of the signal. A modified ResNet-18 was trained to map these feature images to HR values.

The use of spatiotemporal features offers numerous advantages over traditional video-based approaches. These methods achieve higher temporal resolution compared to raw video frames, providing detailed insights into skin changes caused by blood volume fluctuations. They also mitigate the impact of motion and light artifacts by integrating information over time, leading to more robust pulse estimation. The integration of handcrafted preprocessing steps with DL enables models to exploit prior knowledge while benefiting from the automatic feature extraction capabilities of neural networks.

However, there remain challenges and unexplored areas in spatiotemporal approaches. Most existing works rely on pre-defined ROIs, which may overlook regions that carry sufficient physiological signals, limiting the accuracy of pulse estimation. Furthermore, averaging information from multiple frames into a single spatiotemporal image can suppress significant signal variations, potentially discarding important physiological dynamics. These limitations highlight the need for more adaptive and dynamic strategies to identify and utilize the most informative regions in videos.

1.6 Are We Certain?

In medical and ML applications, certainty in predictions is crucial, especially when dealing with noisy, real-world data. Conformal predictions offer a principled way to quantify uncertainty, ensuring that pulse estimations are not only accurate but also reliable.

conformal preditions is a statistical framework that provides reliable uncertainty estimates for ML predictions. By generating prediction sets or intervals with a guaranteed level of confidence, conformal preditions ensures coverage for new, unseen data under minimal distributional assumptions. It has been applied across various domains, including healthcare, time-series analysis and computer vision, which are relevant to rPPG applications.

In time-series forecasting, Stankeviciute et al. (2021) applied conformal preditions to multivariate time-series for robust uncertainty estimation. The temporal nature of rPPG signals aligns with such work, suggesting that conformal preditions could enhance trust in pulse predictions by quantifying uncertainty in HR estimations.

conformal preditions methods have been used to provide confidence intervals in clinical settings, particularly for medical imaging tasks. Papadopoulos et al. (2017) presented a method for providing reliable confidence measures in stroke risk estimation using ultrasound carotid plaque imaging. Gade et al. (2024) applied conformal preditions to a DL-based prostate segmentation model, flagging uncertain pixel predictions at a user-defined confidence level. Lu et al. (2022) explored how conformal preditions can complement existing DL approaches by providing intuitive uncertainty measures and facilitating greater transparency in clinical use. They modified conformal preditions methods to be more adaptive to subgroup differences in patient skin tones through equalized coverage.

To the best of our knowledge, at the time of writing this thesis, no prior studies have applied conformal preditions to rPPG. Given the challenges in rPPG, such as sensitivity to motion, illumination changes as we discover in previous work, conformal preditions could offer a valuable framework for improving reliability and quantifying certainty in pulse estimation. In medical applications, particularly in remote healthcare monitoring, ensuring trustworthy predictions is essential for clinical adoption. By leveraging conformal preditions, future rPPG research could achieve more reliable and interpretable physiological measurements, paving the way for broader adoption in real-world healthcare and wellness applications.

1.7 Video Quality Matters

Several studies have explored how motion, illumination, compression, and resolution affect video quality for rPPG. However, no comprehensive framework currently exists to evaluate these factors holistically. Existing works often focus on one or two specific factors, such as the impact of motion artifacts as we saw in multiple works previously, without accounting for the full range of video quality dimensions critical for robust pulse estimation. Motion artifacts are quantified in some datasets like MAHNOB-HCI Soleymani et al. (2011), but these evaluations lack integration with other quality measures, such as occlusions or blur. Compression studies, such as McDuff et al. (2017), primarily analyze bitrates and codecs but do not address interactions with environmental factors like illumination or motion. Hanfland and Paul (2016) also investigated the impact of video compression formats - Motion JPEG, MPEG-4 and Motion JPEG 2000 - on the quality of rPPG signals, comparing correlation indices between raw and compressed.

The lack of a unified metric limits our ability to systematically evaluate video quality for rPPG across diverse settings. Factors such as occlusions (e.g., glasses, masks), blur (e.g., out-of-focus regions), illumination (e.g., uneven lighting or flicker) and motion (e.g., head movement) are known to degrade rPPG signals, but they are rarely quantified together in a single framework.

To date and to the best of our knowledge, no study has proposed a metric or evaluation framework that captures the combined effects of all these factors on rPPG signal quality. A comprehensive video quality metric for rPPG, incorporating motion robustness, illumination consistency, focus clarity and occlusion detection, would bridge this gap and provide a standardized tool to assess the suitability of videos for pulse extraction. This thesis addresses this limitation by presenting a novel set of metrics designed to evaluate video quality for rPPG holistically, enabling better understanding and selection of data for robust pulse estimation.

1.8 The (Challenging) Road Ahead

The evolution of rPPG has been marked by significant advances in signal processing, ML and spatiotemporal methods, each addressing different aspects of the challenges in remote pulse estimation. Traditional signal processing techniques laid the foundation for rPPG by identifying the physiological principles behind pulse extraction from facial videos. These methods, however, are constrained by their sensitivity to motion artifacts, illumination variability and static ROI selection, which prompted the transition to ML-based approaches.

DL has significantly improved rPPG performance by learning robust feature representations from raw video data. CNNs and transformers have demonstrated superior accuracy by leveraging spatial and temporal dependencies in 3D, while spatiotemporal methods balance the strengths of both signal processing and DL-based rPPG pipelines. Despite these advances, key challenges remain, particularly in dataset limitations, generalization across diverse populations and computational efficiency. While synthetic data and self-supervised learning have been explored to mitigate data scarcity, there is still a gap in establishing standardized evaluation metrics that account for real-world variations in video quality.

One of the most pressing concerns in rPPG is its reliability under diverse conditions. Factors such as lighting variability, skin tone and motion artifacts continue to pose significant challenges. To ensure equitable performance, models must be rigorously tested across diverse populations, avoiding biases that could disproportionately affect individuals with darker skin tones or those in non-ideal lighting environments. That being said, this level of rigor requires access to large, diverse and well-annotated datasets, which are currently scarce. Most publicly available datasets are limited in size and demographic representation, often lacking sufficient variations in skin tone, age and environmental conditions. More inclusive data collection efforts and availability for researchers are paramount for the future of rPPG research.

Despite the remaining challenges, rPPG has expanded beyond medical applications in fitness tracking, human-computer interaction and biometric security. However, for rPPG to become a truly reliable and accessible tool, the field must continue addressing biases. The research in this thesis aims to address these gaps by improving the robustness, interpretability and generalizability of rPPG. By tackling these challenges, we move closer to establishing rPPG as a reliable tool for healthcare and beyond.

Chapter 2

Input Layer: Foundations of Signal Processing for Remote Photoplethysmography

Signal processing marks the beginning of the remote photoplethysmography journey. Like the input layer of a machine learning model, it serves as the gateway, transforming raw, noisy data into structured insights and laying the groundwork for everything that follows.

Despite the rise of deep learning in biomedical signal analysis, signal processing remains the backbone of rPPG research due to its interpretability, efficiency and physiological focus. Unlike deep learning models, often criticized as 'black boxes,' signal processing techniques provide transparency, which allows researchers to better handle challenges like noise, motion artifacts and illumination variability. Another key advantage of signal processing is its computational efficiency, requiring fewer resources and training data compared to deep models, making them particularly attractive for initial analysis and scenarios where simplicity and speed are crucial.

Historically, the first approach used to extract vital signs from video data was a signal processing one. In the early 2000s, when deep learning was still in its infancy and computational resources limited, signal processing provided a robust and efficient framework for this emerging field. The revolutionary study by Verkruysse et al. (2008) demonstrated that ambient light captured using a standard consumer-grade camera

could reveal pulse information. This study established the foundation of rPPG, proving that subtle changes in reflected light on the skin could be analyzed to extract physiological signals. While early research focused on analyzing reflected light, following studies such as Balakrishnan et al. (2013), explored the Newtonian reaction of micro motion caused by blood flow, a work that broadened the scope of rPPG research.

Our decision to focus on signal processing techniques in this chapter is driven by both practical considerations and the desire to investigate their strengths and limitations. While deep learning methods have achieved impressive results in rPPG, the underlying factors influencing their performance are often less interpretable. In contrast, signal processing enables a straightforward analysis of the core aspects of the problem, such as the structure of the rPPG signal and how it is influenced by light and motion.

Signal processing methods also operate on well-established principles, with tools such as ICA, Fast Fourier Transform (FFT) and bandpass filtering. These techniques were key in the early days of rPPG research, providing practical ways to separate meaningful signals from noise. By building a strong foundation in signal processing, following research can benefit from a deeper understanding of the problem, ensuring more robust, effective and tailored solutions.

Chapter Contributions:
In this chapter we investigate the role of signal processing in rPPG. This work is our foundation for understanding the core principles of rPPG signal extraction, on which the later approaches and evaluation methods will rely. More specifically, our contributions are as follows:

- We implement two core signal processing methods for rPPG-based on facial displacements. These methods establish a baseline for pulse estimation, but we find their performance is heavily influenced by dataset characteristics, particularly motion artifacts.
- We study the use of Particle Video Point Trajectories for feature tracking compared to traditional optical flow algorithms. Our findings show that this improves tracking accuracy, resulting in lower errors compared to traditional feature tracking methods.
- We evaluate the impact of feature clustering using K-Means clustering to improve computational efficiency and performance. Our experiments demonstrate that clustering features reduces the error in prediction.
- We analyze the effect of dataset bias on signal processing methods and find that limited representation of women and darker skin tones leads to poorer model generalizability. This finding highlights the importance of diverse datasets to improve inclusion and performance.

- We investigate the limitations of handcrafted features in signal processing. Our results indicate that these features, while interpretable, require extensive tuning, struggle to generalize across datasets and are less effective in handling significant motion and illumination variability.

The rest of the chapter is organized as follows: Section 2.1 introduces the datasets and experimental setup, detailing participant demographics, data collection protocols and evaluation metrics. Section 2.2 describes the signal processing frameworks, along with advancements like Persistent Independent Particles for feature tracking. Section 2.3 presents our optimization insights, covering parameter tuning and its impact on performance. It discusses the results, comparing the methods across datasets and conditions, while Section 2.4 concludes with a discussion of the findings, limitations and implications for future work.

2.1 Datasets and Experimental Foundations

This chapter utilizes two datasets: one obtained from Ostankovich et al. (2018), referred to as O-HR, and a subset of the MMSE (Multimodal Spontaneous Expression) dataset [Zhang et al. (2016)], referred to as MMSE-HR. These datasets were selected due to their accessibility and their suitability for evaluating rPPG methods under varied conditions. The O-HR dataset was immediately accessible, providing a good starting point for the experiments. In contrast, acquiring the MMSE-HR dataset required nearly a year of effort, highlighting the broader issue of data accessibility in rPPG research. Many other datasets either did not respond to access requests or presented financial barriers, highlighting the challenges of data accessibility in rPPG research. Despite these challenges, O-HR and MMSE-HR together offer insights into rPPG performance across controlled, diverse and dynamic scenarios.

We conduct all experiments on a MacBook Pro equipped with the Apple M1 Pro chip, 16GB of memory and a 500GB SSD.

2.1.1 O-HR Dataset

The O-HR dataset [Ostankovich et al. (2018)] consists of 30 RGB videos of 15 participants (Figure 2.2) filmed in a seated position before (normal) and after physical exercise (physical). Each video has an average duration of 20 seconds, a frame rate of 25 frames per second (fps) and a resolution of 1920 by 1080 pixels. Videos are captured indoors under fluorescent lighting. Participant demographics include:

- **Gender:** 86.6% male

- **Facial Hair:** 60% of participants have facial hair
- **Skin Tone:** Based on the Fitzpatrick (1988) scale, participants fall within skin types II-IV, with 86.6% in types II-III and one participant in type IV as seen in Figure 2.1.

We conduct this classification manually as it provides a reasonable indication of the dataset's diversity. While the Fitzpatrick skin type scale was used to approximate participant skin tone, it is important to recognise that this classification system has significant limitations. Originally developed for assessing skin reactivity to ultraviolet light rather than for colourimetric or imaging purposes, the Fitzpatrick scale simplifies a complex, continuous spectrum of skin reflectance into six broad categories. This reduction overlooks crucial variations in undertone, melanin distribution and illumination-dependent appearance that directly affect image-based physiological sensing such as rPPG. Furthermore, the scale was derived from predominantly lighter-skinned populations and lacks balanced representation across global skin tones, leading to potential bias when used for evaluating model generalisability.

Manual classification compounds these limitations, as it is subjective and dependent on viewing conditions, monitor calibration and individual perception. As a result, the use of Fitzpatrick types in this context should be interpreted as a coarse approximation rather than an accurate ground truth. Future work should adopt more objective, imaging-based skin tone estimation methods (e.g., colour-space mapping or standardised digital scales to better capture the diversity relevant to rPPG performance.

FIGURE 2.1: The Fitzpatrick skin scale [Charlton et al. (2020)].

The ground truth heart rate is collected using a 6-lead ECG device (I, II, III, avR, avL and avF) sampled at 250 Hz. For this study, we use lead I as it provides reliable measurements for heart rate estimation, after consultation with a medical professional. Ground truth data is available in two formats:

- .cardio files, requiring specialized software ("ECG Control") for extraction
- .txt files

FIGURE 2.2: Example frames from the O-HR dataset, showcasing diverse participants in seated conditions.

The original heart rate recordings were stored in the `.cardio` format. This format is not directly compatible with standard data analysis tools and offers limited documentation, making automated parsing and integration with Python-based workflows more challenging. For processing ease and reproducibility, we opt for the `.txt` format as it is widely supported. Using SciPy's peak detection algorithm we identify the R-peaks in the ECG signal, representing each heartbeat. This involves analyzing the signal to locate local maxima within the QRS complex, where the R-peaks correspond to the highest amplitude points, making them ideal markers for heart rate calculation, as illustrated in Figure 2.3.

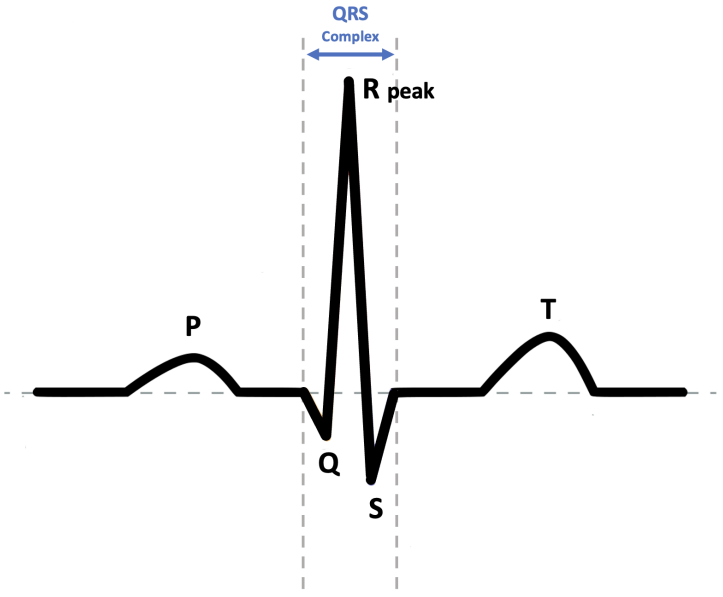


FIGURE 2.3: Illustration of an ECG waveform; the R-peak represents the highest point in the QRS complex, typically used for heart rate calculation through peak detection algorithms.

Respiration, voluntary or involuntary movement or electrode contact issues may introduce low-frequency artifacts. Figure 2.4 illustrates the raw and filtered ECG signals for visual comparison, amplitude representing the strength of the electrical signal. We filter the raw signal with a high-pass Butterworth filter (1 Hz) to remove baseline drift and isolate the cardiac signal. Filtering is applied only for visualization purposes to enhance the clarity of the ground truth ECG signal. Since the R-peaks in the raw signal are already prominent, filtering is not necessary for peak detection and heart rate extraction later in the chapter. More detailed descriptions on filtering and its impact can be found in Section 2.2.1.1.

When cropping videos into shorter segments for analysis, we adjust the ground truth by selecting the corresponding number of samples from the ECG data, aligning them with video duration.

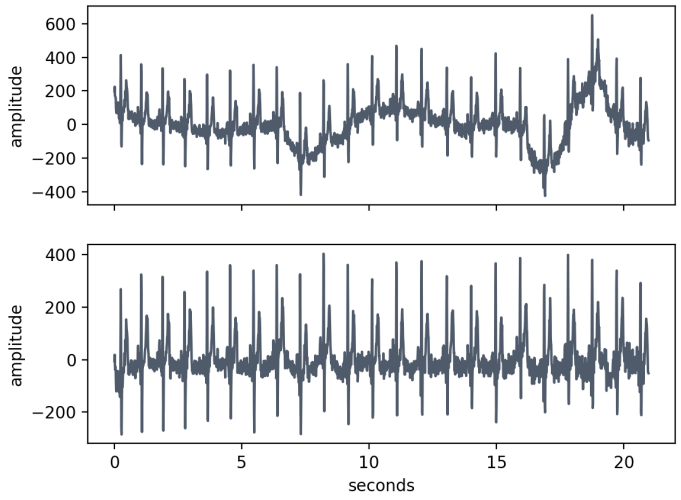


FIGURE 2.4: Comparison of the raw and filtered ECG signals, illustrating the removal of low-frequency artifacts.

2.1.2 Multimodal Spontaneous Expression - HR Dataset

MMSE-HR is a subset of the MMSE database [Zhang et al. (2016)], specifically designed to test the robustness of heart rate estimation methods under conditions with diverse emotional and physical responses. It consists of 98 RGB videos from 40 participants, recorded at a resolution of 1040 by 1392 pixels with a frame rate of 25 fps. Each video duration ranges from 30 seconds to 1 minute. During the recordings, participants are exposed to various stimuli, including videos, sounds and smells, to evoke a variety of emotions.

Similar to the O-HR dataset, we perform a manual participant classification by skin tone using the Fitzpatrick (1988) scale. This is based on percentage split information

provided with the dataset and visual inspection. Table 2.1 summarizes the distribution of participants by skin tone and gender. It must be noted that the Fitzpatrick skin tone types I–III were merged into a single category for analysis. They were visually difficult to distinguish reliably based on the available images. The original Fitzpatrick classification was developed for assessing skin reactivity to ultraviolet light rather than for colour-based categorisation and its distinctions between lighter tones are subtle and often indiscernible in standard RGB video, particularly under varied illumination and camera settings. Given these limitations, separating types I, II and III would have introduced unnecessary subjectivity and inconsistency in manual labeling. Grouping them into a single "lighter tone" category therefore ensured greater reliability and reproducibility in the classification process while still allowing meaningful comparison with darker tones (IV–VI), where differences in melanin concentration have a clearer effect on signal quality.

TABLE 2.1: Participant distribution by skin tone and gender for the MMSE-HR dataset, based on the Fitzpatrick (1988) scale.

Skin tone/ Gender	I-III	IV	V	VI	Total per Gender
Male	14	2	1	-	17
Female	21	1	-	1	23
Total per Skin tone	35	3	1	1	**40**

The dataset includes six emotion-evoking tasks, transitioning participants between positive and negative emotional states. Each task is followed by a brief pause. The distribution of participants across these tasks is summarized in Table 2.2. A detailed description of the activities is provided in Table 2.3.

TABLE 2.2: Participant distribution by activity type for the MMSE-HR dataset.

Skin tone/ Gender	T1	T8	T9	T10	T11	T14	Total per Gender
Male	2	2	-	17	16	-	37
Female	8	7	1	23	21	1	61
Total	10	9	1	40	37	1	**98**

Heart rate ground truth is recorded via a 1 kHz contact sensor, providing pulse measurements that align with the video frames. The data is provided in a .txt format, containing heart rate measurements aligned with each frame of the video. For experiments involving shorter video segments, we adjust the ground truth by calculating the number of frames corresponding to the desired duration and averaging the pulse measurements over those frames. This ensures alignment between video data and heart rate measurements.

TABLE 2.3: Description of activities in the MMSE-HR dataset, summarizing the stimuli used.

Task	Activity
T1	Listen to a funny joke
T8	Improvise a silly song
T9	Follow-up task similar to "Improvise a silly song"
T10	Experience physical threat in dart game
T11	Cold pressor: Submerge hand into ice water
T14	Experience smelly odor

2.1.3 Evaluation Metrics

We evaluate our methods using Mean Absolute Error (MAE), defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2.1}$$

where y_i and $\hat{y}_i$ denote the ground truth labels and predictions, respectively.

We also compute the Standard Deviation (SD) of the errors, defined as:

$$\text{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

where x_i represents the individual errors and $\bar{x}$ is the mean error.

MAE and SD are particularly relevant for evaluating rPPG methods because they quantify the accuracy and consistency of our predictions. MAE measures the deviation between the predicted and ground truth heart rates, quantifying prediction accuracy. Meanwhile, SD reflects the variability in the errors, which is crucial for assessing the robustness of the method across different conditions and participants. Together, these metrics offer a strong evaluation of our methods' performance, highlighting both accuracy and reliability.

2.2 Tracking the Pulse with Motion-Based rPPG Models

Both motion-based methods we build upon rely on the idea that subtle motion changes in facial regions caused by blood flow can be used to estimate pulse. While the principles are similar, the two approaches differ in their signal decomposition strategies.

Below, we describe the two methods in detail, highlighting the adaptations made for our experiments.

We build on the frameworks proposed by Balakrishnan et al. (2013) and Ostankovich et al. (2018), adapting their methods to align with our dataset characteristics. While our general workflow remains consistent with their proposed approaches, specific implementation details such as parameters, ROI selection, feature tracking and filtering techniques have been adjusted, either because details were not provided by the authors, to address dataset specific challenges or to improve performance. These frameworks were chosen because they represent key advancements in motion-based pulse estimation, illustrating the progress made in reducing noise and improving accuracy. For instance, Balakrishnan et al. (2013) introduced the first framework to detect subtle motion changes caused by blood flow, while Ostankovich et al. (2018) further refined this by enhancing feature selection and integrating more advanced signal processing techniques.

Together, these methods provide a robust baseline while also uncovering the strengths and limitations of traditional signal processing approaches for pulse estimation.

2.2.1 Motion-Pulse rPPG (MP-rPPG)

The first method we implement follows the framework proposed by Balakrishnan et al. (2013), referred to as Motion-Pulse rPPG (MP-rPPG). The pipeline begins with applying the Viola and Jones (2001) algorithm to the gray-scale version of the frame to identify a bounding box containing the face.

To ensure only relevant facial regions are included and to minimize motion artifacts, we reduce the size of the bounding box by 50% of its width, 90% of its height and exclude areas such as the neck, hair and eyes (20%-50% of the bounding box height), as seen in Figure 2.5.

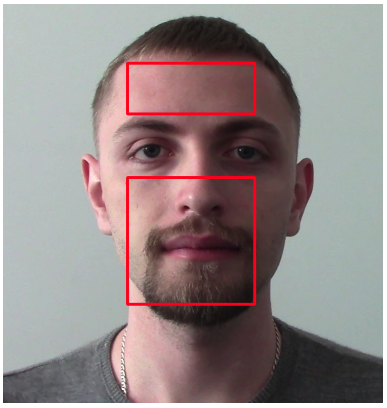


FIGURE 2.5: Selected ROIs focusing on the forehead and mouth/cheeks, optimized for pulse estimation.

As illustrated in Figure 2.6, we test a variety of ROIs, from individual areas such as the forehead and mouth/cheek regions to their combinations, as prior research suggests that combining these regions provides optimal results [Lempe et al. (2013), Tasli et al. (2014)]. In addition to testing the above standard ROIs, we explore the neck area and a region including only the cheeks and nose. This allows us to evaluate the impact of ROIs on the quality and reliability of the extracted signals.

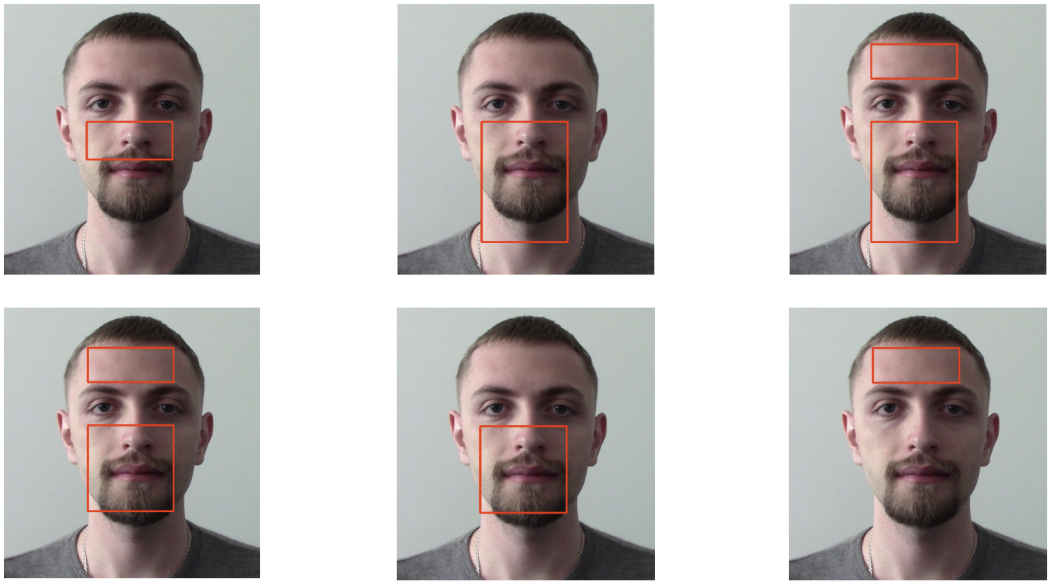


FIGURE 2.6: Examples of different ROIs tested during experiments, including combinations of the forehead, mouth/cheeks area and neck.

Within the selected ROI, we detect features using the Shi-Tomasi corner detection algorithm [Tomasi and Kanade (1991)]. This algorithm identifies points of high texture variation, which are ideal for tracking motion. We track the detected features across video frames using the Lucas-Kanade optical flow algorithm [Shi et al. (1994)] to estimate vertical displacements, as these are most correlated with cardiovascular activity according to Balakrishnan et al. This algorithm estimates the motion of features by analyzing their displacement between consecutive frames. We specifically focus on the vertical displacement, as it "captures the majority of motion caused by cardiovascular activity" according to Balakrishnan et al. (2013).

We apply cubic spline interpolation to align the extracted features with ground truth data by matching their sampling rates, ensuring accurate temporal correspondence between video frames and ECG measurements, since the ground truth heart rate data is sampled at a higher frequency than the video frame rate. For the O-HR dataset, the ECG ground truth is sampled at 250 Hz, while the video frame rate is 25 fps. Similarly, for the MMSE-HR dataset, the ECG device operates at 1 kHz and the video at 30 fps. With interpolation we ensure temporal consistency for more accurate analysis. In Figure 2.7, we present a visual representation of an interpolated feature.

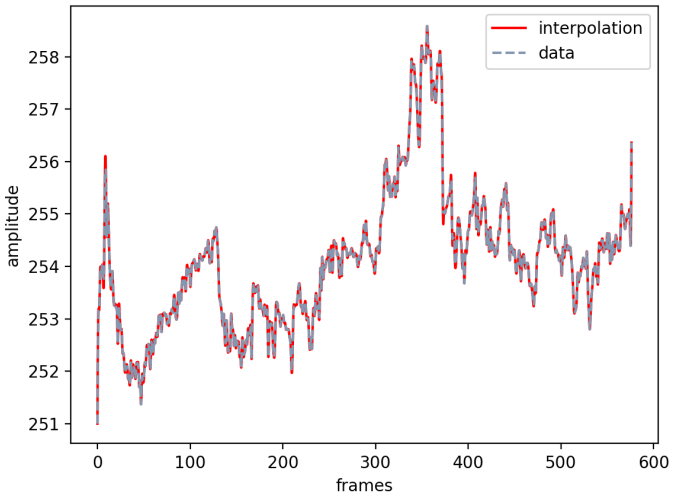


FIGURE 2.7: Comparison between the original feature signal (dashed line) and its interpolated version (solid red line). The close alignment illustrates how cubic spline interpolation effectively reconstructs a smooth, continuous signal from sparsely sampled video data, ensuring temporal consistency with higher-frequency ground truth signals.

Essentially with cubic spline interpolation, we create multiple small cubic equations (curves) between consecutive data points. Each cubic equation smoothly connects one point to the next.

In many cases, the features include noise from voluntary or involuntary head and facial movements. Since the pulse is only causing a minimal displacement, we can safely remove unstable and erratic features. To achieve this, we calculate the maximum distance traveled by each feature across frames and remove those that exceed the average maximum distance, ensuring that only stable features are retained.

Once the features are filtered, we proceed to apply temporal filtering in order to isolate the frequency range corresponding to heart rates. Typically, we only need to consider frequencies of approximately [0.75, 2] Hz, which correspond to 40-120 beats per minute (bpm). However we decide to follow Balakrishnan et al. (2013)'s approach, who suggests to consider a wider range [0.75- 5] Hz, to capture the fundamental pulse frequency (directly corresponding to heart rate) and its harmonics (multiples of the fundamental frequency). Harmonics provide additional signal components that can reinforce the true pulse frequency, particularly in cases where motion artifacts or noise weaken the fundamental frequency. We select a 5th-order filter as it effectively isolates the desired signal by smoothly passing relevant frequencies while sharply blocking noise.

After filtering, we are left with a set of features that contain a mix of cardiovascular, respiration, facial expression and natural head motion signals. We decompose the mixed

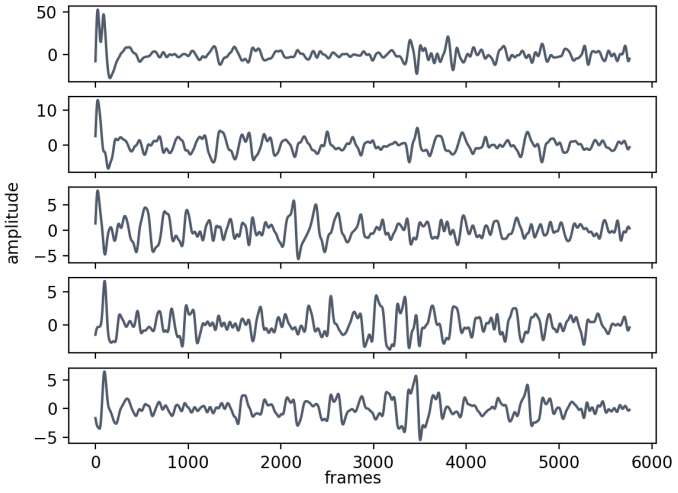


FIGURE 2.8: Extracted principal components from PCA analysis. Each component represents a different signal derived from the mixed input data, isolating distinct features such as cardiovascular activity, motion and noise.

signal into separate ones to isolate the pulse using PCA (Figure 2.8). An alternative approach we consider is ICA, but it assumes statistically independent sources, which may not always be true in rPPG where motion and physiological signals can be correlated. PCA is more effective at capturing the dominant variance in the signal.

Based on Balakrishnan et al. (2013), we focus on the first five PCA components, which capture the majority of the variance in the data. We calculate the periodicity of each component, which is defined as the proportion of spectral power concentrated at the dominant frequency and its harmonic, indicating the regularity of the signal. The component with the highest periodicity is selected for heart rate estimation (Figure 2.9).

The heart rate is computed as:

$$\text{Heart Rate} = \frac{60}{f_{pulse}} \tag{2.3}$$

where f_{pulse} is the dominant frequency of the chosen component.

This approach is computationally efficient, leveraging well-established signal processing techniques that require minimal resources.

2.2.2 Blind-Signal rPPG (BS-rPPG)

Building on Balakrishnan et al. (2013)'s framework, Ostankovich et al. (2018) introduces refinements to improve robustness and accuracy. We incorporate those refinements and

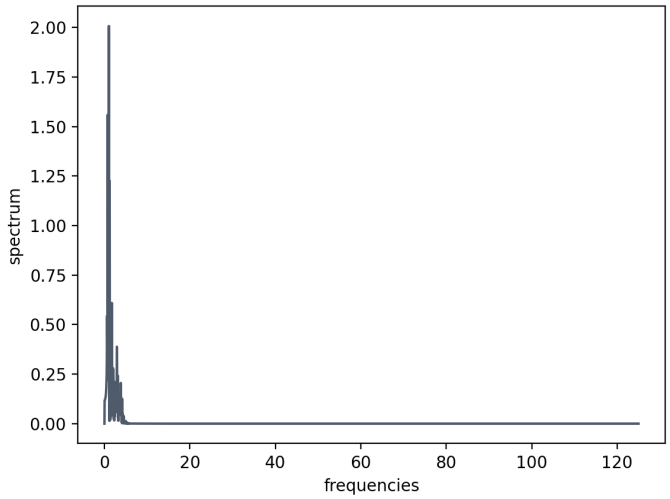


FIGURE 2.9: Periodogram of one extracted PCA component, showing the power spectrum across frequencies. The dominant peak corresponds to the pulse frequency, which is used for heart rate estimation.

we will be referring to our version of their method as Blind-Signal rPPG (BS-rPPG). While the initial steps of face detection, ROI selection and feature tracking are identical to those in MP-rPPG, several key differences distinguish this approach.

After we extract and filter features to retain only stable ones, the BS-rPPG method introduces an additional smoothing step to improve the extracted signal quality. Specifically, Singular Spectrum Analysis (SSA) is applied to each of the five PCA components to decompose them further (Figure 2.11). SSA decomposes the signal into principal components, allowing for the extraction of dominant patterns while filtering out noise. The top three components from SSA are then recombined to create smoother signals with reduced noise. This additional decomposition step enhances the separation of cardiovascular signals from noise caused by motion and illumination changes.

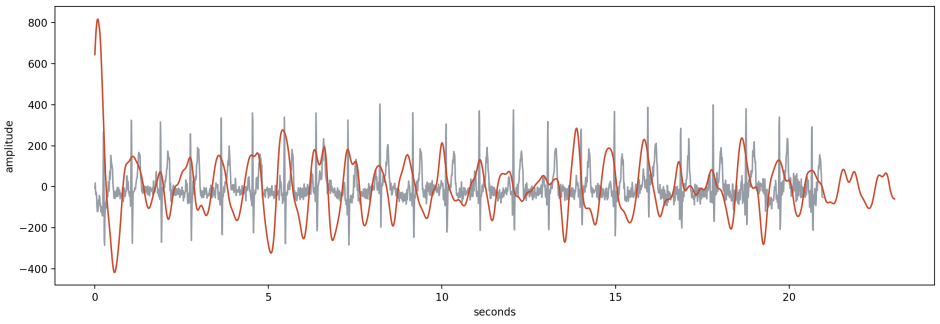


FIGURE 2.10: Extracted signal after SSA plotted alongside the ground truth ECG signal. The alignment of peaks demonstrates the effectiveness of SSA in isolating pulse-related components.

Once the signals are smoothed, the BS-rPPG approach, instead of relying on periodicity to select the best component, uses Moving Dynamic Time Warping (MDTW). Dynamic Time Warping (DTW) is a time-series alignment algorithm that compensates for temporal distortions, such as variations in heart rate over time. However, DTW is sensitive to noise and struggles with local fluctuations, which can be problematic in rPPG, where heart rate signals naturally vary.

To address these limitations, MDTW applies DTW within sliding windows rather than across the entire sequence, allowing it to dynamically adjust to temporal variations in shorter intervals. This localized adaptation enhances its robustness to signal fluctuations caused by motion artifacts or physiological variability. Additionally, MDTW improves heartbeat peak detection by transforming the signal representation into a more periodic space, making subtle pulse patterns more distinct while reducing the impact of noise and irregularities.

The heart rate is then estimated using the formula:

$$\text{Heart Rate} = \frac{60}{t_2 - t_1} * N_p \tag{2.4}$$

where t_1 and t_2 are the timestamps in seconds of the first and last detected peaks, respectively and N_p is the number of peaks within this interval. This peak-based approach is particularly effective in datasets with high motion artifacts, as it leverages the temporal consistency of the pulse signal.

Below we present a comprehensive visualization of the two methods.

2.2.3 Persistent Independent Particles for Motion-Robust Pulse Detection

In the context of estimating pulse signals from facial videos, the identification of ROIs and features within them is a fundamental step. However, the accurate tracking of these features becomes challenging when video instability is substantial. Accurate feature tracking leads to more reliable separation of cardiovascular signals from noise during subsequent processing steps. In our initial experiments we apply the Lucas-Kanade optical flow algorithm [Shi et al. (1994)]. However optical flow comes with limitations that, in scenarios like pulse estimation, could be detrimental.

The Shi et al. (1994) optical flow algorithm faces challenges when tracking features with partial facial occlusions, significant motion or changes in the feature appearance. It heavily relies on local gradients making it sensitive to noise. This can lead to inaccurate tracking, especially for subtle signals. Additionally, over time, optical flow can accumulate errors (drift), especially in long video sequences, resulting in feature misalignment. Optical flow primarily focuses on pixel intensity gradients, limiting its

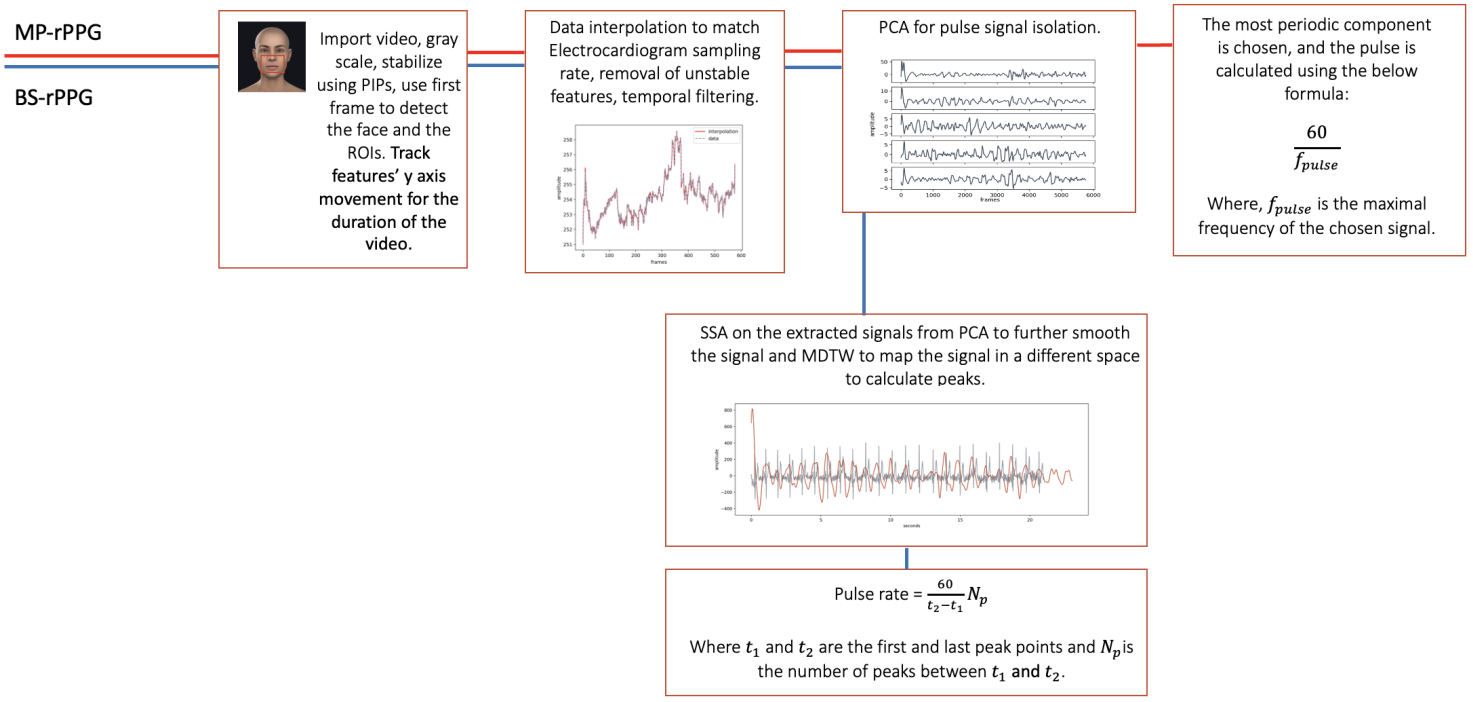


FIGURE 2.11: Diagram of MP-rPPG (Balakrishnan et al. (2013)) and BS-rPPG (Ostankovich et al. (2018)). Both methods share the initial stages of feature tracking, stabilization, interpolation and PCA-based pulse signal extraction. MP-rPPG estimates heart rate by selecting the most periodic principal component, whereas BS-rPPG applies Singular Spectrum Analysis and Moving Dynamic Time Warping to further smooth the signal and detect peaks for heart rate estimation. The diagram highlights the common and distinct stages between the two approaches.

ability to capture complex patterns in motion. Finally, optical flow algorithms operate under the assumption that the motion of pixels in a small area of an image (or a neighborhood as it is called) is consistent and follows a predictable, rigid pattern. This assumption works well for objects that do not change shape, but not when faced with facial expressions or skin movements where different parts of the face move in varying directions and with different intensities.

To address this concern, we propose the application of the Persistent Independent Particles (PIPs) algorithm for feature tracking. Our inspiration is drawn from the work of [Sand and Teller (2008)] and the follow up work of [Harley et al. (2022)], who introduce a novel motion representation paradigm referred to as "particle video". The idea is that the video is represented as a set of particles that traverse across multiple frames and we leverage long-range temporal priors while tracking them, not just current and previous frame information. It is important to note that PIPs does not require fine-tuning, making it a ready-to-use solution for improving feature tracking without dataset-specific adjustments. While PIPs utilizes deep learning components, it functions as a motion-tracking enhancement rather than a full model-based learning system, making it a compatible addition to our signal processing framework.

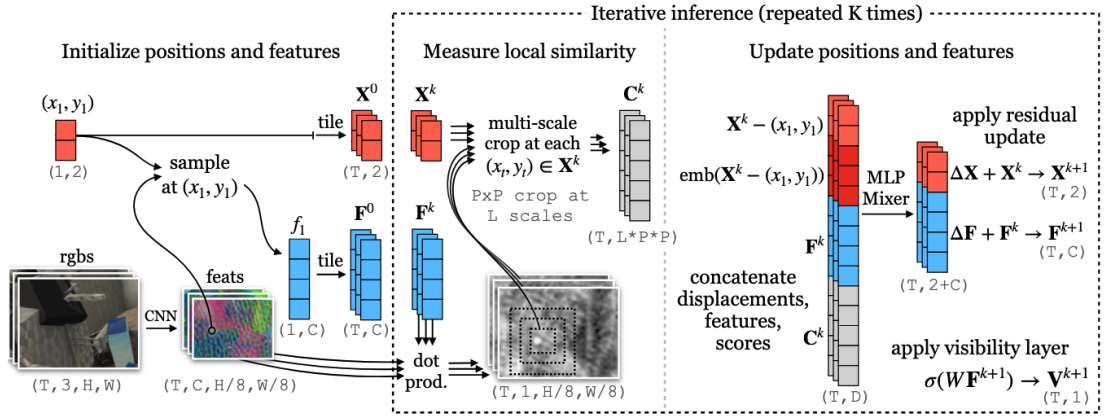


FIGURE 2.12: Persistent Independent Particles architecture: Given an RGB video as input along with a location of a feature to track, the model initializes a multi-frame trajectory, then computes features and correlation maps and iteratively updates the trajectory and its corresponding sequence of features, with a deep MLP-Mixer model [Harley et al. (2022)].

Below is a detailed description of how the PIPs algorithm works (Figure 2.12):

We consider an RGB video as input, along with the initial coordinates of the target object (in our case a feature within the ROI). The algorithm returns the coordinates of that target per frame. The framework consists of 4 main steps: extracting visual features, initializing a list of positions/features per target, measuring appearance similarity locally and updating the positions/features per target.

- **Visual Features Extraction** : Feature maps are extracted using a 2-dimensional (2D) CNN which processes each frame independently, using the "Basic Encoder" architecture from the Recurrent All-Pairs Field Transforms (RAFT) deep network architecture for optical flow [Teed and Deng (2020)].

- **Positions/Features Initialization** : After the feature maps are extracted, the features and positions for the target are initialized. The first feature map is selected and samples a feature vector at the given coordinate of the target. This feature vector represents the appearance of the target at the initial frame. To initialize the trajectory of features, this feature vector is repeated across all time frames. This initialization assumes that the appearance of the target remains constant throughout the video. To initialize the positions of the target, the initial position of the target across all time frames is copied. This initialization assumes that the target does not have any initial motion and stays at the same position throughout the video. During the final step, the trajectory of features is updated to capture variations in the target's appearance. The trajectory of positions to track the target's motion across frames is also updated.

- **Appearance Similarity Measurement** : To evaluate how well the positions and features match the pre-computed feature maps, visual similarity maps are calculated for each feature in the current iteration with the corresponding feature map of the same time frame. The local similarity scores are extracted by sampling a crop centered at the position (x_t, y_t) associated with the feature. This crop represents a small region around the target. The result is a set of patches containing un-normalized similarity scores. Larger positive values indicate higher similarity between the target's feature and the features in that particular region of the image. It has been found beneficial to create a spatial pyramid of these score patches. This allows to obtain similarity measurements at multiple scales, capturing different levels of details.

- **Iterative Updates** : In the main inference step, the sequences of positions and features are updated. Displacements from the initial positions are computed and encoded using sinusoidal position encodings. These displacements, along with the features, are concatenated and processed by an Multi-layer Perceptron Mixer (MLP-Mixer) architecture, which produces updates for the positions and features. The updates are applied iteratively and after the final update, the positions are considered the final trajectory. Visibility scores for each time step are estimated using a linear layer and sigmoid activation. The model is supervised during training using the L1 distance between the ground-truth trajectory and the estimated trajectory.

This addresses all our previous concerns with optical flow; PIPs is specifically designed to handle inconsistencies through frames even if features are temporarily occluded or

distorted. At the same time, PIPs leverages long-range temporal priors, which allows it to maintain more stable trajectories over time, even in noisy conditions or under varying lighting. It also minimizes the effects of drift. PIPs integrates modern feature extraction methods, such as CNN-based encoders, which allows for more robust tracking by incorporating richer visual information. At the same time, it can handle non-rigid motion more effectively by treating particles as independent entities. Finally, despite the fact that PIPs leverages advanced CNN-based feature extraction, its computational efficiency remains manageable for real time applications due to its iterative and localized update process.

In this chapter, we take advantage of the pre-trained PIPs algorithm to track the extracted spatial features in our motion-based approach. The output of PIPs is the positions of the features through the frames for the duration of the video. We compare this approach to the optical flow tracking algorithm in our results section.

2.3 Evaluating Signal Processing for rPPG

We perform extensive parameter tuning to optimize the performance of the signal processing algorithms. This is one of the first issues we come across; a significant amount of manual labor. It is known that signal processing algorithms require substantial parameter tuning to optimize performance because they are handcrafted methods designed to target specific features or characteristics of the input data.

Various configurations are evaluated to minimize the MAE between the ground truth and the predicted pulse signal with the key categories being: Corner Detection Parameters, Filtering Parameters, ROIs and Quality Levels.

Below we describe in more detail what parameters each category includes along with their description:

Corner Detection Parameters:

- **Maximum Corners:** This specifies the maximum number of features that the algorithm will detect within the frame. A higher number increases the density of features, improving robustness in complex scenes but potentially introducing redundant or less meaningful features.
- **Minimum Distance:** Defines the minimum Euclidean distance between detected corners. A smaller value allows more closely spaced features to be detected, while a larger value ensures that features are spread out, reducing redundancy. The Mahalanobis distance was also tested in our experiments.

- **Block Size:** Determines the size of the neighborhood used for corner detection. A larger block size averages over more pixels, improving stability in noisy images but potentially missing finer details.

Filtering Parameters:

- **Butterworth Filter Order:** The filter order controls the sharpness of the frequency cutoff. Higher-order filters have steeper roll-offs, effectively isolating the frequency band of interest (e.g., 0.65–4 Hz for heart rate signals).
- **Frequency Range:** Specifies the band of frequencies to preserve. As mentioned in Section 2.2.1.1 the band is chosen to include typical heart rate frequencies while suppressing noise from motion or environmental artifacts.

Regions of Interest:

- **ROI Selection:** This parameter refers to the areas of the face analyzed for pulse extraction, such as forehead, cheeks, nose, mouth and neck areas and their combinations.

Quality Levels:

- **Feature Quality Threshold:** Represents the minimum quality score for detected features, ensuring that only high-confidence features are used. A higher threshold results in more reliable tracking but may exclude useful features in noisy conditions.
- **Distance Constraints:** Defines limits on the spatial relationships between features to ensure consistency across frames. For example, maintaining a minimum distance prevents overlapping or unstable feature points. It must be noted that Distance Constraints are different from Minimum Distance, which is about selecting features within a single frame to avoid redundancy and ensure even coverage. Distance constraints apply across frames to maintain stable tracking and avoid large or unrealistic deviations due to noise, motion or tracking errors.

Each configuration is evaluated on the O-HR dataset comprising of both normal and physical activity videos, with results recorded as MAEs to facilitate direct comparison.

The default parameter configuration provides a baseline average MAE of **15.82** for normal videos and **16.12** for physical activity videos. This setup uses standard values for corner detection and filtering parameters without region-specific adjustments. After a number of experiments we derive the optimal parameters

2.3.1 Optimizing Performance

2.3.1.1 Where to Look? The Impact of ROI Selection

The analysis of the performance across ROIs provides insights into the variability of rPPG performance depending on the selected facial region. We compute the error for: the forehead, mouth/cheeks/nose, cheeks/nose and combinations, such as mouth/cheek/nose/neck and mouth/cheek/nose/forehead/neck, as can also be seen in Figure 2.6. The results for these areas for the BS-rPPG algorithm are presented in the accompanying bar chart, which illustrates the mean MAE for each ROI along with error bars for SDs.

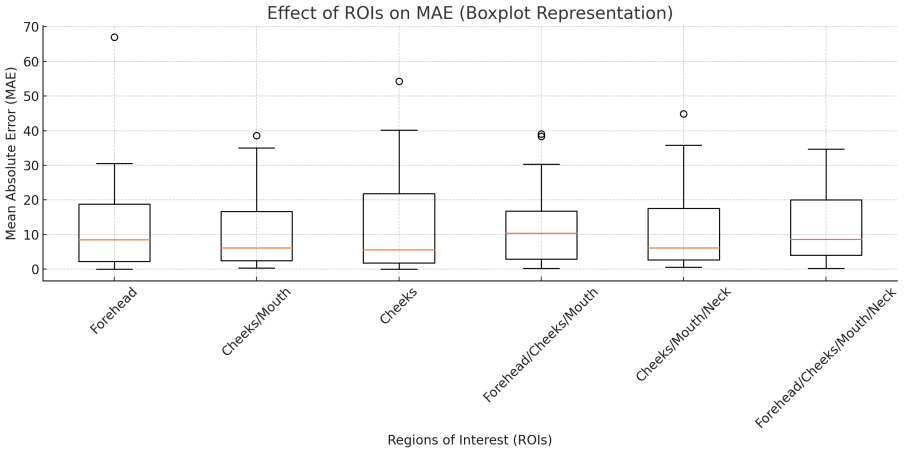


FIGURE 2.13: Boxplot comparison of heart rate estimation MAE across different ROIs for BS-rPPG on the O-HR dataset. The boxes represent the range, the orange lines indicate the median and whiskers show the full range of errors. Combining multiple facial regions generally results in slightly lower and more stable MAE values compared to using single regions.

According to Figure 2.13, among all ROIs, the cheeks/mouth region has the lowest average MAE, indicating its better average performance for pulse signal extraction using the BS-rPPG algorithm. This region likely benefits from limited occlusion (compared to areas like the forehead, which might be occluded by hair) and consistent illumination. However, while the cheeks/mouth region achieves the lowest mean MAE, the improvement over other regions is not substantial, suggesting that all regions carry valuable information for pulse estimation.

One particularly interesting insight is the significant variability in performance for each region, as indicated as indicated by the spread in the boxplots. These error bars reflect how consistent or inconsistent the performance of each ROI is across participants. The cheeks/mouth region shows slightly lower variability compared to the forehead, indicating that it may provide more consistent results. Nonetheless, the variability is not negligible, highlighting that the optimal ROI can differ significantly between individuals.

This aligns with real-world considerations: some participants may have occlusions (e.g., hair covering the forehead, facial hair, glasses), skin tone variations, facial structure or lighting conditions which can affect the signal quality from specific regions. In other words, while the mouth/cheeks region may be optimal on average for these experiments, it is not universally the best-performing region for all participants.

This observation is further supported by Figure 2.14, which evaluates the same dataset but with a different algorithm, MP-rPPG. In this case, the forehead appears to be the best-performing region, achieving the lowest mean MAE and similar or slightly better consistency compared to the cheeks/mouth region. This demonstrates that the algorithm used for signal extraction influences which ROI performs best. The forehead, for instance, may be more effective for MP-rPPG due to its relatively stable surface and reduced motion artifacts compared to the cheeks/mouth region, which might be affected by speaking or expressions.

Our findings align well with literature suggesting that the effectiveness of different ROIs can vary depending on both the algorithmic approach and recording conditions. In our experiments, the cheeks and mouth region performed best for BS-rPPG, while the forehead was more reliable for MP-rPPG, which aligns with observations and highlights that the optimal ROI is context and method-dependent. These results therefore confirm previous findings that combining multiple well-lit, stable facial areas generally improves signal robustness and consistency across subjects. However, they also illustrate that algorithmic design such as whether spatial filtering or blind source separation is used, can shift the balance between motion sensitivity and signal strength, influencing which ROI performs best in practice.

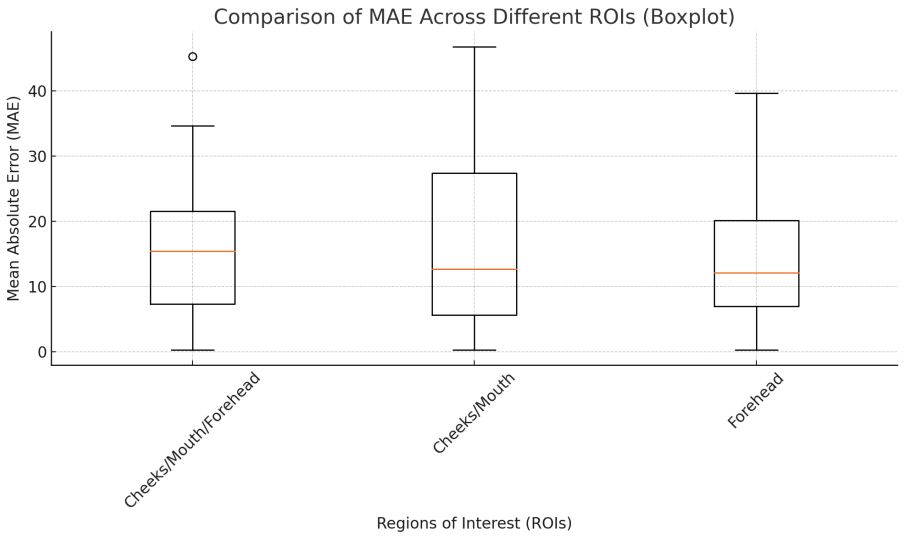


FIGURE 2.14: Boxplot comparison of heart rate estimation MAE across different ROIs for the O-HR dataset using the MP-rPPG algorithm.

The findings from both figures suggest that there is no universal "one-size-fits-all" ROI for pulse signal extraction. Instead, the performance of an ROI depends on both the algorithm and the individual characteristics of the participant. This underscores the need for a dynamic approach to ROI selection, where the optimal ROI is chosen based on the specific algorithm and the conditions of the input video. Such an adaptive strategy could improve the accuracy and reliability of pulse estimation across diverse participants and settings - something we explore in subsequent chapters.

2.3.1.2 More Features, Better Pulse?

Figure 2.15 illustrates the relationship between the number of features used in the analysis and the MAE for a sample participant. The graph reveals a clear trend: as the number of features increases, the MAE decreases sharply in the initial range. This suggests that the addition of features significantly enhances the model's ability to extract pulse signals accurately, particularly when moving from a very limited to a larger feature set.

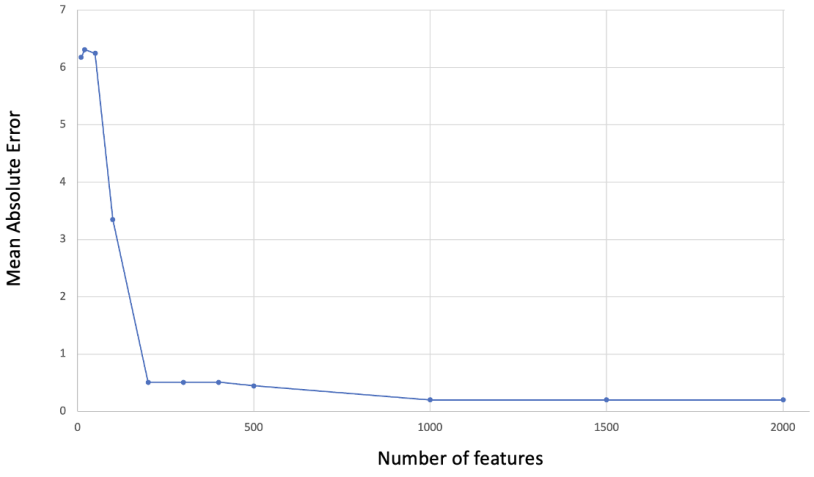


FIGURE 2.15: The effect of the number of features on MAE. The graph highlights the plateau observed after 1000 features.

However, beyond the point of 1000 features, the MAE plateaus. Adding more features at this point does not lead to substantial error decrease, if anything it was observed that, in some cases, it slightly degraded performance, suggesting the possibility of overfitting or an increase in noise introduced by redundant or irrelevant features. This underscores the importance of balancing the number of features to optimize accuracy while avoiding unnecessary computational overhead.

Although Figure 2.15 represents a single participant, similar trends are observed across multiple subjects. It is worth noting, however, that there is variation in the exact number of features at which the MAE stabilizes. Factors such as individual physiological characteristics and video quality can influence this threshold.

To ensure a balance between computational efficiency and accuracy, the number of features is capped at 1000 across experiments. We select this value based on the observed trends and we believe it provides a robust compromise, allowing the model to perform optimally without adding excessive computational weight.

2.3.1.3 Filtering the Noise

We observe that the choice of filter parameters has an impact on the accuracy of our methods. The results for three different Butterworth filter configurations are summarized below:

0.65–4 Hz (2nd-Order Butterworth Filter)
The 2nd-order Butterworth filter shows the highest average MAE of 19.81, with an SD of 15.74. While this filter captures the frequency band associated with pulse, its lower order appears it is not effectively filtering out noise and artifacts, leading to relatively higher errors. The results indicate that while the 2nd-order filter is functional, it is suboptimal for robust and accurate signal extraction.

0.75–5 Hz (5th-Order Butterworth Filter)
The 5th-order Butterworth filter is the best-performing configuration, achieving the lowest average MAE of 15.36 and the smallest SD of 11.29. This demonstrates its ability to balance capturing the desired frequency range and cutting out noise. Its higher performance suggests that this configuration is more reliable and suitable for pulse signal extraction across varying conditions.

0.75–5 Hz (10th-Order Butterworth Filter)
The 10th-order Butterworth filter exhibits intermediate performance, with an average MAE of 16.12 and a SD of 13.93. Although its higher order allows for greater sensitivity to subtle signal features, it may be amplifying noise or irrelevant details, which could explain its slightly higher errors and variability compared to the 5th-order filter.

These findings align well with Balakrishnan et al. (2013), which highlights the importance of selecting appropriate filter parameters for optimizing accuracy. The 5th-order Butterworth filter provides a better trade-off between accuracy, robustness and noise suppression, making it an ideal choice for our problem.

2.3.1.4 Tuning Supporting Parameters

The additional parameters, including the quality threshold, distance between corners and block size, are adjusted empirically based on initial exploratory experiments. While their influence on the results is observed to be less noticeable compared to ROI selection and filter design, they are optimized to ensure consistency, as well as to balance computational efficiency with accuracy.

2.3.2 Benchmarking Against Existing Work

This section presents a comparison of our proposed methods with existing approaches, as summarized in Table 2.4. The performance of each method is evaluated using MAE and SD across normal, physical and all activity conditions, alongside the identified optimal ROIs for the O-HR dataset.

We compare our algorithms, MP-rPPG and BS-rPPG, only against the work of Ostankovich et al. (2018), as this is the only prior signal processing approach using the same dataset. No deep learning models have been trained on this dataset due to its small size, and given the fundamental differences between deep learning and signal processing approaches, a direct comparison would not be meaningful. Another reason behind this choice is that the method proposed by Balakrishnan et al. (2013) utilizes a private dataset, which we could not obtain access to for direct comparison.

TABLE 2.4: Comparison of the MAE and standard deviation across normal, physical, and all activity conditions for the method proposed by Ostankovich et al. (2018) and our approaches (MP-rPPG and BS-rPPG). The table also highlights the optimal ROIs identified by each method.

Method	Normal		Physical		All Activities		Optimal ROI
	MAE	STD	MAE	STD	MAE	STD	
Ostankovich et al. (2018)	11.7 ± 3	12.0	12.7 ± 3	13.0	12.2 ± 2	13.0	Forehead/ Cheeks/ Mouth
MP-rPPG	11.2 ± 2	8.4	17.2 ± 3	12.1	14.2 ± 2	10.7	Forehead
BS-rPPG	**10.6 ± 3**	**10.5**	**11.2 ± 3**	**13.1**	**10.9 ± 2**	**11.7**	**Cheeks/ Mouth**

For normal activity conditions, BS-rPPG achieves the lowest MAE, demonstrating the highest accuracy compared to both MP-rPPG and Ostankovich et al. (2018). However, its higher variability suggests that while it performs better on average, its results are less consistent than those of MP-rPPG in the normal scenario. MP-rPPG showcases similar average performance but with lower variability, highlighting its robustness in normal conditions. In contrast, Ostankovich et al. (2018) showcases reasonable accuracy but with higher fluctuations between participants.

Post-physical activity, the performance of all methods declines due to the increased motion noise and physiological variability. Despite this, BS-rPPG maintains its position as the best-performing method, reflecting its ability to adapt to such conditions. MP-rPPG experiences a more pronounced drop in accuracy, suggesting that it is more sensitive to the effects of physical activity. Ostankovich et al. (2018) demonstrates relatively stable performance but with persistent high variability.

When considering all activities, BS-rPPG stands out once more with the lowest overall MAE, underscoring its effectiveness. MP-rPPG performs slightly worse overall but maintains lower variability compared to Ostankovich et al. (2018). The results suggest that while BS-rPPG performs better in terms of accuracy, MP-rPPG offers more consistent performance across diverse conditions.

The analysis of optimal ROIs reveals interesting differences between the methods, which we discussed extensively in Section 2.3.1.1.

Overall, BS-rPPG outperforms MP-rPPG in both normal and physical conditions, achieving the lowest MAE across all activities. However, MP-rPPG demonstrates more stable performance, especially in controlled conditions. The choice of ROI significantly influences performance, with BS-rPPG benefiting from cheeks/mouth, while MP-rPPG performs better with forehead signals. These findings suggest that adaptable ROI selection could further enhance accuracy.

2.3.3 The Impact of PIPs Feature Tracking

Table 2.5 provides an analysis of the BS-rPPG method with and without enhancements such as PIPs for feature tracking and K-means clustering for feature clustering. The detailed description of PIPs implementation can be found in Section 2.2.2.

When using K-means clustering, the 1000 features per ROI are grouped into 25 clusters. The centroids of these clusters serve as the new features, significantly reducing the total feature number while preserving the most important information. This clustering approach not only reduces computational complexity but also improves the robustness of feature trajectories by focusing on the behavior of the cluster centroids.

The results in Table 2.5 demonstrate how these enhancements impact the performance of the BS-rPPG method, showcasing the effectiveness of PIPs and K-means clustering in improving accuracy and optimizing feature utilization.

Introducing PIPs for feature tracking significantly improves the results, as seen in Table 2.5, reducing the MAE for normal, physical activities separately and across all activities. This demonstrates the effectiveness of PIPs in enhancing feature selection and tracking, particularly for normal activities. Additionally, the identified optimal ROIs

TABLE 2.5: Comparison of the MAE and standard deviation across normal, physical, and all activity conditions for our BS-rPPG approach with and without PIPs for feature tracking and K-means clustering for feature clustering. The table also highlights the optimal ROIs identified by each method.

Method	Normal		Physical		All Activities		Optimal ROI
	MAE	STD	MAE	STD	MAE	STD	
BS-rPPG	10.6 $\pm$ 3	10.5	11.2 $\pm$ 3	13.1	10.9 $\pm$ 2	11.7	Cheeks/ Mouth
BS-rPPG with PIPs	7.7 ± 2	6.4	10.5 $\pm$ 2	8	9.1 ± 1	7.3	Forehead/ Cheeks/ Mouth
BS-rPPG with K-means and PIPs	$\mathbf{7.5 \pm 2}$	**6.3**	$\mathbf{8 \pm 1}$	**5**	$\mathbf{7.8 \pm 1}$	**5.7**	**Forehead/ Cheeks/ Mouth**

expand to include the forehead, cheeks and mouth, suggesting a better utilization of facial regions.

Integrating K-means clustering with PIPs showcases the best performance among all configurations. These results indicate that the combination of PIPs and K-means clustering enables more robust and precise feature clustering and tracking, leading to substantial improvements in signal extraction and computational overhead. The optimal ROIs remain the forehead, cheeks and mouth.

Overall, the improvements from the baseline BS-rPPG method to the integration of PIPs and K-means clustering demonstrate the benefits of these techniques in improving both accuracy and robustness. The results highlight the benefits of clustering features, particularly in challenging conditions such as physical activities and underscore the importance of multiple ROIs for optimal performance.

On the other hand, the high performance variability across participants in the dataset reveals the significant impact of appearance and behavior on the accuracy of pulse signal extraction. Participants who remain relatively still during the video recordings consistently outperform those who exhibited more dynamic behavior, such as talking, laughing or making large gestures. Performance gaps are significant, with the best performing participants consistently showcasing MAEs lower than 2, in contrast with worst performing participants who showcase MAEs over 10, occasionally reaching 30. This observation aligns well with the known sensitivity of rPPG methods to motion artifacts. When participants move excessively or make abrupt gestures, the captured signal is more likely to be noisy, making it harder for the model to accurately extract pulse information.

After physical activity, participants may exhibit increased heart rates and more pronounced facial movements, which can introduce noise and distort the physiological signals being measured. This highlights another key challenge on extracted signal quality.

However, the inclusion of PIPs for feature tracking appears to mitigate these issues to some extent. By leveraging PIPs, the model benefits from more stable features, as proven by the improved performance in scenarios with higher motion. PIPs allows for a more robust tracking, even when participants exhibit some movement, reducing the impact of motion artifacts.

These findings emphasize the importance of mitigating the effects of motion artifacts and physiological variability in order to improve the robustness of rPPG methods. While static scenarios may allow models to perform optimally, more challenging conditions reflect real life cases best, but require more advanced stabilization techniques or adaptive modeling approaches to maintain accuracy. This highlights the need for dynamic ROI selection, advanced filtering methods and techniques to moderate motion, all of which can help enhance performance.

Based on the results, we observe trends related to gender and skin tone. Female participants generally exhibit higher MAEs compared to their male counterparts in both normal and physical conditions. With females comprising less than 15% of the dataset's participants, this observation aligns with previous findings that models trained predominantly on male physiology struggle to generalize effectively to female physiology due to differences in facial features and skin reflectance. Makeup is a significant suppressant of signal as it covers the facial changes necessary to extract pulse, however, to our knowledge, participants in this dataset do not use foundation. This also highlights the need for dynamic ROI selection. Similarly, the single identified person of color in this dataset also demonstrates higher MAEs across multiple configurations.

These findings highlight significant limitations in the generalizability of signal processing-based methods, which are influenced by the demographic distribution of the data. With the original dataset being predominantly male and light-skinned, the parameter tuning does not adequately represent the diversity of real-world populations.

2.3.4 Generalizing to New Datasets

To assess the generalizability of the calibrated BS-rPPG method, we apply it on the MMSE-HR dataset. Unlike previous experiments, this evaluation does not incorporate PIPs or K-means clustering. Instead, this experiment aims to showcase the substantial manual effort required to adapt signal processing approaches to new datasets.

Results by Skin Tone and Gender

Table 2.1 summarizes the results on the MMSE-HR dataset, grouped by skin tone (Fitzpatrick scale) and gender. The MAE values indicate a substantial decline in performance compared to O-HR. For example, male participants with skin type III achieved

TABLE 2.6: Comparison of MAE on the MMSE-HR dataset across different skin tones (Fitzpatrick scale [Fitzpatrick (1988)]), broken down by gender with standard deviation presented in the parenthesis. The table provides insight into performance variation across skin tones III–VI and highlights overall MAE per gender and per skin tone.

Skin tone/ Gender	III	IV	V	VI	MAE per Gender
Male	20.9 ± 3(16)	26.8 ± 6(11)	62.6	-	$23.3 \pm 3(15)$
Female	26.7 ± 3(20)	36.8 ± 2(3)	-	27.1	$27.4 \pm 3(20)$
MAE per Skin tone	24.7 ± 2(19)	31.8 ± 3(9)	62.6	27.1	$25.9 \pm 2(19)$

TABLE 2.7: Comparison of MAE on the MMSE-HR dataset across different activity types (T1–T14 - activity table can be found in 2.3), disaggregated by gender with standard deviation presented in the parenthesis. This table highlights how task type affect rPPG performance and provides the average MAE per activity and per gender.

Activity/ Gender	T1	T8	T9	T10	T11	T14	MAE per Gender
Male	4.1 ± 4(4)	40.7 ± 2(2)	-	18.5 ± 3(11)	28.9 ± 6(23)	-	23.3 ± 3(19)
Female	23.9 ± 7(19)	38.3 ± 12(31)	10.1	26.7 ± 4(18)	26.6 ± 4(16)	34.3	27.4 ± 3(20)
MAE per Activity	19.9 ± 6(19)	38.8 ± 9(27)	10.1	23.1 ± 3(16)	27.6 ± 3(19)	34.3	25.9 ± 2(19)

a MAE of 20.9±16, while those with skin type IV a MAE of 26.8±11. Performance deteriorated significantly for participants with darker skin tones (type V), highlighting the challenges for individuals with higher melanin levels. For females, the performance had similar variations, with an overall MAE of 27.4±20, further proving that participant demographics have a significant impact on performance. The overall trend shows that participants with lighter skin tones performed better.

Results by Activity

Table 2.7 presents the results on MMSE-HR by activity. T1, which involved listening to a funny joke, yielded the lowest MAE, while tasks such as T8 (improvising a silly song) and T14 (experiencing a smelly odour) showed significantly higher errors, reflecting the challenges posed by motion noise.

Overall, the results highlight that participants who moved their heads naturally while talking contributed to higher error rates, as these movements introduced significant motion artifacts. This aligns with findings from O-HR, where static participants showcased better performance. The MMSE-HR dataset inherently introduces more motion

noise due to its design, which explains the higher MAE across tasks and participants, especially when the model is tuned to a more static dataset like O-HR.

2.4 Challenges, Insights & Future Directions

Signal processing has long been the foundation of rPPG research, offering transparent, efficient, and interpretable solutions for extracting physiological signals from video. The methods explored in this chapter - MP-rPPG and BS-rPPG - demonstrate strong performance under controlled conditions and continue to be valuable tools for pulse estimation. Our results show that these techniques can achieve low MAE values when carefully tuned. Their computational efficiency makes them well-suited for real-time applications, even on standard hardware.

A key advantage of signal processing methods is that they do not require large amounts of training data, unlike deep learning models, which depend on extensive datasets for generalization. This makes them particularly useful in scenarios where data collection is challenging or when working with smaller datasets. In particular, BS-rPPG performed exceptionally well across both normal and physical activity conditions, with further improvements when incorporating PIPs for feature tracking and K-means clustering for feature selection. Additionally, ROI selection played a crucial role, with different regions (forehead vs. cheeks/mouth) proving optimal depending on the method. These findings reinforce the strengths of signal processing in providing reliable and explainable solutions for rPPG, especially in scenarios where speed and efficiency are crucial.

However, real-world deployment presents new challenges, as demonstrated in our results. While signal processing techniques are highly effective, they require manual parameter tuning, making them less adaptable to new datasets, diverse demographics or high-motion conditions. For instance, our results indicate that performance can vary significantly across skin tones and facial movements, particularly in more dynamic environments like MMSE-HR, where head motion and lighting changes introduce variability. Motion artifacts remain one of the biggest obstacles, as even small facial movements can introduce noise that distorts pulse estimation. The limited diversity in publicly available datasets further magnifies this issue, underscoring the need for broader data collection efforts and increased availability for researchers.

Despite these challenges, signal processing methods remain an essential component of rPPG research. They provide strong baselines, valuable insights into the nature of pulse signals and efficient alternatives for real-time applications. However, as rPPG expands into more complex settings such as real-world monitoring, wearables and multi-person scenarios, more adaptive solutions are needed.

The next chapter explores spatiotemporal methods, which aim to address the challenges of motion artifacts and generalization by learning representations directly from data. While deep learning introduces new trade-offs such as increased computational complexity, its ability to adapt to different conditions without requiring manual tuning makes it a promising complement to traditional signal processing techniques. By integrating the efficiency and interpretability of signal processing with the adaptability of deep learning, we move towards more robust and scalable rPPG solutions.

Chapter 3

Hidden Layers: Capturing Spatiotemporal Patterns

This chapter embodies the hidden layers of an ML model, where the complex interplay of light and motion is learned. These layers work to uncover the intricate patterns in the data, laying the foundation for robust predictions.

In the previous chapter, we focused on signal processing approaches, which form the basis of the rPPG field and we uncovered their strengths but most importantly their shortcomings. Despite the fact that they are computationally light, easily interpretable and intuitive, they present significant limitations.

Some of the most critical challenges of such algorithms involve sensitivity to motion artifacts and variations in illumination, reliance on handcrafted features, which requires labor-intensive and time-consuming tuning, and their limited generalizability. While advancements such as PIPs for feature tracking and feature clustering using K-means improve accuracy, these methods still struggle in dynamic settings.

A recurring challenge in our experiments was that participants post-exercise or exhibiting natural behaviors like speaking and moving, introduce significant noise to the extracted signals, degrading performance. Participant demographics also prove to be a challenge, with female participants and people of color performing significantly worse. With signal processing approaches relying on dataset specific parameter tuning, their ability to generalize across populations with diverse demographics or conditions is limited, especially when these populations are under-representated in the data. This

challenge is not exclusive to signal processing methods. There is a known under representation of certain populations in datasets currently available for research, which creates biased results. This is a broader problem and conversation that must be addressed by researchers, with the collection of inclusive datasets or the use of generative AI to combat such inequalities. Another interesting theme that was observed in our previous experiments was the ROI selection. As we discussed, there is no "one size fits all" approach, as physiology, appearance and the choice of model can have an effect on the optimal area selection.

Over the years, as ML research advanced and as resources became more easily accessible, researchers started incorporating simple ML algorithms for pulse estimation to signal processing frameworks, until this day, where complex, DL solutions dominate the field. Unlike signal processing, ML techniques automate feature extraction, eliminating the need for handcrafted features and such intense "hands-on" parameter tuning. They are able to capture complex, subtle patterns in video data that are not immediately apparent to humans, even with extensive knowledge of physiology, helping them learn representations that are robust to noise, motion and illumination changes. ML solutions can generalize better to unseen data and their scalability and adaptability makes them particularly suited to modern, real-world rPPG applications, where diversity and variability are the norms. They can easily handle large and diverse datasets, leveraging parallel computing, distributed processing and Graphics Processing Unit (GPU) acceleration. This automation of the pulse extraction process was very attractive to enthusiasts of the field, who started continuously developing more complex but accurate solutions.

However, as models get more complex, a lot of their interpretability is lost, making it challenging to understand how the models make decisions and when they will fail. This lack of transparency is incredibly problematic, especially in healthcare applications, where accuracy is of the essence. As models become deeper, their processing needs grow with them. Models can no longer be as easily implemented on small devices, real-time processing becomes computationally expensive and their training requires extensive amounts of labeled data, which are timely and expensive to collect.

Our research interests have always focused on developing practical and accessible solutions rather than increasing complexity. By combining the transparency and simplicity of signal processing approaches with the powerful performance of ML methods, we believe it is possible to strike a balance - creating solutions that are not only effective but also easy to implement in the real-world.

Hybrid methods extract handcrafted features, that are then processed by ML or DL networks for pulse estimation. They take advantage of spatiotemporal maps which are subsequently fed through a machine or DL framework to extract the pulse rate.

The use of spatiotemporal features as a means to estimate heart rate has many benefits compared to the traditional use of videos. Such approaches can result in higher temporal resolution compared to video frames, providing more detailed information about the skin changes over time and can help mitigate the impact of motion and light artifacts by integrating information over time, leading to more robust pulse estimation. Some of the works we presented in chapter 1 have demonstrated promising results, however there are still components that have not been addressed. All these works rely on pre-defined ROIs, which could neglect regions with sufficient signal to assist in the increased accuracy of pulse estimation as we proved in chapter 2. Additionally, averaging information from multiple frames in a single image can result in significant signal variations being suppressed.

Chapter Contributions:
Building on the advancements presented in chapter 1 and our observations from chapter 2, we introduce the Spatiotemporal Two-Stage Learning Approach (ST2S-rPPG), a framework designed to address the limitations of signal processing methods while leveraging the strengths of spatiotemporal models. ST2S-rPPG combines the interpretability of traditional techniques with the automation and adaptability of modern ML models. More specifically, our contributions are as follows:

- We propose a stabilization method using the PIPs algorithm to address motion artifacts. Building on PIPs for feature tracking as described in chapter 2, we leverage the same algorithmic foundation for video stabilization.
- We develop a novel spatiotemporal representation of video to images. This methodology could be particularly valuable in healthcare settings where data availability and computational resources are often limited, allowing for more robust training of ML algorithms.
- We design a two stage learning framework to optimise estimation accuracy by selecting the most informative spatiotemporal images. The two-stage learning approach can lead to recommendations that are more accurate and calibrated to each individual, ultimately improving outcomes.
- We propose representing the ground truth as beats per video segment rather than BPM. This approach simplifies the task for the ML model, enabling it to directly identify peaks in the spatiotemporal images without the added complexity of converting segment data into a per-minute metric.
- We test the Eulerian video magnification method to enhance subtle changes occurring on the skin during the cardiac cycle.

The rest of the chapter is organized as follows: Section 3.1 introduces the datasets and experimental setup, detailing participant demographics, data collection protocols and

evaluation metrics. Section 3.2 describes ST2S-rPPG framework, along with advancements like PIPs for video stabilization. Section 3.3 presents results, comparing the methods across datasets and conditions, while Section 3.4 concludes with a discussion of the findings, limitations and implications for future work. Finding of this chapter were published in Machine Learning for Health (PMLR, 259:550–562, 2024).

3.1 Datasets and Experimental Framework

We evaluate the performance of ST2S-rPPG on two benchmark datasets, MMSE-HR, as described in Section 2.1.2 and the more recently obtained Université Bourgogne Franche-Comté dataset for rPPG (UBFC-rPPG) [Sabour et al. (2021)]. These datasets are widely used in the field of rPPG and allow us to compare our methodology with state-of-the-art approaches. For that reason we decide to discard O-HR, as there are no significant works outside of Ostankovich et al. (2018)'s signal processing and thus, we cannot benchmark our current work against it.

We must note that MMSE-HR and UBFC-rPPG have different baselines due to their different characteristics (frame rate, resolution, collection protocol). Our work uses subsets of the original datasets, specifically formatted for rPPG. Despite our continuous efforts to expand our study with additional datasets such as PURE and VIPL-HR - both of which are also commonly used in rPPG research - we were unable to secure access to them. Despite this, MMSE-HR and UBFC-rPPG provide sufficient variability to validate the generalizability of our approach.

We implement ST2S-rPPG using PyTorch and one NVIDIA GeForce GTX 1080 GPU. Below we present UBFC-rPPG's characteristics in detail. MMSE-HR's characteristics can be found in Section 2.1.2.

3.1.1 The UBFC-rPPG Dataset: A Closer Look

This is a subset of the UBFC-Phys database, designed to test the accuracy of rPPG algorithms. It consists of 40 RGB videos and corresponding ground truth heart-rate data obtained from 40 participants. Each video provided is recorded at a resolution of 640x480 pixels and a frame rate of 30fps. This is significantly lower than MMSE-HR which gives us a good indication of our algorithm's accuracy in varying video qualities. Each video duration varies from 46 seconds to 1 minute 8 seconds, with most videos closer to the minute mark. Each video is synchronized with a pulse oximeter finger clip sensor to collect the ground truth.

Participant demographics include:

- **Gender:** 80% male
- **Facial Hair:** 27.5%
- **Skin Tone:** 92.5% of participants fall within skin types I-V, with 92.5% in I-III, 5% in IV and 2.5% in V.

3.1.2 Ground Truth: From Raw Signals to Usable Data

Ground truth measurements are provided for both datasets. In these experiments we convert these measurements to beats per 10 second segments using the procedure below:

For the MMSE-HR dataset, ground truth heart rate data is acquired through a contact sensor operating at a sample rate of 1 KHz, providing pulse measurements per frame. In MMSE-HR, the definition of the heart rate ground truth data is that the measurement changes every time there is a heartbeat. We define each 10 second time segment as $[t_{start}, t_{end}]$, where t_{start} is the starting time and t_{end} the ending time of the segment. To identify the location of these segments we multiply the start and end time with the sampling rate. Within each segment we count the number of changes in the provided ground truth files, each change is a heartbeat.

For the UBFC-rPPG dataset, we use the raw signal data and the scipy library *find_peaks* to identify the beats. With the same process as above we count the number of peaks per segment. This process provides granular information regarding the pulse variability within each 10 second segment, which is not necessarily visible by using the average measurements for the whole video. The idea behind this choice is that the pulse estimation CNN will be able to distinguish beats easier than extrapolated bpm in each 10 second segment. After we compile our results we multiply the predicted value by 6 to extract the BPM measurement and compare performance with existing methodologies.

3.2 Building the ST2S-rPPG Framework

The proposed ST2S-rPPG framework is divided into five steps, face identification and video stabilization, Eulerian Video Magnification, spatiotemporal image generation, pulse estimation using a CNN for regression and a second learning component to improve estimation. In the following sections, these steps are described in detail.

3.2.1 Face Tracking and Video Stabilization

In the context of estimating pulse from facial video data, the identification of ROIs is a fundamental step. However, the accurate tracking of these regions becomes challenging when confronted with video instability, stemming from voluntary or involuntary movements. To address this concern, we observe the need for an accurate stabilization tool, with the primary objective being to facilitate the tracking of the facial region within video sequences.

3.2.1.1 Video Pre-processing for Precision

Prior to applying the stabilization step, we observe that even within the same video, different segments may exhibit unique physiological patterns. We segment the original video V into discrete 10-second segments to not only maximize the utilization of available data but also to reduce computational costs, making the process more efficient and resource-effective. Equation 3.1 represents each of the 10-second segments V_c.

$$V_c = \lfloor \frac{V}{10} \rfloor \tag{3.1}$$

We isolate the first frame of each video and apply the Viola-Jones algorithm [Viola and Jones (2001)] to extract the precise facial location within the frame. We assume the dimensions of the Viola Jones bounding box are (h, w), where h is the box's height and w is its width in pixels. We then identify the box's central point (x_0, y_0):

$$(x_0, y_0) = (l + \frac{h}{2}, c + \frac{w}{2}) \tag{3.2}$$

where (l, c) denote the pixel coordinates of the top-right corner of the bounding box.

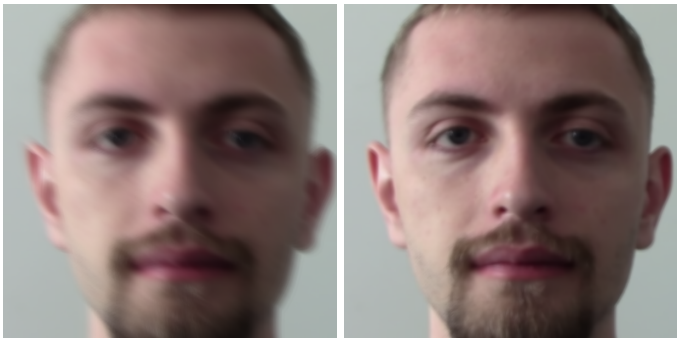
3.2.1.2 Central Point Stabilization

We repeat the process of identifying the central point (x_0, y_0) for each subsequent frame, detecting the face and calculating its central point $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where n is the frame index. For each frame, we define the desired size of the cropped region (h,w), which are the dimensions of the first frame bounding box. Using the central point of each frame, we extract a sub-frame that places the face at the center using the formula:

$$Ic_n(x_n, y_n, z) = Io_n(x_n - \frac{w}{2} : x_n + \frac{w}{2}, y_n - \frac{h}{2} : y_n + \frac{h}{2}, z) \tag{3.3}$$

$Ic_n(x_n, y_n, z)$ represents the cropped frame at index n, Io_n represents the original full frame, **(x_n, y_n)** denote the x and y-coordinates of the central point in frame n, **(w, h)** represent the width and height of the desired crop and z represents each color channel (RGB) (z=3).

To visualize our results, we overlay the video frames with some transparency to highlight the improvement in motion artifacts from non-stabilized to stabilized using central point videos. motion of the videos and the improvement of our methods (Figure 3.1).



(A) No Stabilization Technique (B) Central Point Stabilization

FIGURE 3.1: Overlaid video frames with some transparency to visualize the motion through frames for the original video and using Central Point Stabilization

We evaluate the stabilization using Mean Squared Error (MSE). MSE was calculated between consecutive frames to quantify the level of frame-to-frame variation in the video, as seen in Equation 3.4.

$$MSE_n = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} (I_{n-1}(x,y) - I_n(x,y))^2 \tag{3.4}$$

where $I_{n-1}(x,y)$ represents the pixel value at position (x, y) in the previous frame, $I_n(x,y)$ represents the pixel value in the current frame, w is the width of the frames in pixels, h is the height of the frames in pixels, $\sum_x$ denotes the summation 1 to width, $\sum_y$ denotes the summation from 1 to height. After calculating the MSE for each pair of consecutive frames, the overall stability measure can be obtained by averaging the MSE values across the entire video sequence, as seen in Equation 3.5.

$$MSE_{video} = \frac{1}{n} \sum_{1}^{n} MSE_n \tag{3.5}$$

Using the baseline non-stabilized video, results in a MSE of 32.9. Using the Central Point Stabilization approach results in a 32.2% increase in stability, as seen in Table 3.1.

TABLE 3.1: Improvement in motion artifacts using the Central Point Stabilization approach.

Method	MSE	Improvement percentage
Original video	32.9	-
Central Point Stabilization	22.1	32.7%

This method attempts to ensure that the face remains centered in each cropped frame, helping to stabilize the video. However, a significant limitation arises because the Viola-Jones algorithm detects the face independently in each frame, without considering its position in previous frames. This frame-by-frame independence introduces inconsistencies in the bounding box dimensions and location, leading to motion noise in the cropped frames. As a result, while this approach provides some level of stabilization, it does not fully address motion artifacts.

To achieve better stabilization, we explore an alternative approach that tracks the central point of the face continuously throughout the video rather than treating each frame independently, PIPs. This enables temporal consistency by maintaining the face's position relative to previous frames. By tracking the face's motion over time, PIPs significantly reduces the noise introduced by frame-by-frame variations, providing a more stable and reliable output. In the previous chapter, we demonstrated that PIPs outperforms optical flow in handling motion artifacts, which is why we adopt it here. This approach ensures that the face remains consistent across frames, even in the presence of natural motion artifacts.

3.2.1.3 Particle Video Point Trajectories Stabilization

The PIPs algorithm was proven to be efficient and accurate in tracking features through frames. We rely on its accuracy to track the identified central point (x_0, y_0) in the first frame. In other words, instead of repeating the process of identifying the face and its central point for the entirety of the frames in the video, we feed that first frame central point to the PIPs algorithm and we extract it's trajectory throughout the frames of the video. For each frame and position of the feature in that frame, we use 3.3 to place the bounding box around the central point.

We notice a 74.2% improvement of stabilization using the PIPs method compared to the original video. We overlay the video frames with some transparency to visualize the motion of the videos and the improvement of our methods (Figure 3.2).

The results of our experiments demonstrate the improvements in video stabilization achieved by transitioning from the original video to Central Point Stabilization and,

TABLE 3.2: Summary of results for stabilization methods

Method	MSE	Improvement percentage
Original Video	32.9	-
Central Point Stabilization	22.1	32.7%
PIPs Stabilization	8.5	74.2%

(A) No Stabilization Technique (B) Central point Stabilization (C) PIPs Video Stabilization

FIGURE 3.2: Overlaid video frames with some transparency to visualize the motion through frames for each method

finally, to the PIPs-based stabilization method. While the Central Point Stabilization approach reduces motion artifacts to some extent, its reliance on frame-by-frame face detection introduces inconsistencies, limiting its effectiveness.

The PIPs algorithm, in contrast, offers a robust solution by tracking the central point of the face throughout the video, ensuring temporal consistency and significantly reducing motion noise. With a 74.2% improvement in stability compared to the original video, the PIPs method clearly outperforms both the non-stabilized and Central Point Stabilization approaches.

3.2.2 Enhancing Visibility with Eulerian Video Magnification

Building upon the stabilization techniques discussed earlier, the next step in our approach focuses on enhancing the subtle changes occurring on the skin, specifically motion and light variations. These changes are critical for improving the performance of rPPG algorithms, as they are directly tied to the physiological signals we aim to extract. To achieve this, we utilize the Eulerian Video Magnification method [Wu et al. (2012)], a technique for amplifying temporal variations that are otherwise invisible to the naked eye.

This method takes a video as input and enhances its temporal variations by amplifying specific frequency bands. This process involves four steps:

- **Building a Laplacian pyramid:** Each frame is decomposed into multiple frequency bands using a Laplacian pyramid. We use five pyramid levels to balance the capture of fine details and computational efficiency.
- **Applying a band-pass filter:** We apply a band-pass filter to the Gaussian pyramid of the video frames. We use the [0.75, 5] Hz frequency range identified in our signal processing experiments, as this range corresponds to the important physiological frequencies.
- **Amplifying the filtered pyramid:** The filtered frequencies are amplified by a factor a=10. This amplification factor provides visible enhancement, without distorting the video.
- **Reconstructing the video:** The amplified Gaussian pyramid is combined with the original Laplacian pyramid to reconstruct the enhanced video.

In the color amplification approach, the color channels of each level are amplified, in contrast with the motion amplification approach, where the temporal gradients (changes in pixel values over time) are computed and amplified. This dataset was processed to generate three sets of videos: the original stabilized videos, stabilized color amplified videos and stabilized motion amplified videos.

To illustrate the results of the magnification, we provide an example from the O-HR dataset, as its open-source nature permits the inclusion of visual examples (Figure 3.3).

3.2.3 Spatiotemporal Image Generation

With the stabilized videos and enhanced temporal variations from Eulerian Video Magnification, the next step focuses on generating spatiotemporal images. These images encapsulate both spatial and temporal information from the video, providing a robust input representation for ML models. Utilizing spatiotemporal images offers several advantages over analyzing a single continuous video stream. Firstly, it increases the dataset size, as each spatiotemporal image encapsulates a temporal sequence of a single facial region. This is especially beneficial since each facial region consists of slightly distinct features. This spatiotemporal transformation facilitates more robust training of ML models, enhancing their ability to distinguish subtle changes in pulse signals over time. Additionally, the process of stabilizing the images ensures consistent tracking of specific facial areas across the temporal sequence. By maintaining alignment between consecutive frames, the analysis remains focused on the same regions, enabling more

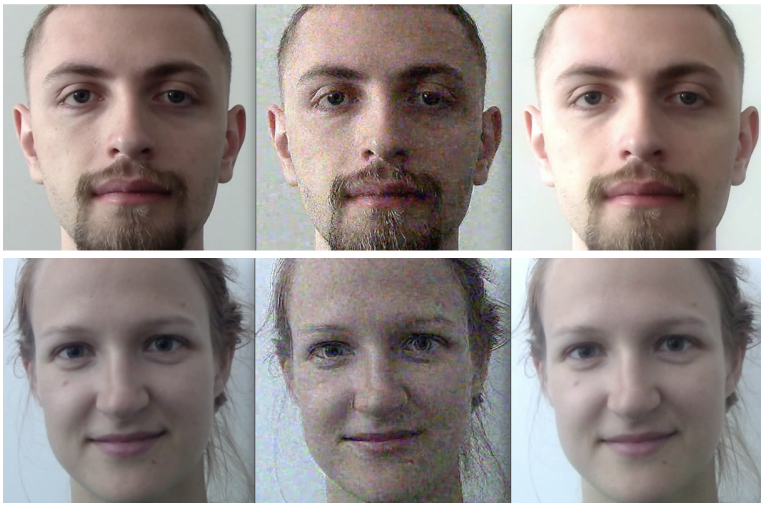


FIGURE 3.3: Example of the Eulerian Video Magnifications results. The first column shows the original stabilized video. The second column presents the motion amplified video, where edges of the face and subtle movements become more prominent. The third column shows the color amplified video, where brighter regions highlight changes in skin color due to blood flow.

precise examination of physiological variations. By transforming video sequences into structured spatial-temporal representations, these images preserve critical signal variations while reducing the computational overhead of continuous video processing.

In order to create spatiotemporal images we employ a technique that involves the division of the stabilized videos into six equal vertical segments. Subsequently, the first and last segments are discarded to exclude any residual background or non-essential facial regions that may not have been adequately eliminated by the pre-processing steps. Then, we segment each remaining frame into L vertical segments of three pixels:

$$L = \lfloor \frac{w}{3} \rfloor \tag{3.6}$$

where w represents the width of each frame in pixels.

Through empirical testing, we find that three-pixel-wide slices provide an optimal balance between spatial resolution and computational efficiency. At the same time, facial features relevant to pulse estimation may exhibit variations on the order of a few pixels. By segmenting the frames into three-pixel-wide segments, we aim to capture these subtle variations more effectively.

We can represent each frame as:

$$Frame^{(k)} = \begin{bmatrix} S_{0,0}^{(k)} & S_{0,1}^{(k)} & \cdots & S_{0,L-1}^{(k)} \\ S_{1,0}^{(k)} & S_{1,1}^{(k)} & \cdots & S_{1,L-1}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_{i,0}^{(k)} & S_{i,1}^{(k)} & \cdots & S_{i,L-1}^{(k)} \end{bmatrix} \tag{3.7}$$

for

$$S_{i,n}^{(k)} = (P_{i,3j}^{(k)}, P_{i,3j+1}^{(k)}, P_{i,3j+2}^{(k)}) \tag{3.8}$$

where k is the k^{th} frame, (i,j) are the height and width of the frame in pixels respectively, P represents the pixel values and S each 3 pixel slice values. For clarity, j refers to the width of the frame in terms of pixel groups or slices and n is the time dimension or different frames in the sequence.

In order to construct the spatiotemporal images, we arrange the corresponding vertical segments from each frame sequentially, frame by frame. The m^{th} spatiotemporal image, generated by the m^{th} vertical slice is represented by:

$$ST_m = \begin{bmatrix} S_{0,m}^{(0)} & S_{0,m}^{(1)} & \cdots & S_{0,m}^{(k)} \\ S_{1,m}^{(0)} & S_{1,m}^{(1)} & \cdots & S_{1,m}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_{i,m}^{(0)} & S_{i,m}^{(1)} & \cdots & S_{i,m}^{(k)} \end{bmatrix} \tag{3.9}$$

These resulting images provide a comprehensive representation of the video's content, with each image capturing a distinct set of three-pixel-wide segments spanning the entire duration of the video (Figure 3.4). It must be noted that the number of images per subject can vary, depending on their approximate location to the camera or their facial size.

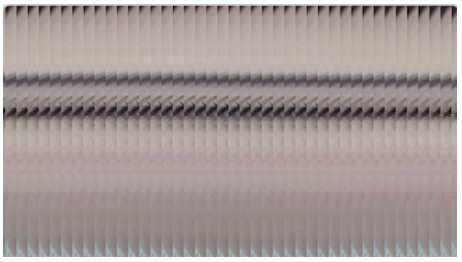
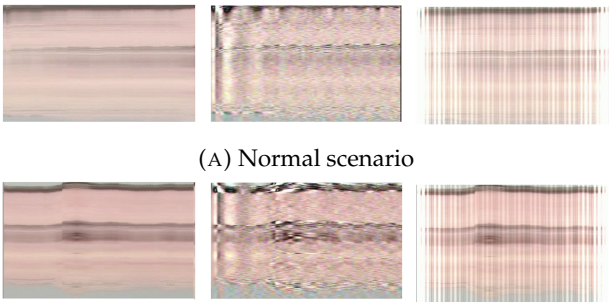


FIGURE 3.4: Example of a spatiotemporal image

In Figure 3.5, we present examples of spatiotemporal images generated without stabilization, highlighting the differences in scenarios with and without physical exercise. While these images are for visualization purposes, stabilized versions of these images are used in our experiments to improve consistency and accuracy.

3.2.4 Pulse Estimation with a Convolutional Neural Network

Let ST_m represent the input spatiotemporal image, with dimensions $w \times h \times z$, where w is the width, h is the height and z is the number of channels - here the channels are 3. The CNN is designed to process the input spatiotemporal image ST_m and predict the



(A) Normal scenario

(B) After physical exercise scenario

FIGURE 3.5: Examples of spatiotemporal images without stabilization. Column 1: No amplification; Column 2: Motion amplification; Column 3: Color amplification. (A) Normal scenario; (B) After physical exercise. Motion amplification in (B) highlights increased movement compared to (A).

number of beats $\hat{y}$. The architecture consists of three convolutional layers, a flattening step and fully connected layers, as summarized in Table 3.3. During the forward pass, the input ST_m is reshaped and passed through each layer of the CNN in sequence, with ReLU activation functions applied after each convolutional and fully connected layer. The optimizer used is the Adam optimizer with a learning rate of 0.001 and the loss function is the L1 loss (MAE) between the predicted number of beats $\hat{y}$ and the ground truth number of beats. A visual representation of the CNN can be found in Figure 3.6.

TABLE 3.3: Parameters of the CNN Architecture

Layer	**Input**	**Parameters/Output**
Conv1	ST_m	Kernel: $K_1 = 3 \times 3$, Stride: $S_1 = 1$, Activation: ReLU, Output: O_1
Conv2	O_1	Kernel: $K_2 = 3 \times 3$, Stride: $S_2 = 3$, Activation: ReLU, Output: O_2
Conv3	O_2	Kernel: $K_3 = 3 \times 3$, Stride: $S_3 = 3$, Activation: ReLU, Output: O_3
Flatten	O_3	Flattened Output: F
Fully Connected	F	Units: $H_1 = 128$, Output: H_1
Output Layer	H_1	Predicted Beats: $\hat{y}$
Optimizer	Adam, Learning Rate: 0.001	
Loss Function	L1 Loss (MAE)	

3.2.5 Second-Stage Learning: Refining Signal Selection

It is apparent that not all spatiotemporal images exhibit similar performance and certain regions within them may contain significant noise. Rather than making the assumption on which areas the CNN finds the most informative based on convention,

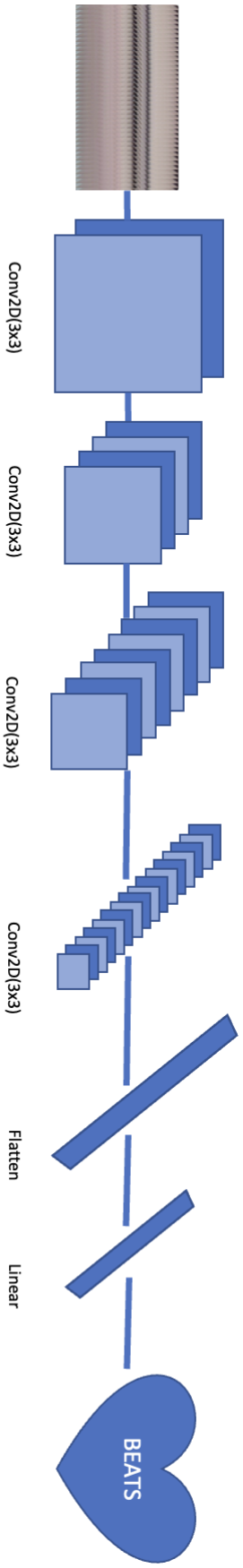


FIGURE 3.6: Illustration of a spatiotemporal image capturing temporal changes in a specific facial region, constructed by sequentially arranging three-pixel-wide slices from consecutive video frames.

we have implemented a second learning stage. Following the CNN's pulse prediction on individual images, we construct a new binary dataset. This dataset is formed based on the MAE between the CNN's predictions and the ground truth on number of beats. Utilizing a predefined threshold, t=0.5, corresponding to a MAE of 3 beats per minute (bpm), we categorize the images into two classes according to whether their MAE surpasses or remains below the threshold. The 3 bpm criterion for categorizing images automatically distinguishes "good" images from "bad" ones. This threshold was chosen as it represents an acceptable margin of error for pulse estimation. A Multi-Layer Perceptron (MLP) is trained to classify the images in the custom binary dataset, ensuring that only the most informative "good" images are utilized for further analysis. This automated selection process eliminates the need for subjective assumptions about image quality, enhancing the robustness of the pipeline. An MLP comprising of 5 layers with 200 neurons per layer, is trained to classify the spatiotemporal images, using the custom "good" and "bad" image dataset as described above. A 10-fold cross-validation experiment is conducted, selecting images that the classifier categorizes as "good" 70% of the time. The evaluation metrics presented in the subsequent section are estimated based on the MLP's predictions for the "good image" class. An overview of the ST2S-rPPG framework can be found in Figure 3.7.

3.3 Evaluation and Performance Analysis

3.3.1 Metrics for Success - How We Measure Performance

We evaluate the performance of ST2S-rPPG using five metric indicators commonly utilized to assess rPPG regression approaches, namely Mean Absolute Error (MAE), Mean Error (ME), Standard Deviation (SD), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as defined the equations below. It is worth noting that some entries in the comparison tables include missing values due to the absence of reported metrics in the referenced literature. Our work provides a comprehensive evaluation across all relevant metrics, ensuring a complete and consistent comparison.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.10}$$

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{3.11}$$

$$\text{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((y_i - \hat{y}_i) - \text{ME}\right)^2} \tag{3.12}$$

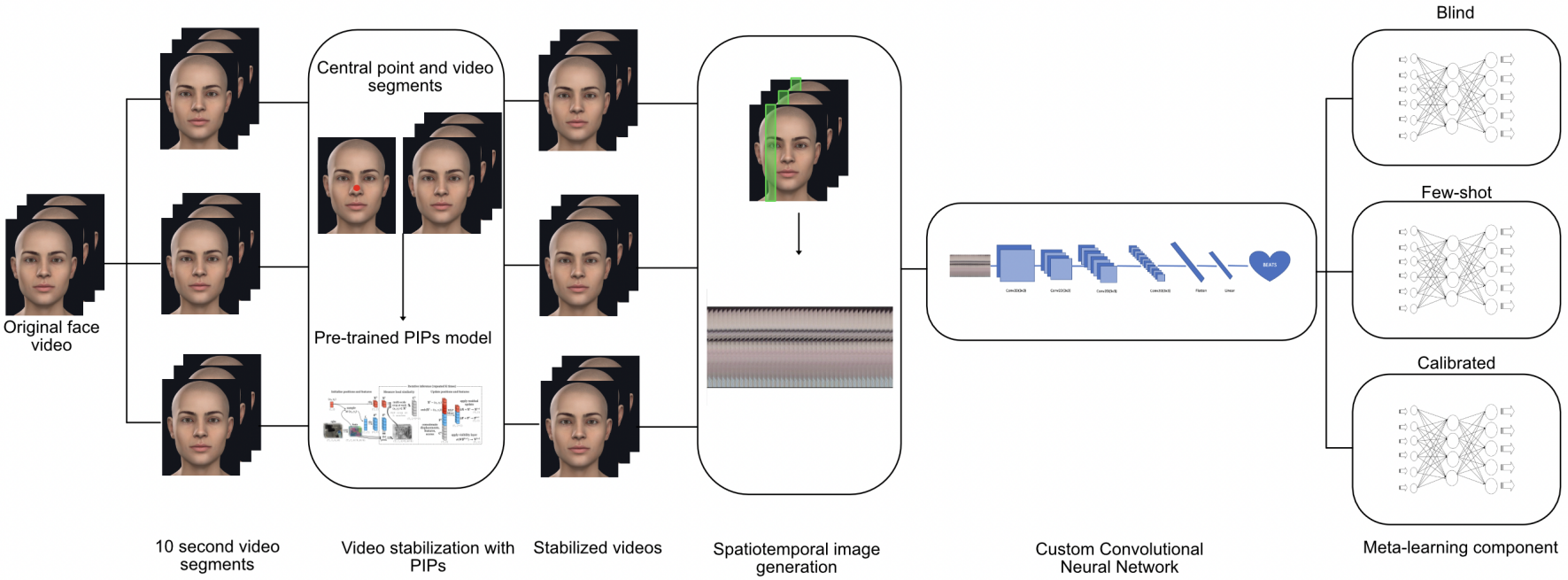


FIGURE 3.7: Overview of ST2S-rPPG for rPPG-based measurement of HR via spatiotemporal images. Key steps include video stabilization using PIPs to reduce motion artifacts, spatiotemporal image generation for capturing combined spatial and temporal features and the second-stage learning process for automated selection of high-quality, informative images. Each stage is designed to optimize pulse estimation accuracy while minimizing computational complexity.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.13}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3.14}$$

Each metric offers a unique perspective on the error between predicted values ($\hat{y}_i$) and ground truth values (y_i), allowing for a deeper understanding of our model's performance.

MAE provides a straightforward and interpretable measure of the model's absolute prediction error, which is why it is widely used to evaluate overall accuracy, offering an intuitive sense of how close predictions are to the true values on average.

ME, on the other hand, measures the mean of the raw differences between predictions and ground truth values. Unlike MAE, ME retains the sign of the differences, which allows it to capture any bias in the model's predictions. For example, a consistently positive or negative ME indicates that the model is systematically overestimating or underestimating, respectively. It's important to mention that while it is provided by some papers, its informativeness can be misleading. In our analysis, we prioritize MAE as the most informative metric, as it directly measures the average error without biasing negative or positive deviations.

SD assesses the variability of the errors, with a lower SD indicating that the model's errors are more consistent and predictable, whereas a higher SD suggests that the errors vary significantly. This metric is particularly useful for understanding the reliability of the model's predictions under varying conditions.

RMSE is similar to MAE but gives greater weight to larger errors due to the squaring of differences. This sensitivity to large errors makes it particularly useful in scenarios where significant deviations from the ground truth are critical. It complements MAE by emphasizing the impact of outliers or large prediction errors.

Finally, MAPE expresses the average error as a percentage of the actual values. This makes it especially valuable for comparing performance across datasets with different scales.

By using these metrics in combination, we can gain a holistic understanding of the model's strengths and weaknesses, guiding further optimization and refinement.

3.3.2 Implementation Details

For both datasets, during the CNN prediction, we implement the Leave-One-Subject-Out (LOSO) method. This is due to the fact that individual variability is significant and LOSO ensures that the model is trained on a wide variety of subjects and tested on a completely independent individual while preventing overfitting to individual characteristics. For the second-stage learning component, we conduct three distinct experiments: **Blind Scenario:** The classifier is withheld samples of the individual it is predicting on (i.e. LOSO), ensuring no data leakage. **Few-Shot Scenario:** The classifier is provided with data from all participants and only 6 random samples from the individual it is predicting on (3 per class). **Calibrated Scenario:** The classifier is trained and tested using data from the same individual to simulate a personalized or user-specific model. To ensure fair evaluation and avoid data leakage, we first balance the dataset (e.g., equal number of samples per class) and then split it into 80% for training and 20% for testing, making sure that no overlapping data points appear in both sets.

3.3.3 Eulerian Video Magnification

Before presenting our results, it is important to address the impact of Eulerian Video Magnification on motion and light amplification. Our experiments revealed that its use for enhancing motion and light variations did not improve performance; in some cases it even led to worse results compared to using the raw stabilized videos. This suggests that the amplification process may have introduced additional noise or distortions that interfered with the pulse estimation.

One possible explanation for this observation is that Eulerian Video Magnification, while effective in amplifying subtle temporal variations, also enhances artifacts such as minor head movements, compression noise and illumination changes that are not relevant to pulse estimation. These additional artifacts may have led to reduced performance. Additionally, the motion amplification process could have exaggerated micro-movements that are already captured adequately in the original frames, causing redundant or misleading information to be introduced.

Another factor to consider is that Eulerian Video Magnification increases computational complexity. The additional processing required for building the Laplacian pyramid, applying band-pass filtering and reconstructing the video raised the computational cost without offering any benefits in terms of performance accuracy.

Based on these findings, we decide to exclude Eulerian Video Magnification from our final methodology. Instead, we rely on the stabilized videos without amplification, as they provide comparable or better results while maintaining efficiency.

3.3.4 Experimenting on MMSE-HR

We compare our proposed method with several state-of-the-art methods, ranging from approaches mitigating motion artifacts [Li et al. (2014), Tulyakov et al. (2016)] to other spatiotemporal approaches [Niu et al. (2019a), Jaiswal and Meenpal (2022)]. To ensure the validity of the comparison, we report on work that has been evaluated on the same dataset. All related results are presented in Table 3.4.

TABLE 3.4: A summary of average HR estimation per video for ST2S-rPPG on the MMSE-HR dataset. Bold numbers indicate best performance and underlined numbers indicate second best performance.

Method	HR_{MAE}	HR_{ME}	HR_{SD}	HR_{RMSE}	HR_{MAPE}
Li et al. (2014)	-	11.56	20.02	19.95	14.64%
SAMC [Tulyakov et al. (2016)]	-	7.61	12.24	11.37	10.84%
RythmNet [Niu et al. (2019a)]	-	-0.85	4.99	**5.03**	**3.67%**
Niu et al. (2019b)	-	-3.10	9.66	10.10	6.61%
Jaiswal and Meenpal (2022)	6.4	-	6.63	6.82	-
ST2S-rPPG - No second-stage learning	10.21	1.59	5.58	11.75	14.94%
ST2S-rPPG - Blind (ours)	<u>5.94</u>	<u>0.65</u>	<u>4.78</u>	7.67	7.66%
ST2S-rPPG - Few-shot (ours)	**5.13**	**-0.39**	**4.11**	<u>6.57</u>	<u>6.16</u>%
ST2S-rPPG - Calibrated (ours)	*2.06*	*-0.23*	*2.35*	*3.11*	*2.88%*

We decide to keep the calibrated results separate from the evaluation as our classifier is trained with personalized data and comparison would not be fair. Our MAE excluding the second-stage learning highlights the challenges of rPPG without the selection of informative data regions. Despite this, the standard deviation of the first-stage learning is favorable compared to literature, which indicates that our approach produces more consistent predictions with lower variability across individuals. Experiments without stabilization were not conducted because inconsistent facial tracking would prevent reliable spatiotemporal image generation. We can observe that ST2S-rPPG achieves promising results on most commonly used metrics.

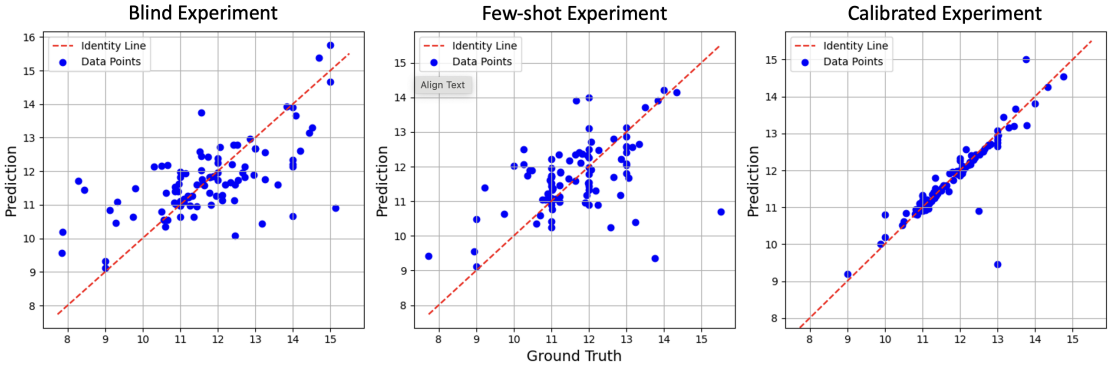


FIGURE 3.8: Scatter plot between ground truth HR and predicted HR for the MMSE-HR dataset.

Specifically, both blind and few-shot two-stage learning approaches achieve the best results in $HR_{MAE}, HR_{ME}, HR_{SD}$. The few shot two-stage learning also achieves second best performance in HR_{RMSE}, HR_{MAPE}. We demonstrate the most significant improvement compared to Li et al. (2014) and Tulyakov et al. (2016). These methods do not use spatiotemporal representations, further proving their efficiency. They also use adaptive band-pass filters for noise reduction, proving our stabilization method's capability. Our advantage over Niu et al. (2019a,b) is that instead of using an aggregate signal from multiple ROIs, we take advantage of all regions of the face, do not aggregate the spatiotemporal signal and do not choose the optimal images (ROIs) empirically. Finally, compared to all spatiotemporal approaches in Table 3.4, ST2S-rPPG does not perform any RGB transformations, since the lighting in the MMSE-HR database is not heterogeneous.

Compared to Jaiswal and Meenpal (2022), the HR_{MAE} error was reduced by 13.64%. At the same time we have achieved the lowest standard deviation, suggesting more consistent predictions across individuals. Our calibrated two-stage learning experiment achieves the best performance across all metrics, keeping in mind that the classifier is trained with personalized data. However the significance of such an approach holds a lot of potential for real-life settings, where personalization can lead to substantial improvement in model performance, leading to better health outcomes.

We also present a modified Bland-Altman plot between the difference of the ground truth and the predicted beat values across the ground truth in Figure 3.9 for all three experiments. The two dashed red lines represent the upper and lower 95% limits of agreement and are calculated as the mean difference plus and minus 1.96 times the standard deviation of the differences. Each data point corresponds to each subject in the dataset. The majority of the points fall within the 95% limits of agreement, indicating accuracy within acceptable clinical limits. We also present the scatter plot between the ground truth HR and predicted HR in Figure 3.8.

FIGURE 3.9: Bland-Altman plot with adjustments for ST2S-rPPG on the MMSE-HR dataset, the black line represents the mean and the red lines the 95% limits of agreement.

Our analysis of model performance across gender groups reveals some unexpected trends. Despite women making up a larger portion of the dataset (23 women, 17 men), their initial performance in pulse estimation was worse than men. As seen in Table 3.5, the original model without the second-stage learning, the women's MAE was 11.47, significantly higher than 8.38 for men. This performance gap gradually decreased with the introduction of second-stage learning methods, with MAE dropping to 6.51 (women) vs. 5.11 (men) in the blind approach, 5.38 vs. 4.79 in few-shot learning and 2.02 vs. 2.14 in the calibrated approach. By the final stage, performance across genders had nearly equalized.

TABLE 3.5: A summary of results across genders for the MMSE-HR dataset.

Method	**Women**	**Men**	**MAE Gap (W-M)**
ST2S-rPPG - No second-stage learning	11.47 ± 5.9	8.38 ± 5.13	3.09
ST2S-rPPG - Blind	6.51 ± 5.78	5.11 ± 3.35	1.4
ST2S-rPPG - Few-shot	5.38 ± 4.36	4.8 ± 3.79	0.58
ST2S-rPPG - Calibrated	2.02 ± 2.32	2.14 ± 2.4	-0.12

This is an interesting observation; although the dataset used in this study was balanced in terms of gender and skin tone distribution the initial results showed a consistent performance gap, with higher error rates for female participants. This was unexpected, as a larger sample size typically improves model generalization, suggesting that dataset size alone was not the determining factor. Several potential explanations may contribute to this discrepancy. We observed that women exhibited more facial movement during recordings which potentially affected the rPPG signal extraction, even though we attempt to minimize that with stabilization. Another factor could be subtle differences in facial blood flow visibility, like variations in skin composition, texture or even makeup application. For the latter we have no way of examining this as it is not mentioned in the dataset characteristics but it is a possibility. Additionally, it is possible that the model struggled with feature representation differences between male and female participants. Without controlled experiments designed to isolate these factors, it is not possible to attribute the performance gap to any single cause with certainty.

Despite this, our second-stage learning significantly improved model performance across both genders, eventually eliminating the gap in our calibrated experiments. This suggests that adaptive learning strategies can successfully mitigate biases, making them an essential component for improving generalization in video based pulse estimation.

3.3.5 Experimenting on UBFC-rPPG

We compare ST2S-rPPG to several state-of-the-art approaches that have been evaluated on the UBFC-rPPG database.

TABLE 3.6: A summary of average HR estimation per video for ST2S-rPPG on the UBFC-rPPG dataset. Bold numbers indicate best performance and underlined numbers indicate second best performance

Method	HR_{MAE}	HR_{ME}	HR_{SD}	HR_{RMSE}	HR_{MAPE}
ICA[Poh et al. (2010a)]	8.43	-	18.6	18.8	-
CHROM [Wang et al. (2016)]	10.6	6.78	19.1	20.3	-
3D CNN [Bousefsaf et al. (2019)]	5.45	-1.31	8.55	8.64	-
Meta-rPPG [Lee et al. (2020)]	5.97	-	7.12	7.42	-
TransPhys [Shao et al. (2023)]	**4.66**	-	7.22	7.36	-
ST2S-rPPG - No second-stage learning	8.51	-1.93	4.75	9.84	8.25%
ST2S-rPPG - Blind (ours)	5.62	<u>0.04</u>	<u>4.76</u>	<u>7.24</u>	<u>5.6%</u>
ST2S-rPPG - Few-shot (ours)	<u>5.24</u>	**-0.02**	**3.81**	**6.36**	**5.21%**
ST2S-rPPG - Calibrated (ours)	*3.05*	*-1.04*	*2.82*	*3.98*	*2.95%*

In Table 3.6, we observe a similar trend with Table 3.4 regarding our results without the second-stage learning. Our ST2S-rPPG blind and few-shot method achieves the best results most reported metrics ($HR_{ME}, HR_{SD}, HR_{RMSE}, HR_{MAPE}$) and the second best results in HR_{MAE}, demonstrating its efficiency in accurately estimating heart rate even with limited training data. Additionally, ST2S-rPPG exhibits improvements in HR_{SD}, indicating more precise predictions and reduced variability in heart rate estimations. Similarly to the previous database, we demonstrate the most significant improvement of performance against non spatiotemporal traditional approaches [Poh et al. (2010a); Wang et al. (2016)] and 3D CNN approaches [Bousefsaf et al. (2019)]. TransPhys [Shao et al. (2023)] seems to be performing best in the HR_{MAE} metric, suggesting that spatiotemporal transformers show promising results, but can be computationally expensive. Finally, Meta-rPPG [Lee et al. (2020)], showcases slightly lower estimation accuracy, potentially indicating that a second stage learning component, trained on predictions captures more valuable information for estimation.

In Figure 3.10, we also present the modified Bland-Altman plots for this database and in Figure 3.11 the scatter plots between the ground truth and the predicted HR.

The UBFC-rPPG dataset presents an interesting contrast to the MMSE-HR dataset, as 80% of participants in this dataset are male, creating an imbalance in gender representation. Despite this skewed distribution, a similar initial performance gap is observed, with higher error rates for women in the early stages of the model. However, as with MMSE-HR, our second-stage learning approach reduces this gap, ultimately leading to almost equivalent performance in the calibrated approach.

A general observation across the selected spatiotemporal slices suggests that regions from the central facial area and the sides of the face tend to be more frequently chosen, whereas slices that include the eyes are rarely selected. This aligns with prior findings in the rPPG literature, where areas with stable illumination and strong blood perfusion typically yield stronger pulse signals. The lower selection rate of slices that contain

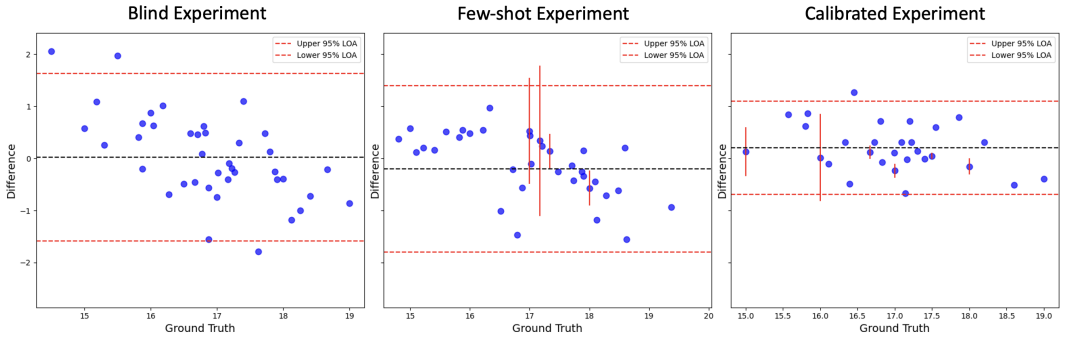


FIGURE 3.10: Bland-Altman plot with adjustments for ST2S-rPPG on the UBFC-rPPG dataset, the black line represents the mean and the red lines the 95% limits of agreement

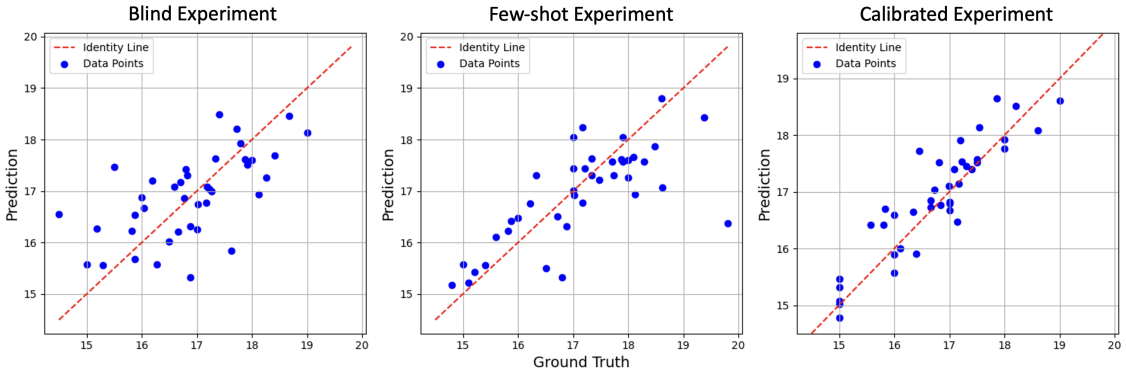


FIGURE 3.11: Scatter plot between ground truth HR and predicted HR for the UBFC-rPPG dataset

TABLE 3.7: A summary of results across genders for the UBFC-rPPG dataset.

Method	Women	Men	MAE Gap (W-M)
ST2S-rPPG - No second-stage learning	9.95 ± 4.93	8.14 ± 4.71	1.81
ST2S-rPPG - Blind	6.41 ± 5.32	5.42 ± 4.62	0.99
ST2S-rPPG - Few-shot	6.52 ± 5.03	4.86 ± 3.42	1.66
ST2S-rPPG - Calibrated	2.4 ± 2.54	3.04 ± 2.87	-0.64

the eyes may be due to increased motion artifacts from blinking or the lower signal intensity in these regions. Although not explicitly analyzed, this trend suggests that future work could refine spatial feature selection for improved model robustness.

3.4 Challenges, Insights & Future Directions

The rapid growth of telehealth, accelerated by the COVID-19 pandemic, has driven the adoption of remote physiological monitoring methods. While DL techniques dominate the field, their complexity often limits interpretability and accessibility, particularly

in clinical applications where trust in model decisions is crucial. This chapter introduces a spatiotemporal approach that balances accuracy, computational efficiency and interpretability. By stabilizing videos using PIPs and leveraging spatiotemporal representations, we mitigate motion artifacts and enhance signal consistency. The second-stage learning framework further refines pulse estimation by automatically selecting the most informative regions, improving overall performance across diverse demographics.

Our evaluation demonstrates that ST2S-rPPG achieves competitive performance compared to state-of-the-art methods, particularly in reducing MAE and improving prediction consistency. The incorporation of second-stage learning significantly enhances accuracy by filtering out low-quality spatiotemporal images, particularly in datasets with diverse participant demographics. While our blind and few-shot learning scenarios significantly improved pulse estimation accuracy, our fully calibrated experiments demonstrated the strongest performance across all metrics. By training the model with subject-specific data, we observed a substantial reduction in error, nearly eliminating demographic-based performance discrepancies. This suggests that personalized calibration can be a powerful tool for improving rPPG accuracy in real-world applications, particularly in healthcare settings where individual physiological differences must be accounted for.

One notable challenge observed in our experiments is the disparity in performance across genders, particularly in MMSE-HR. Initially, female participants exhibited higher error rates compared to male participants, despite a balanced dataset composition. This difference persisted across multiple trials, suggesting that motion artifacts, physiological factors and potential dataset biases (e.g., makeup, facial expressions, movement patterns) could contribute to the gap. However, our second-stage learning approach mitigated these discrepancies, nearly equalizing performance across genders in the calibrated setting. This highlights the importance of adaptive learning strategies in reducing demographic biases, reinforcing the need for more inclusive datasets and model evaluation across diverse populations.

Our experiments highlight discrepancies in performance differences across skin tones. Participants with darker skin tones tended to exhibit higher initial error rates compared to participants with lighter skin tones. This highlights an ongoing challenge in rPPG research, as the optical properties of melanin can affect light absorption and reflection, complicating pulse signal extraction. However, it is important to note that the number of participants with darker skin tones in our evaluation was relatively small and thus these observations should be interpreted with caution. More extensive validation on larger and more diverse datasets is necessary to confirm these trends. However, the second-stage learning framework again mitigated much of the performance gap, reducing discrepancies across skin tones similarly to how it reduced gender-based performance differences. This suggests that adaptive, quality-aware learning strategies

can play a crucial role in addressing demographic biases in rPPG systems, although further work is needed to ensure fairness and robustness across diverse populations.

Another key observation is that spatiotemporal transformations provide a robust feature representation for pulse estimation, but pre-defined ROI selection can limit their effectiveness. Our work moves beyond this constraint by allowing data-driven selection of regions that contribute the most to accurate predictions. However, a trade-off exists between fine-grained spatial segmentation and computational efficiency, as increasing resolution and spatial detail can introduce redundant or noisy information. Finally, our experiments with Eulerian Video Magnification revealed that amplification of motion and color variations did not improve model performance. While Eulerian Video Magnification has been useful in prior work for enhancing subtle skin tone fluctuations, our results indicate that it also amplifies irrelevant artifacts such as head movements and environmental lighting changes, leading to noisier predictions. Given the increased computational cost of Eulerian Video Magnification without measurable benefits, we excluded it from our final methodology.

Despite its strengths, ST2S-rPPG is not without limitations. While our second-stage learning improves accuracy by selecting high-quality images, it introduces an additional classification step, which may add latency in real-time applications. Optimizing this process for real-time inference remains an area for future work. Another limitation is the restricted dataset availability. Although we evaluated our method on MMSE-HR and UBFC-rPPG, our inability to access larger datasets such as PURE and VIPL-HR limits broader generalization. Future work should include testing on more datasets and real-world conditions, particularly in low-light environments or cases with occlusions (e.g., glasses, masks). Additionally, our gender analysis revealed disparities in initial model performance, which suggests the need for further investigation into demographic-specific biases in rPPG datasets.

A critical takeaway from this research is that rPPG performance is highly dependent on video quality. While our stabilization and second-stage learning techniques mitigate some sources of error, factors such as resolution, frame rate, compression artifacts and lighting conditions still impact pulse estimation accuracy. These factors vary significantly across datasets and real-world applications, yet standardized evaluation metrics for video quality in rPPG are almost entirely lacking. Current research often evaluates rPPG models on individual datasets without accounting for how video quality variations directly impact pulse estimation performance. In the next chapter, we shift our focus to quantifying video quality and establishing video evaluation metrics. By systematically analyzing the impact of different video characteristics, we aim to develop a video quality assessment framework that correlates with rPPG accuracy, providing a structured approach for improving dataset selection, preprocessing strategies and model robustness.

Chapter 4

Output Layer: Assessing Video Quality and Developing Metrics

This chapter serves as the output layer of the model, where the insights gained from analyzing video quality are transformed into structured, quantitative metrics.

As we established in previous chapters, rPPG's performance is highly dependent on the quality of the input video. Both signal processing and ML methods, while effective in controlled environments, are vulnerable to videos affected by motion artifacts, poor illumination, significant occlusions or other quality issues. This leads to reduced performance and reliability in rPPG models.

The dependence on video quality creates a bottleneck for using rPPG systems in real-world applications. Traditional video quality metrics, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), have been widely used to evaluate video quality in video compression, streaming and general computer vision tasks. These metrics, however, measure compression and perceptual quality but ignore rPPG-specific distortions, such as motion or uneven illumination that disrupt blood volume signal extraction. Since pulse estimation depends on skin reflectance, traditional metrics fail to indicate rPPG suitability reliably.

To this day, very few, if any, researchers have focused on creating tailored video quality metrics for rPPG. Existing studies often evaluate individual factors, such as motion or resolution, similar to what we did in previous chapters. The lack of a comprehensive

approach, has left a gap in the field, which makes it challenging to develop robust rPPG systems that can generalize effectively to diverse real-world conditions.

A dedicated video quality metric for pulse estimation would offer a standardized and objective way to assess video suitability in real-world scenarios. By aggregating multiple video quality factors (such as motion artifacts, illumination and resolution), such a metric could quantify how these elements influence rPPG performance. This would allow researchers to better understand dataset limitations, assess whether additional data collection is needed for diversity and ensure more generalizable models. In practical applications, it could help clinicians, researchers or individuals identify quality issues in their recordings, offering insights into how video conditions may impact measurement accuracy. By establishing a standardized benchmark, this metric would enhance reproducibility and provide a structured way to evaluate and compare video datasets across different studies and applications.

Chapter Contributions:
Building on this motivation, this chapter presents a framework for video quality evaluation in rPPG. The insights gained here form the foundation for developing metrics specifically tailored to rPPG applications. By bridging the gap between video quality and rPPG performance, we aim to support the creation of more accurate and inclusive rPPG systems. More specifically, our contributions are as follows:

- We systematically analyze how different video quality factors affect rPPG models and quantify their impact on pulse estimation accuracy. Our findings reveal that motion artifacts, occlusions, and resolution drops significantly degrade performance, while color space variations have a lesser effect.
- We introduce the first integrated metrics for assessing video quality tailored to rPPG. By combining multiple degradation factors into a structured score, we enable more reliable dataset selection and preprocessing.
- We validate our metrics by analyzing their correlation with rPPG model performance. Our results indicate that these metrics can serve as reliable predictors of performance degradation caused by video quality issues.

The remainder of this chapter is organized as follows:

4.1 Breaking Down the Noise: How Video Quality Shapes rPPG

To evaluate and quantify the effects of video quality on rPPG performance, we modify videos based on ten quality factors that commonly impact or could impact performance. This approach is necessary because most datasets are collected under optimal

conditions, such as controlled lab environments, and often lack sufficient variability. While some factors have been previously studied, leading to efforts in developing more robust models, others remain largely unexplored. In the following sections, we describe our experimental setup, including the models selected, the datasets used and the quality-based edits applied to the videos. We then present a comprehensive analysis of how these factors influence model performance. This analysis serves as the foundation for developing the proposed video quality metrics.

4.1.1 Experimental Setup

4.1.1.1 Models

For our study, we select four benchmark rPPG models: Plane-Orthogonal-to-Skin (POS), ICA, DeepPhys and Temporal Shift Convolutional Attention Network (TSCAN). These models represent both traditional signal processing approaches and more recent DL based frameworks, which allows us to explore a broad spectrum of methods commonly used in rPPG. POS and ICA are established signal processing techniques known for their robustness. They are widely used in rPPG studies due to their simplicity and effectiveness. DeepPhys and TSCAN are DL models that have demonstrated superior performance and can handle complex, noisy video inputs. For the implementation of these models, we use the rPPG-Toolbox [Liu et al. (2024)], which provides pre-implemented versions of several state-of-the-art rPPG algorithms and pre-trained models. The pre-trained DeepPhys and TSCAN models were trained on the PURE dataset [Stricker et al. (2014)]. Our work builds directly on this framework to ensure consistency with existing methods. Below, we provide a detailed breakdown of their design.

POS

The POS algorithm [Wang et al. (2016)] is a signal processing technique that leverages subtle color changes in the skin caused by blood flow. The algorithm identifies ROIs on the face, typically areas with consistent lighting and minimal occlusion, such as the cheeks or forehead to extract rPPG signals. It then projects the normalized RGB signals onto a plane orthogonal to the direction of the skin tone. This projection enhances the pulse signals while minimizing noise from lighting conditions or skin tone variations. The projected signals are filtered to isolate frequencies corresponding to the typical heart rate range and the dominant frequency within the filtered signal is extracted as the estimated pulse rate. In literature, POS is valued for its simplicity, computational efficiency and robustness under moderate noise conditions. However, it struggles with significant motion artifacts or lighting variations.

ICA

ICA [Poh et al. (2010a)] is another signal processing-based technique widely used in rPPG research. It is a blind source separation method, meaning it isolates and extracts independent source signals from a mixture without prior knowledge of the source characteristics or how they were combined. Similar to POS, ICA extracts signals (treated as mixtures of independent sources, including the pulse signal) from selected ROIs. These mixed signals are separated into independent components by maximizing their statistical independence using mathematical criteria like kurtosis (how ”peaked” or ”flat” a distribution is) or negentropy (a measure of how far a distribution is from being purely random or Gaussian). Finally, a frequency analysis is performed on the independent components to identify the one with dominant pulsatile frequencies corresponding to heart rate. ICA is effective in separating physiological signals from noise and is robust under conditions with minimal motion. However, it requires careful tuning and preprocessing to work effectively.

DeepPhys

DeepPhys [Chen and McDuff (2018)] represents a DL approach that utilizes a CNN and attention mechanisms to improve robustness under challenging conditions. The input video frames are preprocessed to extract ROIs and to create a spatiotemporal representation by stacking frames to form a 3D input. The attention mechanism focuses on facial regions with strong pulsatile signals, such as the cheeks or forehead, while ignoring noisy or occluded areas. The CNN processes the input representation, learning features that correspond to the rPPG signal by identifying patterns in temporal color variations. The output of DeepPhys is the predicted heart rate after mapping the learned features to a time-series signal, which is then analyzed for the dominant frequency. DeepPhys is more robust in scenarios with complex noise and motion but is computationally intensive and requires large datasets for training.

TSCAN

TSCAN [Liu et al. (2020)] builds on the foundation of DeepPhys but introduces temporal shift modules to process temporal information without significantly increasing computational complexity. These modules shift small amounts of information across adjacent frames, enabling the model to handle dynamic inputs like motion and illumination changes effectively, by helping it understand changes over time better. The network has two branches, a motion branch that focuses on capturing motion patterns caused by blood flow and an appearance branch that analyzes spatial features, such as skin texture and color changes. Similar to DeepPhys, TSCAN includes an attention mechanism to prioritize regions with clear rPPG signals, reducing the influence of noisy or occluded areas. The convolutional layers in TSCAN process the output from the temporal shift and attention modules to extract spatiotemporal features. Finally, the network predicts the heart rate by analyzing the spatiotemporal features and

identifying the dominant frequency. TSCAN is designed for real time applications and exhibits improved performance under challenging conditions like motion and illumination changes. However, it does require high-quality and large amounts of data for optimal performance.

4.1.1.2 Datasets

The rPPG-Toolbox [Liu et al. (2024)] was originally configured to work only with UBFC-rPPG. Given that editing and analyzing videos for quality assessment is both time and computationally intensive, UBFC-rPPG was selected due to its well-controlled conditions, allowing us to systematically introduce distortions and analyze their effects. Additionally, its compatibility with the rPPG-Toolbox streamlined implementation. While datasets like MMSE-HR contain more naturalistic variations, they introduce uncontrolled factors that complicate targeted evaluation of specific quality distortions. Details on the UBFC-rPPG specifications can be found in chapter 3.

4.1.1.3 Simulating Real-World Challenges

We design a series of experiments focusing on 10 quality factors. These selected quality factors represent common degradations encountered in real-world rPPG applications, including telehealth and remote monitoring. Motion artifacts, occlusions and illumination changes are particularly critical since they alter the skin's appearance, impacting pulse estimation accuracy. Compression and resolution reductions, while commonly studied in general video quality, have not been systematically assessed in rPPG.

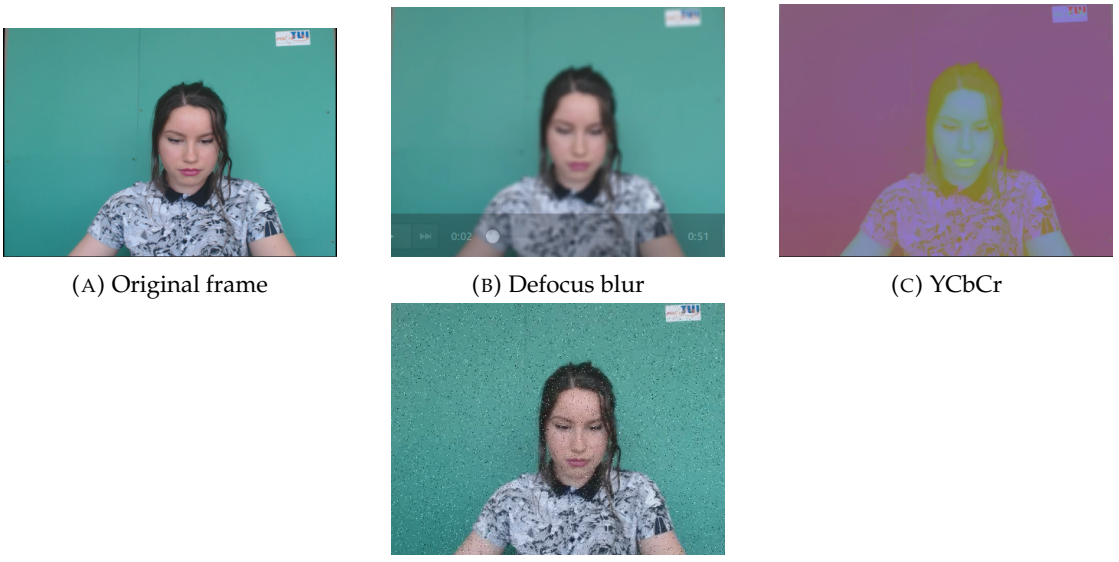
A detailed view of the experimental categories, conditions and their purpose can be found in Table 4.1.

Blur is simulated using defocus and Gaussian blur, both of which reduce fine details critical for capturing subtle skin color variations. Compression (H.264) introduces block artifacts and detail loss, which can affect pulse estimation, especially in telehealth applications where videos are often compressed for storage and streaming. Duration and FPS reductions test the model's ability to function with limited temporal data, simulating conditions where short recordings or low frame rates restrict the number of cardiac cycles available for analysis. Illumination variations explore the effects of dim lighting, fluctuating brightness and extreme contrast shifts, which can alter skin reflectance and impact signal extraction. Motion artifacts are introduced through artificial camera shake, replicating unstable setups or subject movement, which disrupts pixel-based pulse tracking. Noise is added in the form of Gaussian noise (mimicking sensor imperfections) and salt-and-pepper noise (simulating transmission errors or hardware faults), both of which degrade visual clarity and challenge the model's ability

TABLE 4.1: Experimental Conditions and Their Purposes

Experiment Category	Conditions	Purpose
Blur	Defocus Blur (kernel sizes: 3, 5, 9) Gaussian Blur (kernel sizes: 3, 5, 11)	Simulate out of focus and low quality lens effects
Compression	H.264	Evaluate impact of common video compression
Duration	Half Quarter	Assess performance with shorter video clips
FPS	Half Quarter	Test lower frame rates on temporal resolution
Illumination	Constant illumination and brightness (50%, 25%) Fluctuating illumination and brightness every 10 seconds (50%, 25%) Varying brightness (50%, 25%, 150%) Brightness variation in 10-second intervals (50%, 25%, 150%) Contrast (50%, 25%, 150%, 200%, 400%) Contrast variation in 10-second intervals (50%, 25%, 150%, 200%, 400%)	Simulate various lighting conditions
Motion	Artificial Shake (2, 5, 10 pixels)	Replicate unstable camera
Noise	Gaussian Noise (0.02, 0.05, 0.005) Salt-and-Pepper Noise (density: 0.1-0.3, 0.01-0.5, 0.05-0.7)	Test resilience to sensor and transmission noise
Occlusion	Cheeks Forehead Both	Evaluate impact of facial occlusions (hair/facial hair, hands, face mask)
Resolution	Half Quarter	Assess impact of lower resolution
Color Space	Hue, Saturation, Value (HSV) LAB YCbCr YUV Hue adjustments (5, 25, 50) Saturation adjustments (50%, 150%)	Test impact of color space and skin tone variations

to extract meaningful signals. Occlusion tests the effect of covering key facial regions, such as the cheeks or forehead, replicating real-world scenarios like facial hair, masks or hands obstructing the face. Resolution reductions simulate low-quality cameras or image resizing, limiting spatial detail available for pulse detection. Finally, color space transformations alter the way color information is represented in the video, including adjustments to hue and saturation, which test model sensitivity to variations in skin tone appearance due to lighting or camera settings. Together, these modifications allow us to systematically assess how different video quality factors influence rPPG performance. In Figure 4.1, we present visual examples of how these edits affect the video frames.



(A) Original frame (B) Defocus blur (C) YCbCr

(D) Salt & Pepper noise

FIGURE 4.1: A sample frame of various experimental conditions for one participant.

4.1.2 Analysis

To better interpret the impact of different video quality factors, we categorize the results into four groups: spatial degradations, temporal degradations, illumination and color distortions and motion and occlusions. Each category represents a distinct challenge to rPPG model performance, affecting signal extraction in different ways, whether through loss of spatial detail, disruption of temporal dynamics, alteration of skin reflectance or interference with facial visibility. The following sections analyze these effects in detail. The performance of each model was evaluated using the MAE between the predicted heart rate and the ground truth, with unedited videos serving as the baseline.

4.1.2.1 The Impact of Spatial Degradations

Blur, noise, compression and resolution all impact the spatial quality of the video, directly affecting rPPG model performance. Blur (defocus and Gaussian) degrades fine details, reducing the visibility of subtle color variations. Defocus blur causes a steady increase in error across all models as the blur intensity increases, with DeepPhys showing the highest sensitivity. Once details are sufficiently degraded, the performance drop plateaus, suggesting that models operate with insufficient spatial information at extreme blur levels. Gaussian blur, however, has a more variable impact; POS and TSCAN demonstrate relative stability, while DeepPhys and ICA suffer sharper performance declines. Gaussian blur spreads intensity changes more smoothly across the frame, which may explain why POS and TSCAN, which leverage broader periodic patterns, handle it better than defocus blur, which creates a more uniform loss of detail.

Noise (Gaussian and salt-and-pepper) introduces pixel-level distortions, impacting models differently. Gaussian noise adds random variations in pixel intensity, creating a gradual degradation in performance as noise variance increases. DeepPhys and ICA are particularly vulnerable, as their reliance on fine spatial details makes them sensitive to subtle pixel perturbations. Salt-and-pepper noise, on the other hand, introduces extreme pixel outliers, resulting in sharper performance declines. POS remains the most resilient overall, likely due to its reliance on periodic signals rather than precise spatial textures, making it more tolerant to pixel-level distortions.

Compression (H.264) introduces blocking artifacts, which disrupt spatial continuity and fine-grained details. DeepPhys and ICA suffer the most again, as compression eliminates subtle pixel-level patterns essential for their feature extraction. The error increases more abruptly compared to noise, as compression artifacts disproportionately affect facial regions where rPPG models extract signals from. POS and TSCAN, while affected, show greater robustness, likely because they rely more on broader feature patterns that remain somewhat preserved under compression.

Resolution reductions lead to increasing error across all models, as lower spatial detail makes it harder to track skin color variations. DeepPhys and TSCAN experience the sharpest declines, emphasizing their reliance on high-resolution input. ICA and POS, while also impacted, degrade more gradually, benefiting from broader statistical feature extraction rather than pixel-specific details. At extreme resolution reductions, all models show severe performance loss, indicating a threshold where insufficient spatial information renders pulse extraction ineffective.

(A) Defocus blur

(B) Gaussian blur

(C) Gaussian noise

(D) Salt & pepper noise

(E) Compression artifacts

(F) Resolution reduction

FIGURE 4.2: Impact of spatial degradations (blur, noise, compression, resolution) on the performance of four models: DeepPhys, ICA, POS and TSCAN.

4.1.2.2 Temporal Degradations Insights

Reducing FPS and video duration directly impacts temporal signal stability, limiting the number of cardiac cycles available for analysis. Lower FPS reduces the granularity of temporal changes in skin tone, making it harder for models to track pulse fluctuations accurately. Shorter video durations provide fewer overall pulse cycles for the models to learn from, potentially leading to greater sensitivity to noise and short-term variations.

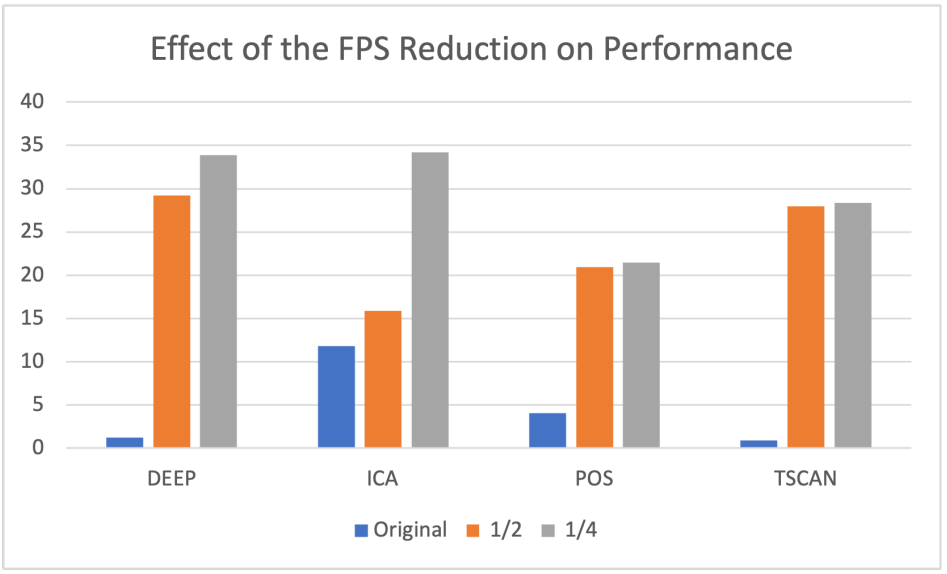
The effect of FPS reduction varies across models, with DeepPhys and TSCAN suffering the most significant performance declines. At half the original FPS, DeepPhys' error increases sharply, highlighting its heavy reliance on fine temporal resolution for feature extraction. TSCAN follows a similar pattern, showing progressive performance loss as FPS decreases, indicating that its feature extraction also depends on a high frame rate. ICA and POS degrade more gradually, suggesting they rely more on broader periodic signals rather than fine-grained temporal details. POS, in particular, remains the most stable, reinforcing its adaptability to different frame rates.

Reducing video duration has a similar effect, but its impact is more uniform across models. All models show increased error rates as the duration is halved and quartered, as fewer frames mean fewer cardiac cycles are available for pulse estimation. DeepPhys is particularly vulnerable, as it depends on longer time-series data to stabilize its predictions. POS and ICA show better resilience, likely because they leverage periodic information from shorter segments more effectively.

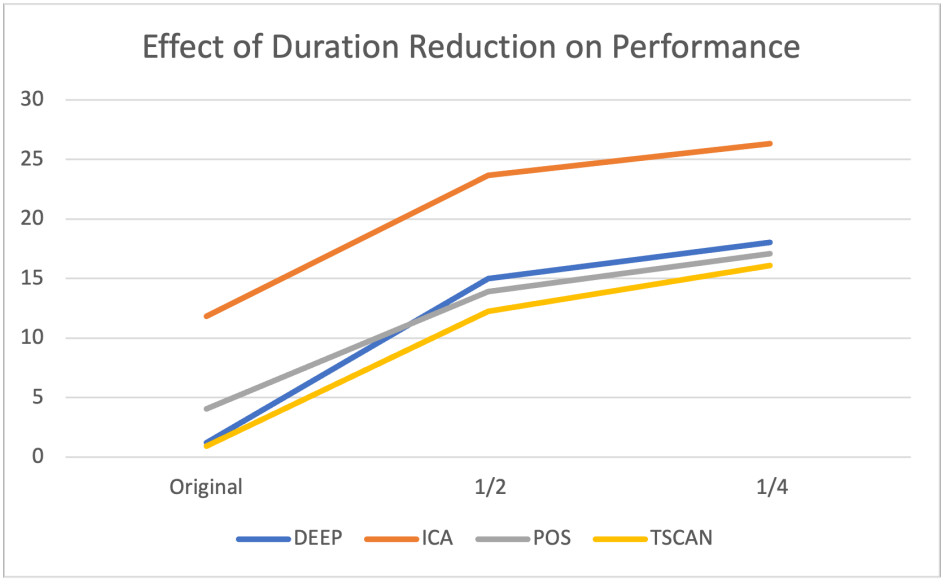
Overall, these results highlight the trade-offs in model architectures. DL-based models (DeepPhys, TSCAN) require high frame rates and longer durations to capture fine temporal variations, making them highly vulnerable to temporal degradation. signal processing models (POS, ICA) handle these conditions better, particularly POS, which remains the most stable model in both FPS and duration reductions.

4.1.2.3 The Role of Lighting & Color

Changes in brightness and contrast significantly alter skin tone representation, impacting rPPG signal extraction. As brightness decreases, models must compensate for reduced contrast between skin regions, while extreme contrast shifts can distort reflectance properties. DeepPhys and TSCAN are highly sensitive to these variations, with error rates increasing sharply under extreme contrast shifts. POS performs moderately well under brightness changes but struggles with overly high contrast. ICA remains the most stable, suggesting that it relies more on statistical features rather than precise skin tone variations, keeping in mind though that it starts from a higher baseline error.



(A) Impact of FPS reduction on model performance.



(B) Impact of reduced video duration on model performance.

FIGURE 4.3: Impact of temporal degradations (FPS and duration) on the performance of four models: DeepPhys, ICA, POS and TSCAN.

Color space transformations (HSV, LAB, YCbCr, YUV) introduce further variability by altering the way color and luminance are represented. DeepPhys is the most affected, particularly in YUV space, where its performance degrades the most. POS and ICA show moderate resistance, suggesting that they rely on broader intensity-based features rather than fine color variations. TSCAN remains consistently vulnerable, likely because its feature extraction is more dependent on the original RGB color structure.

Hue and saturation adjustments present additional challenges. Increasing hue shifts significantly impacts DeepPhys and POS, reflecting their reliance on fine-grained color details for pulse extraction. TSCAN experiences the steepest rise in error, indicating that it depends heavily on stable hue representation. ICA remains the most resilient overall, showing minimal degradation under both hue and saturation changes, which suggests that it prioritizes general intensity-based signals over color-specific features.

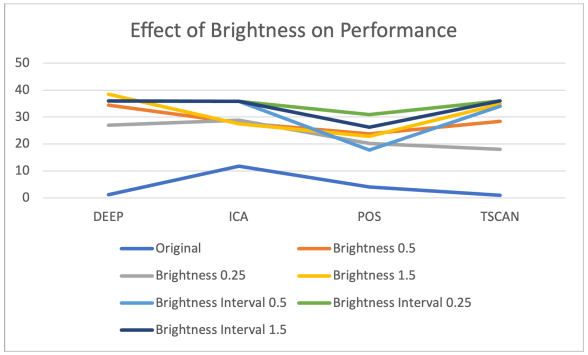
4.1.2.4 Motion & Occlusions

Motion artifacts disrupt rPPG model performance by introducing frame-to-frame inconsistencies, which can cause instability in the extracted pulse signal. DeepPhys and TSCAN experience the most significant performance degradation, as they rely heavily on stable spatial features. With increasing motion intensity, DeepPhys' error rises sharply, reflecting its strong dependence on precise facial feature tracking. TSCAN also deteriorates rapidly, though it shows some stabilization at extreme motion levels, possibly due to temporal averaging.

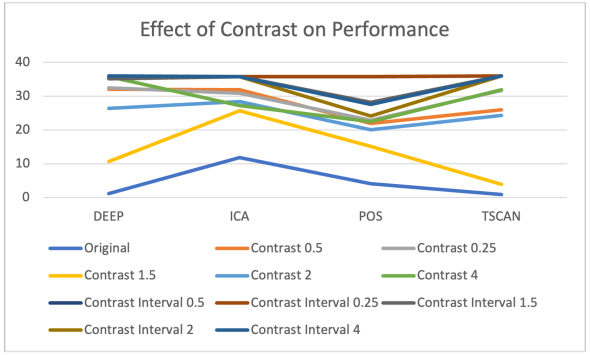
ICA, while also affected, handles motion slightly better because it relies more on broader statistical features rather than precise pixel-level tracking. POS maintains the highest robustness, likely benefiting from its periodic signal processing approach, which makes it less susceptible to motion-induced distortions.

Occlusions introduce another challenge, affecting models differently depending on which facial regions are covered. Cheek occlusions cause the highest performance degradation, followed by combined cheek and forehead occlusions, while forehead-only occlusions have a slightly smaller impact. DeepPhys is the most sensitive, with errors rising steeply under all occlusion types. ICA and POS handle occlusions better, suggesting that they rely on more distributed facial regions for signal extraction, rather than depending on specific areas. TSCAN performs moderately well but struggles when both cheeks and the forehead are occluded.

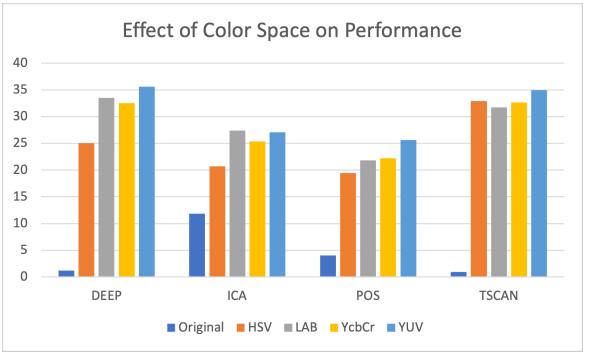
These results highlight a trade-off: DL models like DeepPhys and TSCAN extract fine spatial features, making them more sensitive to disruptions like motion and occlusions. In contrast, signal processing models (POS, ICA) exhibit greater robustness, especially under occlusions.



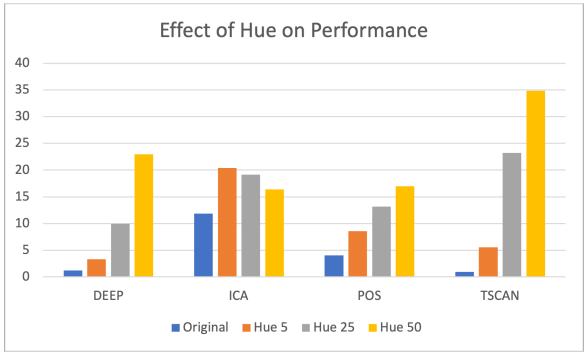
(A) Impact of brightness variations.



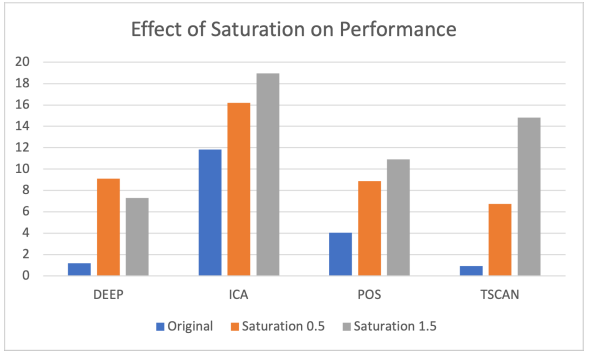
(B) Impact of contrast variations.



(C) Impact of color space transformations.

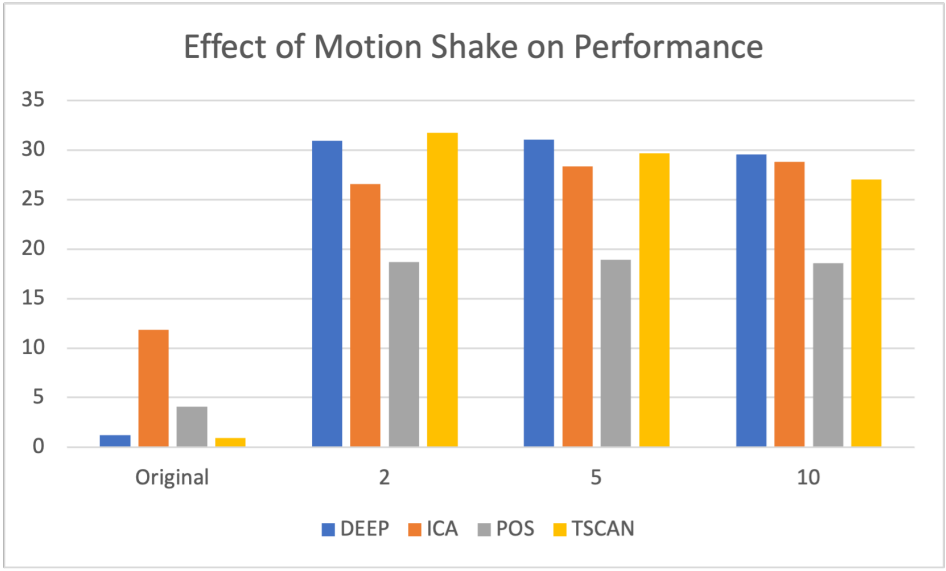


(D) Impact of hue adjustments.

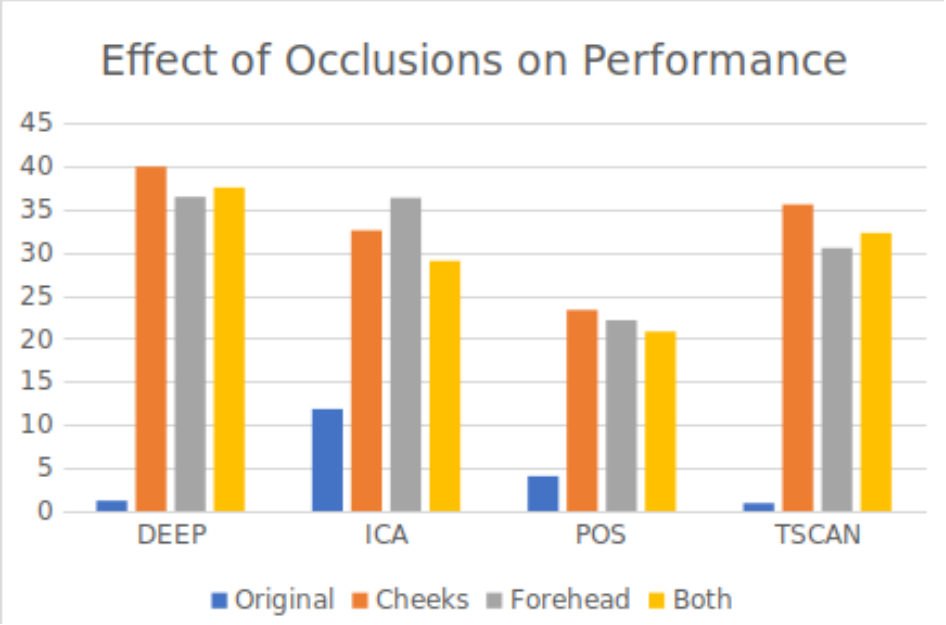


(E) Impact of saturation adjustments.

FIGURE 4.4: Impact of illumination and color distortions (brightness, contrast, color space, hue, saturation) on the performance of four models: DeepPhys, ICA, POS and TSCAN.



(A) Impact of motion artifacts.



(B) Impact of occlusions.

FIGURE 4.5: IImpact of motion and occlusions on the performance of four models: DeepPhys, ICA, POS and TSCAN.

4.1.2.5 Comparing rPPG Models Across Distortions

The evaluation of rPPG models across various video quality conditions highlights key differences in their sensitivity to distortions. Broadly, DL-based models such as DeepPhys and TSCAN exhibit strong dependence on high-quality input data, whereas signal processing-based models like POS and ICA demonstrate greater resilience to degradation. These differences provide important insights into the strengths and limitations of different rPPG approaches and inform their applicability in real-world scenarios where video quality may be suboptimal.

Across all models, motion, resolution, occlusions and illumination changes have the most substantial impact on performance, underscoring the sensitivity of rPPG algorithms to these factors. As seen in Figure 4.6, these distortions consistently increase the MAE across all methods, with DL models showing the most dramatic error spikes. Motion artifacts disrupt temporal stability, occlusions obscure critical skin regions and resolution reductions degrade fine spatial details essential for signal extraction. Illumination variations further challenge the models, as rPPG relies on subtle color fluctuations, which are significantly altered under different lighting conditions. In contrast, changes in color space and shorter video durations have a smaller effect overall, suggesting that these factors are less critical for maintaining rPPG accuracy.

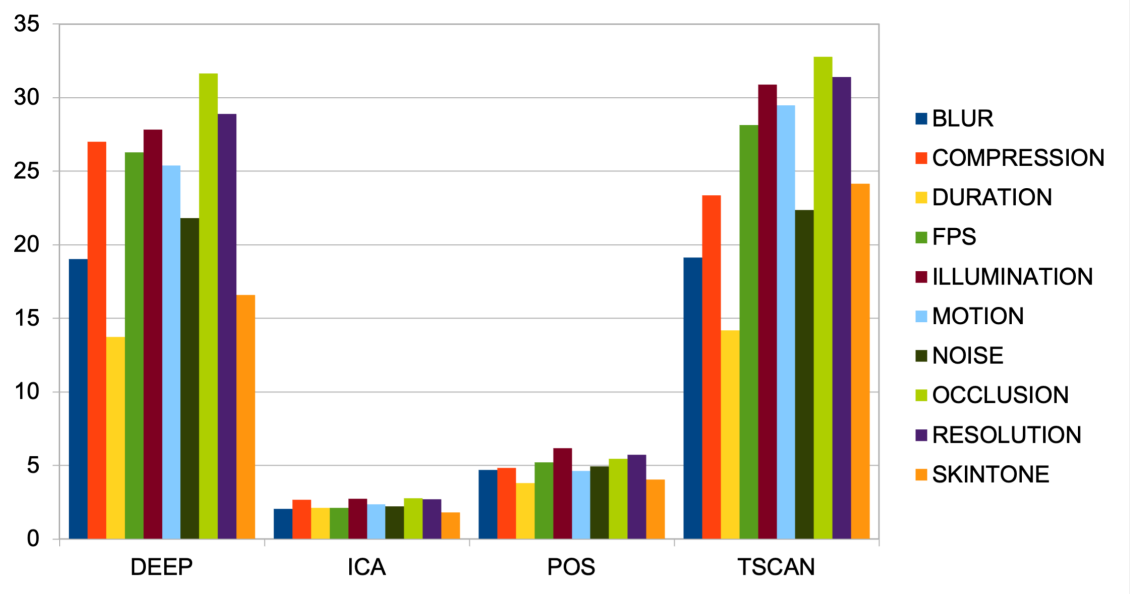


FIGURE 4.6: Bar chart with error increase per quality factor for each model

Among all models, POS achieves the lowest MAE across multiple conditions, including blur, compression, duration, motion and resolution degradation. This suggests that POS can better tolerate both spatial and temporal modifications compared to other models. Unlike DeepPhys and TSCAN, which heavily rely on convolutional feature extraction, POS balances spatial and temporal information without strict dependency on

fine-grained patterns. Its ability to extract periodic signals rather than localized pixel features allows it to maintain a lower error even when video quality is compromised.

	DEEPPHYS	ICA	POS	TSCAN
BLUR	22.85	24.25	18.97	19.15
COMPRESSION	32.41	31.61	19.61	23.37
DURATION	16.51	25.00	15.48	14.17
FPS	31.55	25.04	21.18	28.15
ILLUMINATION	33.42	32.44	24.96	30.89
MOTION	30.49	27.91	18.71	29.47
NOISE	26.20	26.44	19.97	22.37
OCCLUSION	37.98	32.63	22.09	32.78
RESOLUTION	34.68	32.20	23.25	31.41
COLOR SPACE	19.92	21.28	16.40	24.16
ORIGINAL	**1.20**	**11.83**	**4.05**	**0.92**

FIGURE 4.7: MAE values for DeepPhys, ICA, POS and TSCAN across various video quality conditions.

In contrast, DeepPhys and TSCAN exhibit higher MAEs under challenging conditions, particularly when motion, occlusion, illumination or resolution are altered. DeepPhys, in particular, suffers substantial error increases under occlusion and motion artifacts, suggesting strong sensitivity to temporal instability and spatial disruptions. Similarly, TSCAN struggles significantly when video input is unstable, highlighting its reliance on stable frame-to-frame tracking. Both models, despite their sophisticated architectures, demonstrate high vulnerability in real-world, degraded conditions.

One key reason for these performance differences is that DL models like DeepPhys and TSCAN are trained on datasets that contain primarily high-quality, well-lit and stable video recordings. As a result, these models learn to extract features optimized for clean, undistorted input. When presented with distorted data - such as videos with motion artifacts, occlusions or significant blur - these models fail to generalize effectively because the distortions are outside their learned feature space. This explains why their performance deteriorates sharply when video quality degrades. That is not the case for signal processing models like POS and ICA, which do not rely on training data and instead extract pulse signals based on predefined mathematical transformations. This means they do not assume a specific type of video input and are therefore less affected by unexpected distortions. This independence from training data makes signal processing models particularly useful in unpredictable environments where video quality cannot be controlled.

ICA, while maintaining a relatively stable error profile, starts with a higher baseline

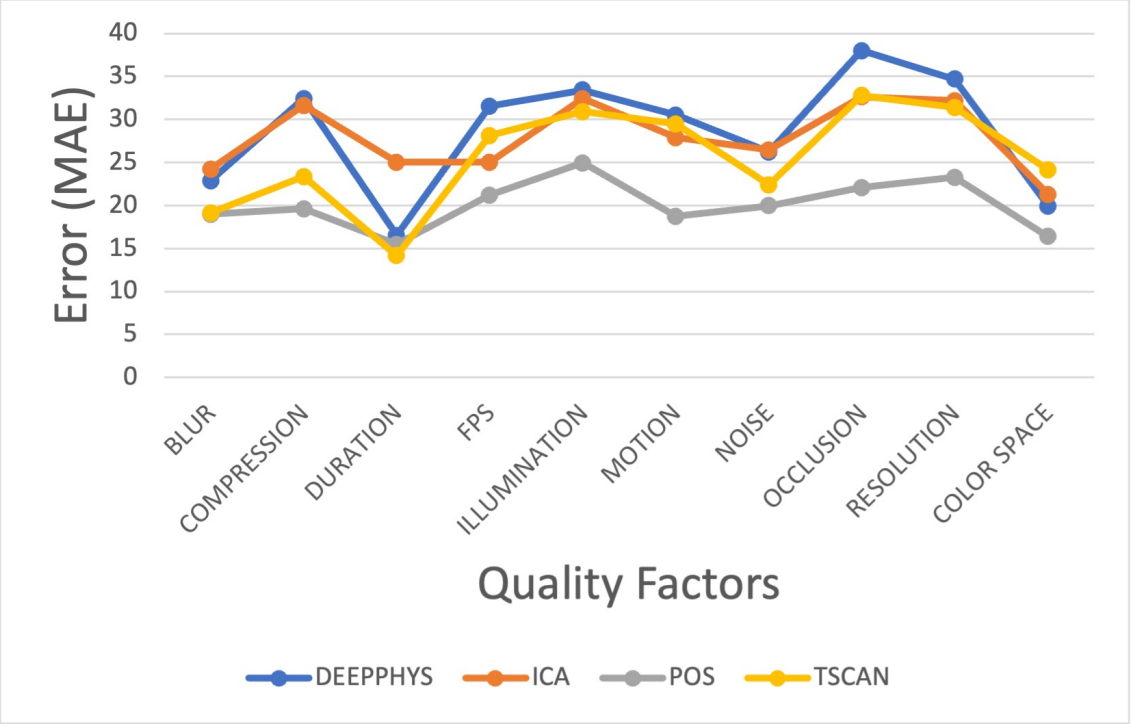


FIGURE 4.8: Line chart with MAE values per quality factor for each model

MAE compared to DeepPhys, POS and TSCAN. This suggests that ICA's signal extraction approach is inherently less precise in ideal conditions but does not degrade as drastically under distortions such as contrast, brightness, blur and FPS reductions. Unlike deep models, which track fine-grained spatial and temporal features, ICA separates independent signal components without optimizing for subtle facial cues. Consequently, ICA's insensitivity to high-quality input makes it more resilient to distortions but also limits its ability to achieve fine-tuned accuracy.

A key observation from Figure 4.6 is that, despite differences in absolute MAE values, all models follow similar relative trends in sensitivity to different video distortions. While DeepPhys, ICA, POS and TSCAN show varying degrees of performance loss, the ranking of which quality factors impact them the most remains consistent across models. This consistency suggests that rPPG distortions can be quantified, reinforcing the feasibility of a video quality metric to assess rPPG suitability across different approaches.

These findings highlight the importance of aligning rPPG models with their intended application. For high-precision settings, such as clinical monitoring or controlled laboratory environments where video quality is optimized, DL models like DeepPhys and TSCAN are ideal due to their superior accuracy in ideal conditions. However, in real-world applications where video quality is variable, signal processing models like POS and ICA offer greater reliability. Their ability to tolerate noise, motion and illumination

	DEEPPHYS	ICA	POS	TSCAN
BLUR	22.85	24.25	18.97	19.15
COMPRESSION	32.41	31.61	19.61	23.37
DURATION	16.51	25.00	15.48	14.17
FPS	31.55	25.04	21.18	28.15
ILLUMINATION	33.42	32.44	24.96	30.89
MOTION	30.49	27.91	18.71	29.47
NOISE	26.20	26.44	19.97	22.37
OCCLUSION	37.98	32.63	22.09	32.78
RESOLUTION	34.68	32.20	23.25	31.41
COLOR SPACE	19.92	21.28	16.40	24.16
ORIGINAL	**1.20**	**11.83**	**4.05**	**0.92**

FIGURE 4.9: MAE values for DeepPhys, ICA, POS and TSCAN across various video quality conditions.

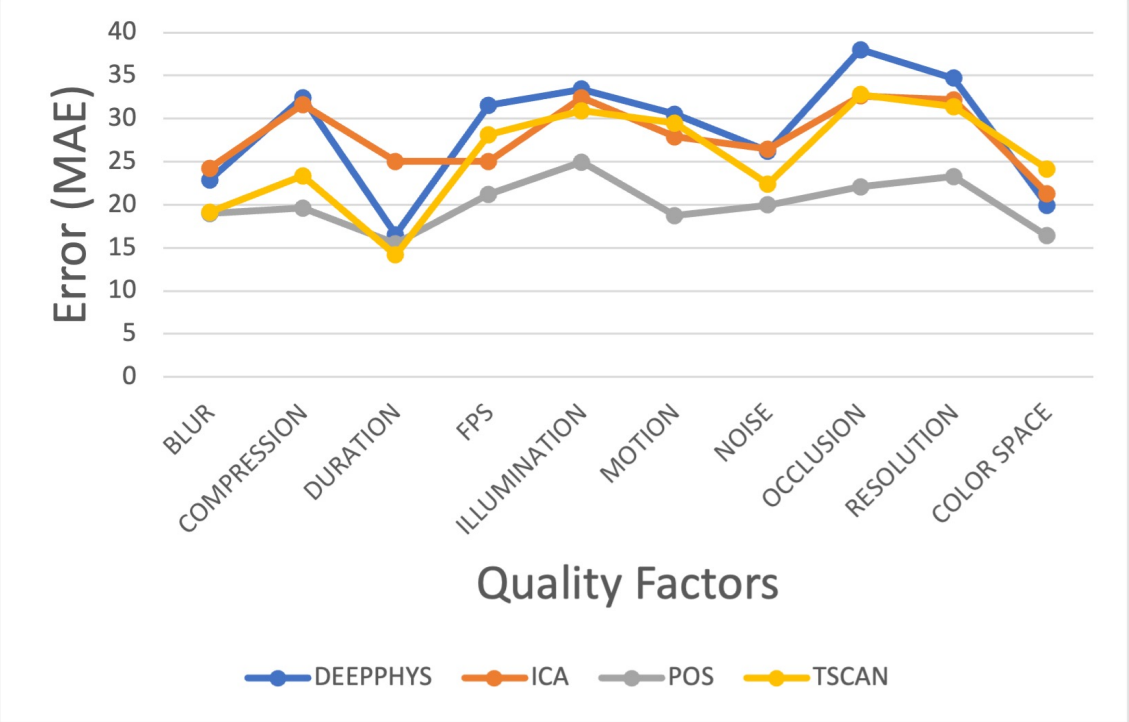


FIGURE 4.10: Line chart with MAE values per quality factor for each model

fluctuations makes them better suited for environments where data quality cannot be guaranteed.

Ultimately, we believe that no single model is universally superior; the best choice depends on the expected video conditions. A promising direction involves hybrid approaches that integrate DL with robust signal processing techniques, allowing models to balance accuracy with resilience. This can enable rPPG systems to maintain high

performance across diverse real-world scenarios, bridging the gap between precision and generalizability.

4.2 From Gut Feeling to Numbers: Building rPPG Quality Metrics

Based on the insights of the model evaluation study, we present the development and evaluation of two metrics designed to assess video quality for rPPG applications. They aggregate individual quality factors (illumination, resolution and motion artifacts among others) into a single score, providing a holistic evaluation of each video.

4.2.1 Experimental Setup

4.2.1.1 Datasets

For this part of our work, along with UBFC-rPPG, we use MMSE-HR and the more recently obtained COHFACE [Heusch et al. (2017)]. The combination of these datasets captures diverse video quality conditions, making them ideal for testing the robustness of the video metrics.

Details on UBFC-rPPG and MMSE-HR can be found in previous chapters. COHFACE, on the other hand, is a dataset designed for research in rPPG and facial video analysis. It contains 40 video recordings of 40 participants, captured under controlled conditions to evaluate the performance of physiological signal estimation algorithms. Each recording includes synchronized facial videos recorded with a Logitech HD Webcam C525 camera and ground truth physiological signals, such as heart rate and respiratory rate, measured using contact-based sensors. The duration of each video is a minute and videos were recorded at 20 fps with a resolution of 640 by 480 pixels. The dataset features variations in facial expressions and slight natural head movements to emulate real-world scenarios, while maintaining a controlled environment to limit extreme quality degradations. COHFACE is particularly valuable for studying the impact of small to moderate facial expressions, head movements and varying illumination conditions on rPPG performance.

As mentioned previously, the rPPG-Toolbox was originally configured to work with UBFC-rPPG. To accommodate for MMSE-HR and COHFACE, we manually reconfigure its pipeline. In this study, we once again use the pre-trained DeepPhys and TSCAN on the PURE dataset to extract pulse measurements for the rest of the datasets.

4.2.1.2 What Features Are We Tracking?

To help us build the video quality metrics metrics, we extract a series of video analytics from the three datasets we selected. These analytics quantify the key aspects of video quality that we analyzed in the previous section. Specifically, we extract the video resolution, frame rate, video duration, occlusion percentage, number of skin pixels, average motion, motion variability, maximum motion, average illumination, maximum illumination, illumination variability and skin tone characteristics.

The resolution of each video is determined by extracting its width and height in pixels and the video duration is calculated by dividing the total number of frames by the frame rate. Occlusion percentage quantifies the extent to which the face is partially or fully obstructed. We leverage SegFormer, a deep facial segmentation model that identifies different facial components, such as the forehead, cheeks, eyes, nose and mouth [Xie et al. (2021)]. The total skin area is estimated by segmenting skin-related labels in the model's output. To ensure that occlusion is assessed in a relevant area for rPPG, we define an elliptical region that captures the face and compute the number of occluded pixels within this region, storing the percentage of occlusion for each participant.

We extract multiple motion-related features to capture different aspects of motion within the video, as it is one of the key factors affecting rPPG performance. Average motion is computed by tracking feature points on the face using optical flow, which estimate the displacement of facial landmarks between consecutive frames. The motion variability is calculated as the standard deviation of the motion values over the duration of the video, capturing the level of fluctuation in movement. A high standard deviation indicates inconsistent or abrupt movements. The maximum motion represents the largest displacement observed within the tracked facial features, which helps identify whether the video contains periods of excessive movement.

We analyze average illumination by computing the mean pixel intensity of each frame in grayscale. Since uniform illumination does not always reflect real-world conditions, we also measure maximum illumination, which captures the highest intensity observed in the video. To quantify the stability of lighting conditions, we compute the illumination variability, defined as the standard deviation of illumination values across frames. This measure provides insight into how consistently the lighting is maintained throughout the video and whether any abrupt fluctuations occur.

Using Stone, a facial skin tone detection model [Rejon Pina and Ma], we estimate the most representative skin tone in each video and classify it into Hexadecimal (Hex) color code and the tone label. We assess blur levels within the videos, using the variance of the Laplacian method, which measures the sharpness of edges in an image. Finally, noise is quantified by computing the standard deviation of high-frequency components

in the image, which highlights random pixel intensity variations that arise from sensor noise or compression artifacts.

4.2.1.3 Feature Transformation

To ensure comparability across datasets while maintaining variability, we apply a two-tiered approach to feature transformation. Features that remain constant within datasets but vary across datasets are normalized globally, whereas features that exhibit variability within datasets are standardized locally. This transformation enables meaningful regression analysis and ensures that our metric accounts for both global dataset characteristics and dataset-specific variations.

The features that remain constant within datasets but vary across datasets are resolution and frame rate, so we apply min-max normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.1}$$

where X represents the raw feature value and X_{min} and X_{max} are the minimum and maximum values across all datasets.

On the other hand, features that exhibit variation within datasets are standardized separately per dataset to preserve distributions:

$$X' = \frac{X - \mu}{\sigma} \tag{4.2}$$

where μ and σ are the mean and standard deviation of the feature within each dataset.

We also perform a Variance Inflation Factor (VIF) analysis on all datasets to detect multicollinearity. VIF quantifies how strongly a predictor variable is correlated with other predictors, with high values (typically VIF $>$10) indicating redundancy that can distort coefficient estimates. Our initial analysis revealed strong multicollinearity between resolution pixels and skin pixels, with VIF values of 73.61 and 61.68, respectively. This suggests that both variables capture the same underlying information. Since skin pixels directly represent facial visibility, we removed resolution pixels to avoid redundancy. Additionally, average illumination exhibited an unusually high VIF ($>$50), indicating a potential non-linear relationship with performance. To account for this, we introduced an illumination squared term, allowing the model to better capture potential non-linear effects.

4.2.2 Weighted Sum Metric (WS_QM)

Given a set of video quality features, the goal is to assign weights to them so that the resulting metric correlates well with rPPG performance, measured using MAE. Based on the analysis we performed in our previous section, we derive initial feature weights from those controlled experiments and use them as priors in the regression-based metric.

To formalize that, the degradation effect of quality feature X_i was computed as:

$$\Delta MAE_i = MAE_{degraded} - MAE_{original} \tag{4.3}$$

where ΔMAE_i represents the increase in error due to degradation i, $MAE_{original}$ is the baseline error before degradation and $MAE_{degraded}$ is the error after introducing the degradation.

Since different video quality features had varying magnitudes of impact, a normalization step (4.4) is applied to ensure comparability. The degradation effects are converted into relative importance scores by normalizing across all features, which ensures that the weights sum to 1.

$$W_i = \frac{\Delta MAE_i}{\Sigma_j \Delta MAE_i} \tag{4.4}$$

where W_i is the normalized weight for feature i and $\Sigma_j \Delta MAE_i$ is the total sum of all MAE increases across all features.

These weights are incorporated as priors for a linear regression model, allowing it to adjust the weights based on dataset-specific performance trends. This approach helps account for differences in data distributions and feature interactions. The regression model is trained with MAE as the target variable, refining the weights to maximize correlation between the predicted video quality score and actual rPPG performance. To address training bias, ensuring that the learned metric is not overly influenced by the behavior of a specific rPPG method, the target MAE is the average MAE across models. The adjusted weights are used to estimate WS_QM:

$$WS\_QM = \sum_{i=1}^{n} W_i^{adj} X_i \tag{4.5}$$

where X_i is each features value, W_i is weight derived from the linear regression model and n the total number of features.

4.2.3 ML Based Metric (ML_QM)

While WS_QM provides initial insights into the relationship between video quality and rPPG performance, it exhibits limitations in generalizing across datasets as we will analyze later in this chapter. To further improve the predictive power of the video quality metric, an ML-based approach is developed, ML_QM. The objective of this approach is to learn non-linear relationships within video quality features, thereby capturing interactions that could not be adequately modeled using linear methods.

We implement a supervised learning framework, where video quality features serve as input variables and a target score as the desired output. To calculate the target score, we average MAE across models and standardize it by applying Min-Max normalization, rescaling the average MAE between 0 and 1. The former helps us address training bias, as with WS_QM, and the latter gives us a reliable proxy for assessing video suitability. The training data consist of video analytics we extract in section 4.2.1.2, including both unedited and synthetically degraded versions of UBFC-rPPG, to introduce greater variability in feature distributions. This augmentation helps mitigate biases arising from dataset-specific characteristics. We test the models predictions on the two remaining datasets.

The models we use for this study include Random Forest, Gradient Boosting, eXtreme Gradient Boosting (XGBoost), KNN and Extra Trees. These approaches are selected due to their ability to capture complex feature interactions and rank feature importance. Unlike linear regression, these models can better adapt to variations in dataset characteristics.

The ML models are trained using a standard regression framework. Given an input feature vector $X = (X_1, X_2, ..., X_n)$, where each X_i represents a video quality factor, the models learn a function $f(X)$ to predict the rPPG error Y.

4.2.4 Results

We assess the accuracy of our video quality metrics by computing the Pearson correlation between the predicted video quality score and the rPPG performance error (MAE). Pearson correlation measures the linear relationship between two variables, where a value of $r = 1$ indicates a perfect positive correlation, $r = -1$ indicates a perfect negative correlation and $r = 0$ implies no correlation. Here we are aiming for a perfect positive correlation, because ML_QM is trained to predict higher quality metric scores for videos with higher errors, and WS_QM assigns higher weights to features that cause greater degradation. As a result, both metrics are expected to increase as rPPG error increases.

Pearson correlation is computed as:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2}\sqrt{\Sigma(Y_i - \bar{Y})^2}} \tag{4.6}$$

where Xi and Yi are the video quality score and MAE and $\bar{X}$ and $\bar{Y}$ are the means of the video quality scores and MAE values.

The strength and significance of this correlation are evaluated using the p-value, which indicates whether the observed relationship is statistically meaningful. A low p-value ($p<0.05$) suggests a significant correlation, while a high p-value implies that the correlation may be due to chance.

4.2.4.1 *WS_QM* Results

Table 4.2 presents the correlation for *WS_QM* across datasets. Our results indicate that the generalizability of *WS_QM* is inconsistent. While weights derived from UBFC-rPPG generalized relatively well to COHFACE and MMSE-HR, the same is not observed for the rest. This suggests that the influence of video quality factors is not uniform across datasets, likely due to differences in recording conditions, subject demographics or preprocessing pipelines. Assuming a strictly linear relationship between individual quality features and rPPG performance may also limit *WS_QM*'s effectiveness, as certain degradations may have non-linear effects or interact in ways that the weighted sum approach cannot fully capture.

Weights	Dataset	Correlation with MAE	p-value
MMSE-HR	MMSE-HR	0.526	7.28E-05
MMSE-HR	COHFACE	-0.204	9.86E-03
MMSE-HR	UBFC-rPPG	0.541	2.62E-04
UBFC-rPPG	MMSE-HR	0.229	3.28E-03
UBFC-rPPG	COHFACE	0.404	3.68E-03
UBFC-rPPG	UBFC-rPPG	0.722	9.75E-08
COHFACE	MMSE-HR	0.246	1.76E-03
COHFACE	COHFACE	0.414	2.54E-03
COHFACE	UBFC-rPPG	0.587	7.01E-07

TABLE 4.2: Correlation between video quality metric and rPPG performance across different datasets.

Within individual datasets, *WS_QM* exhibits strong correlations with rPPG error, with UBFC-rPPG showing the highest correlation, followed by MMSE-HR and COHFACE. This suggests that the metric effectively captures video quality effects when applied within the dataset from which it was derived. However, cross-dataset performance revealed more variability. For example, UBFC-rPPG weights applied to COHFACE produced a low to moderate correlation, whereas MMSE-HR weights applied to COHFACE resulted in a negative correlation. This suggests that video quality effects

in MMSE-HR and COHFACE differ significantly, reinforcing the idea that individual dataset characteristics strongly influence rPPG performance.

The inconsistencies in cross-dataset performance highlight the limitations of using a linear regression model for generalization across datasets. The weaker correlation in COHFACE compared to the strong within-dataset correlation in UBFC-rPPG suggests that the impact of video quality factors is dependent on the dataset, influenced by variations in lighting conditions, resolution and motion artifacts. Additionally, the presence of negative correlations implies that the linear nature of WS_QM may not fully capture complex feature interactions, potentially leading to misleading quality assessments.

A key factor influencing these results is the alignment between dataset characteristics and rPPG model performance. Since the rPPG models used in this study were originally trained on datasets with characteristics more similar to UBFC-rPPG, the impact of video quality factors in this dataset is more predictable. In other words, potentially the rPPG models' errors align more consistently with video quality degradations in UBFC-rPPG. As a result, the WS_QM derived from UBFC-rPPG generalizes better to other datasets, particularly COHFACE, compared to metrics derived from datasets with different statistical properties.

These findings emphasize the importance of dataset characteristics in formalizing quality-performance relationships and reinforce the need for more flexible, data-driven approaches to quality assessment. While WS_QM provides an interpretable baseline, it assumes linear independence of features. However, real-world rPPG degradation is often nonlinear, which WS_QM cannot model accurately.

4.2.4.2 ML_QM

The results presented in Table 4.3 provide proof that data-driven approaches can enhance the generalizability of our video quality metric. When tested on MMSE-HR, ML_QM generally outperformed WS_QM. Extra Trees, for instance, achieves a Pearson correlation of 0.618, which is significantly higher than the correlation observed with WS_QM. Similarly, XGBoost and Gradient Boosting achieve correlations of 0.626 and 0.524 respectively, suggesting that these models can more effectively capture the relationship between video quality and rPPG error.

ML_QM also performs well when tested on COHFACE, with correlations ranging from 0.387 (KNN) to 0.763 (Random Forest), showing a stronger generalization capability compared to WS_QM. The performance of Extra Trees on COHFACE further indicates that ensemble-based models effectively leverage non-linear relationships between video quality factors and rPPG performance.

Training Dataset	Test Dataset	Model	Pearson Correlation (p-value)
UBFC-rPPG	MMSE-HR	Random Forest	0.594 (p<0.00001)
UBFC-rPPG	MMSE-HR	Gradient Boosting	0.524 (p=0.00008)
UBFC-rPPG	**MMSE-HR**	**XGBoost**	**0.626 (p<0.00001)**
UBFC-rPPG	MMSE-HR	KNN	0.270 (p=0.05580)
UBFC-rPPG	MMSE-HR	Extra Trees	0.618 (p<0.00001)
UBFC-rPPG	**COHFACE**	**Random Forest**	**0.763 (p<0.00001)**
UBFC-rPPG	COHFACE	Gradient Boosting	0.607 (p<0.00001)
UBFC-rPPG	COHFACE	XGBoost	0.740 (p<0.00001)
UBFC-rPPG	COHFACE	KNN	0.387 (p<0.00001)
UBFC-rPPG	COHFACE	Extra Trees	0.675 (p<0.00001)
UBFC-rPPG	**UBFC-rPPG**	**Random Forest**	**0.978 (p<0.00001)**
UBFC-rPPG	UBFC-rPPG	Gradient Boosting	0.711 (p<0.00001)
UBFC-rPPG	UBFC-rPPG	XGBoost	0.962 (p<0.00001)
UBFC-rPPG	UBFC-rPPG	KNN	0.480 (p<0.00001)
UBFC-rPPG	UBFC-rPPG	Extra Trees	0.957 (p<0.00001)
UBFC-rPPG - unedited	**MMSE-HR**	**Random Forest**	**0.117 (p=0.41362)**
UBFC-rPPG - unedited	MMSE-HR	Gradient Boosting	-0.071 (p=0.62047)
UBFC-rPPG - unedited	MMSE-HR	XGBoost	-0.039 (p=0.78641)
UBFC-rPPG - unedited	MMSE-HR	KNN	-0.131 (p=0.35897)
UBFC-rPPG - unedited	MMSE-HR	Extra Trees	0.114 (p=0.42517)
UBFC-rPPG - unedited	**COHFACE**	**Random Forest**	**0.204 (p=0.01002)**
UBFC-rPPG - unedited	COHFACE	Gradient Boosting	0.162 (p=0.04131)
UBFC-rPPG - unedited	COHFACE	XGBoost	0.196 (p=0.01307)
UBFC-rPPG - unedited	COHFACE	KNN	0.010 (p=0.89599)
UBFC-rPPG - unedited	COHFACE	Extra Trees	0.168 (p=0.03450)

TABLE 4.3: Performance of ML models trained on UBFC-rPPG and tested on MMSE-HR and COHFACE datasets. Pearson correlation values indicate the strength of association between the predicted video quality metric and rPPG error.

A key advantage of *ML_QM* is the flexibility to learn from diverse data distributions, adjusting to non-linear patterns in the data, allowing them to capture more complex interactions between video quality factors. The stronger generalization of these models suggests that dataset-specific variations, such as differences in illumination, resolution and motion artifacts are better accounted for when training is performed on a dataset that incorporates sufficient variation in quality conditions.

However, the performance of models without the additional edited videos shows a sharp decline, particularly in cross-dataset generalization. For *ML_QM* tested on MMSE-HR all models exhibit near-zero or negative correlations. Similarly, when tested on COHFACE, the highest correlation achieved is 0.204 (Random Forest), which is substantially lower than the results obtained when training included edited videos. This highlights the importance of dataset diversity during training, as models trained on limited variations in video quality struggle to generalize when tested on datasets with different statistical properties.

Both WS_QM and ML_QM highlight the challenges of generalizing video quality assessment across datasets, but ML_QM shows a clear advantage in adaptability. While WS_QM achieves reasonable within-dataset performance, its cross-dataset results indicate inconsistencies, particularly with MMSE-HR and COHFACE, where we even see negative correlations. In contrast, ML_QM with added variations performs significantly better in both within-dataset and cross-dataset evaluations, demonstrating the ability to learn non-linear interactions and adapt to different quality distributions.

4.3 Discussion

The results presented in this study underscore the critical role of video quality in rPPG performance and highlight the need for dedicated evaluation metrics. As established in Chapter 1, prior research has explored individual factors such as motion, illumination, compression and resolution, but no unified framework has been proposed to quantify their combined effects on pulse estimation. By introducing WS_QM and ML_QM, we address this gap, offering a structured and objective method for assessing video suitability in rPPG applications.

Through our video quality analysis, we systematically modified videos across multiple degradation factors, including motion, occlusions, compression, blur, resolution and lighting variations. Our experiments confirmed that motion artifacts, occlusions and resolution reductions were the most detrimental to rPPG accuracy, leading to significant increases in MAE across all models. Illumination changes also had a strong impact, particularly under extreme contrast shifts. In contrast, color space variations and shorter video durations had a lesser effect, though they still introduced some performance degradation. Notably, DL models (DeepPhys, TSCAN) showed greater sensitivity to distortions, while signal processing models (POS, ICA) demonstrated higher robustness to noise and motion but lower baseline accuracy. These findings reinforce the need for a structured quality assessment method, as uncontrolled degradations can severely impact pulse estimation.

WS_QM aggregates multiple video quality factors into a single score. Unlike traditional video quality metrics, WS_QM explicitly models these rPPG-specific degradations, making it a valuable tool for dataset selection, preprocessing and ensuring robustness in real-world conditions. However, its reliance on a linear regression introduces some limitations in generalization, particularly when applied across datasets with differing statistical properties. The variations observed in cross-dataset correlations suggest that the relationship between quality factors and rPPG errors may be more complex than a weighted sum can fully capture.

ML_QM, in contrast, demonstrates superior adaptability and predictive power, leveraging ML to model non-linear interactions between quality factors. Its stronger cross-dataset generalization indicates that learning-based approaches are better suited for capturing the intricate dependencies between video conditions and rPPG performance. The ability of *ML_QM* to maintain high correlation across datasets confirms its potential as a robust and scalable solution for video quality assessment in rPPG. Nonetheless, its performance remains influenced by dataset diversity, as models trained on limited variations struggle to generalize to unseen distributions.

One key factor shaping these results is the alignment between dataset characteristics and rPPG model training data. The rPPG models evaluated in this study were originally trained on datasets with characteristics closely resembling UBFC-rPPG. This may explain why video quality metrics derived from UBFC-rPPG generalized more effectively to other datasets, particularly COHFACE. While this reinforces the relevance of our proposed metrics, it also highlights potential biases in rPPG training data that could affect video quality assessments. Ensuring that rPPG models are trained on diverse datasets remains crucial to mitigating these biases and enhancing the robustness of video quality metrics.

By introducing *WS_QM* and *ML_QM*, this work provides a comprehensive framework for assessing video quality in rPPG applications. These metrics not only help researchers and practitioners understand the impact of video conditions but also serve as tools for dataset evaluation, preprocessing and improving model reliability. The strong correlations observed between these metrics and rPPG errors validate their effectiveness and establish a foundation for more informed video-based physiological monitoring.

While video quality assessment is essential for improving rPPG performance, it does not provide a measure of uncertainty in pulse estimation. In the next chapter, we explore how conformal predictions can be applied to rPPG to quantify confidence in model outputs, further enhancing the reliability of remote physiological measurements.

Chapter 5

Calibration: Confidence with Conformal Predictions

Like in the calibration stage, predictions are fine-tuned and confidence is quantified through conformal predictions. This ensures the model's outputs are not just accurate but also trustworthy.

As we discussed in previous chapters, rPPG has seen rapid advancements, yet one key challenge remains: uncertainty in predictions. Unlike contact-based heart rate monitors, rPPG relies on subtle skin color changes extracted from video, making it sensitive to factors like motion, lighting variations and other artifacts as discussed in the previous chapters. In practical applications, ensuring reliability is just as important as improving accuracy. This becomes a critical need in medical settings where incorrect estimates can lead to incorrect diagnosis and treatment.

Traditional DL and signal processing models for rPPG output point estimates without quantifying their uncertainty. This makes it difficult to trust the predictions with no ground truth data, particularly in cases where the model is uncertain due to unseen variations in data. A solution to this problem is Conformal Predictions (CP), a statistical framework for generating uncertainty intervals that are guaranteed to contain the true value with a specified probability. By leveraging CP, we can enhance the trustworthiness of rPPG systems by providing well-calibrated confidence intervals alongside model predictions.

Chapter Contributions:
This chapter introduces the application of CP for rPPG, quantifying uncertainty in model predictions. By leveraging CP, we establish a framework for evaluating the reliability of rPPG models, ensuring that predictions are accompanied by well-calibrated uncertainty intervals. More specifically, our contributions are as follows:

- We introduce the first application of CP to rPPG, quantifying uncertainty in model predictions and enhancing interpretability.
- We evaluate the effect of different significance levels ($\alpha = 0.1$ and $\alpha = 0.2$) on the prediction intervals, analyzing their impact on interval width and model reliability.
- We compare traditional MAE-based CP with a novel quality-aware CP approach, demonstrating the benefits of incorporating video quality into uncertainty estimation.
- We provide insights into the interpretability of CP in rPPG, showing how prediction intervals can reveal model confidence, making rPPG systems more transparent and usable in practical applications.

The rest of the chapter is organized as follows: Section 5.1 introduces the motivation for applying CP to rPPG and provides an overview of CP principles. Section 5.2 describes the experimental setup, including the datasets, models and methodology for computing conformal intervals. Section 5.3 presents our results, analyzing the effectiveness of CP across different datasets and models. Finally, Section 5.4 discusses the key findings, their implications for rPPG applications and potential future directions.

5.1 Conformal Predictions Background

Given the challenges identified in the previous chapter, where even deep learning models struggle with degraded video quality, it becomes crucial to quantify uncertainty in model predictions. A key question emerges: how confident can we be that rPPG models will maintain performance under all conditions? This question becomes even more pressing in medical applications, where accuracy is essential for reliable diagnosis and decision-making.

An ideal solution would not only provide model predictions but also offer insight into the certainty of those predictions. If we could quantify a model's confidence in its outputs, we could make more informed decisions about whether to trust a given prediction or discard it.

CP provide exactly this capability. They "wrap" around an existing ML model, generating prediction intervals rather than single point estimates. The key principle behind CP is consistency with past data, hence the term "conformal." It evaluates how unusual a new test sample is compared to previously seen data and adjusts the confidence interval accordingly. When a test sample is similar to prior observations, the interval is narrow, indicating higher certainty. Conversely, when the sample deviates significantly from the training distribution, the interval widens, reflecting greater uncertainty.

In more detail, CP is a method that constructs prediction intervals or prediction sets for ML models. It ensures that, under minimal assumptions, the set contains the true label with a guaranteed confidence level.

Applying CP to rPPG represents a significant step forward in building trustworthy, transparent and safe physiological monitoring systems. In clinical settings CP can help flag predictions that are unreliable due to poor video quality or challenging conditions, allowing clinicians to request re-measurements instead of relying on uncertain outputs. In consumer-grade wearables, CP can improve user trust by indicating when measurements are reliable and when external factors (such as movement or lighting) may compromise accuracy. Conformal predictions can transform rPPG systems from "black-box" predictors into transparent tools, enabling users to understand not only what the model predicts, but also how confident it is in each prediction.

Beyond improving reliability, CP opens new avenues for personalizing rPPG systems to individual users and real-time contexts. By continuously monitoring prediction uncertainty, future rPPG systems could dynamically adjust measurement strategies. For example automatically requesting longer recordings when confidence is low or combining multiple short measurements when conditions are unstable. Furthermore, CP could enable the development of self-aware wearable devices that intelligently adapt to user behavior, environment or physiological state, ensuring that the quality of monitoring remains high without manual intervention. In broader terms, integrating CP into rPPG pipelines paves the way for next-generation vital sign monitoring solutions that are not only accurate but also proactive, intelligent and context-sensitive, significantly advancing the field toward real-world adoption.

5.1.1 CP Framework

The theoretical framework of CP for regression, including the formulation of prediction intervals, nonconformity measures and coverage guarantees, follows the approach outlined by Angelopoulos and Bates (2021). While they discuss additional broader aspects of CP, such as full conformal prediction and multi-output prediction sets and adaptive conformal methods, these aspects are not directly applicable to our setting, so they are not presented. Finally in their work they provide a rigorous mathematical treatment of

the topic. In this chapter, we focus on a more applied presentation, emphasizing how CP is integrated into rPPG rather than its formal derivation.

5.1.1.1 Problem Definition

We begin with a dataset of videos along with their corresponding ground truth heart rate values obtained from a device (e.g., ECG or PPG sensor). Let:

$$D = (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) \tag{5.1}$$

where X_i represents the video (or extracted features from the video) and Y_i is the ground truth heart rate from the medical device. We have a trained rPPG model that predicts heart rate, $f(X_i)$. Our goal is to estimate a prediction interval [Lower,Upper] for a new test video X_{n+1} so that:

$$P(Y_{n+1} \in [Lower, Upper]) \geq 1 - \alpha \tag{5.2}$$

where $1 - \alpha$ is the confidence level (e.g., 95% confidence).

5.1.1.2 Nonconformity Measure

To apply conformal prediction, we need a nonconformity score that captures the error between the predicted values and the ground truth. There are several choices on the nonconformity measure, from simpler like MAE to more complex, combining video quality metrics.

1. **Absolute Error-Based Nonconformity:**

$$S_i = |Y_i - \hat{Y}_i| \tag{5.3}$$

   Measures how far the predicted heart rate $\hat{Y}_i$ is from the true value Y_i. It is a simple nonconformity measure, is widely used in regression CP and it does not require additional model modifications.

2. **Uncertainty-Aware Nonconformity:**

$$S_i = \frac{|Y_i - \hat{Y}_i|}{\sigma(X_i)} \tag{5.4}$$

This nonconformity measure incorporates an uncertainty estimate $\sigma(X_i)$ from the model. It can be useful when the model provides variance estimates but requires it to explicitly compute uncertainty, which may not always be available.

3. **Quality-Aware Nonconformity**

$$S_i = |Y_i - \hat{Y}_i| \times (1 + Q(X_i)) \tag{5.5}$$

Here, a score $Q(X_i)$ is introduced based on video quality factors like illumination, motion artifacts and resolution among others. This penalizes predictions made on poor-quality videos by increasing their nonconformity score.

5.1.1.3 Constructing Prediction Intervals

Once the nonconformity scores S_i for a set of calibration samples have been computed, they can be used to construct prediction intervals for new test samples. The fundamental idea behind CP is to determine an appropriate uncertainty range based on past prediction errors. This ensures that new predictions are accompanied by intervals that are statistically calibrated to contain the true value with high probability. To determine this appropriate range for prediction intervals, we compute a statistical value known as the quantile. A quantile represents a cutoff value in a sorted list of numbers. In the context of conformal prediction, it defines the maximum error we expect for a given confidence level.

Mathematically, we estimate the $(1 - \alpha)$ quantile of the distribution of nonconformity scores:

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha}(S_1, S_2, ..., S_n) \tag{5.6}$$

where $S_1, S_2, ..., S_n$ are the nonconformity scores computed from the calibration set. This quantile represents the error threshold that will be used to determine the prediction intervals.

For example, if we set $\alpha = 0.1$, meaning we want a 90% confidence interval, we compute the 90th percentile of all past prediction errors. This value tells us that, in 90% of past cases, the model's error was below this threshold. The intuition is that if the model's past errors followed a consistent pattern, future errors will likely behave similarly.

Once the quantile $q_{1-\alpha}$ is determined, we construct a prediction interval for a new test sample X_{n+1}. If the model predicts a heart rate value $\hat{Y}_{n+1}$ for this sample, we define the prediction interval as:

$$\left[\hat{Y}_{n+1} - q_{1-\alpha}, \hat{Y}_{n+1} + q_{1-\alpha}\right] \tag{5.7}$$

This means that the true heart rate value Y_{n+1} will fall within this interval at least $(1 - \alpha)$ fraction of the time.

5.1.2 Types of Conformal Predictions

Several variants of CP exist, each with different computational trade-offs and practical considerations. These variants primarily differ in how they handle the calibration process and how they construct prediction intervals.

5.1.2.1 Transductive CP (TCP)

The key idea behind TCP is that it considers each new test sample independently and constructs its prediction interval while incorporating the test sample into the calibration process. It is theoretically optimal in terms of validity and coverage guarantees, however it is computationally expensive as it requires recalculating the nonconformity scores for each test sample. This makes it impractical for large datasets, especially in DL.

5.1.2.2 Inductive CP (ICP)

ICP improves efficiency by separating the dataset into three parts: the training set, which is used to train the ML model, the calibration set, used to compute nonconformity scores and determine the quantile threshold and the test set where predictions and confidence intervals are made. By decoupling calibration from testing, ICP is significantly more efficient than TCP. However, it has a slight trade-off in efficiency because it requires withholding part of the dataset for calibration. The choice of the calibration set can also impact interval widths.

5.1.2.3 Split-CP

Split-CP is a simplified and widely used variant of ICP, especially suited for DL. It uses a pre-trained model and applies CP using a separate calibration set. It is the most computationally practical approach for deep models. In this work, we adopt split-CP due to its efficiency and compatibility with DL pipelines. Unlike TCP, split-CP does not require retraining the model for each test point, making it feasible for large-scale applications. Additionally, split-CP allows for quick adjustments of confidence intervals by selecting different nonconformity measures.

5.1.3 Applications of CP in rPPG

Applying CP to rPPG models has the potential to address several challenges that arise in real-world applications, particularly in handling motion artifacts, compression artifacts and other variations in video quality. As we saw in detail in chapter 4, these factors introduce significant uncertainty in heart rate estimation, making it difficult to ensure reliability in diverse settings. By generating prediction intervals, CP allows rPPG models to quantify uncertainty, ensuring that predictions remain reliable even when faced with challenging conditions such as low resolution videos, variations in skin tone, changing lighting conditions and occlusions.

Traditional DL models for rPPG are often trained on high-quality videos, making them vulnerable to distribution shifts when applied in the wild. If an rPPG model encounters a video with degraded quality, CP will produce wider confidence intervals, indicating greater uncertainty in the heart rate estimate. This prevents overconfident decisions in conditions where performance is likely to degrade. If quality is optimal CP will produce smaller intervals providing confidence in the model's predictions. This is particularly crucial for medical and health monitoring applications, where incorrect heart rate estimations can lead to false alarms or missed critical conditions. Its application could improve model trustworthiness and robustness significantly.

5.1.4 Challenges and Open Questions

While CP provide a powerful framework for quantifying uncertainty in rPPG-based heart rate estimation, several practical challenges must be addressed before it can be fully adopted in real-world applications. These challenges stem from computational constraints, assumptions about data distribution and the complexity of physiological signals.

One of the main limitations of CP is its computational cost, particularly in applications requiring real-time heart rate monitoring. The CP framework relies on computing nonconformity scores for a calibration set and estimating the quantile of the error distribution for each new test sample. This step can be computationally expensive, especially when dealing with large-scale video data or real-time processing in resource-limited environments. Unlike static datasets used in traditional ML, rPPG processes continuous video streams. Each new frame or video segment requires computing prediction intervals, which involves sorting and evaluating calibration errors dynamically. Efficient CP variants, such as online or streaming conformal prediction, could help reduce computational overhead by updating confidence intervals incrementally instead of recomputing them from scratch.

CP assumes that the training, calibration and test samples are exchangeable, meaning they come from the same distribution. However, physiological signals are inherently non-stationary, they change over time due to biological rhythms, stress levels and external environmental factors. This is a challenge because patterns seen in the past may not always generalize to future observations. Adaptive CP techniques, such as time-series conformal prediction or weighted conformal prediction, could help by adjusting confidence intervals dynamically based on recent observations rather than assuming a static distribution.

The choice of nonconformity measure significantly impacts the quality of CP prediction intervals. In traditional regression tasks, simple metrics like mean absolute error (MAE) are commonly used. However, in rPPG, heart rate predictions depend on complex spatiotemporal patterns in video data, making it challenging to define a single nonconformity function that effectively captures both spatial and temporal uncertainties. Video quality variations (e.g., motion blur, illumination changes) introduce additional uncertainty that is not well captured by simple error-based nonconformity scores. More advanced quality-aware nonconformity functions could incorporate video quality metrics (such as motion artifacts, contrast variations and occlusions) into the error calculation. Additionally, hybrid approaches that combine model uncertainty (e.g., Bayesian DL) with CP could lead to better-calibrated intervals.

Finally, the performance of CP depends on the calibration set used to estimate the quantiles of nonconformity scores. If the calibration set is too small or not representative of real-world conditions, the resulting prediction intervals may be either too wide (overly conservative) or too narrow (overconfident). In rPPG, calibration data collected in controlled laboratory conditions may not accurately represent real-world settings where subjects experience varying lighting, skin tones, head movements or environmental noise. Using stratified calibration sets that include a diverse range of conditions can improve robustness. Additionally, domain-adaptive CP methods could allow the model to recalibrate dynamically when deployed in new environments.

Despite these challenges, CP remains a promising approach for improving uncertainty quantification in rPPG models. By addressing computational efficiency, adapting to temporal dynamics and designing better nonconformity measures, CP can enhance trust, interpretability and deployment readiness for rPPG-based heart rate monitoring in real-world applications.

5.2 CP in rPPG Experiments

In this section we present the framework for CP, detailing the datasets, the nonconformity measures, the type of CP and then presenting our results.

5.2.1 Datasets

For this study, we select two publicly available datasets that provide video recordings with corresponding ground truth heart rate signals: COHFACE, a widely used dataset for rPPG evaluation that includes facial videos recorded under controlled conditions with synchronized physiological signals and UBFC-rPPG, a well-established dataset in rPPG research that contains videos recorded at different frame rates and lighting conditions, making it particularly useful for assessing the robustness of CP. More details on the specifications of COHFACE can be found in section 4.2.1.1 and on UBFC-rPPG in section 3.1.1.

Although MMSE-HR is a widely used benchmark, we exclude it due to its high motion artifacts, large synchronization errors and substantial distribution shift from training data. These factors would result in overly wide CP intervals, making it difficult to isolate the effect of CP itself.

5.2.2 Models

In this chapter, we apply CP to DL rPPG models, specifically DeepPhys and TSCAN, rather than signal processing models like POS and ICA. The primary reason for this selection is that CP is most effective in scenarios where model predictions exhibit some degree of variability or uncertainty. Unlike DL models, which incorporate stochastic elements such as weight initialization, dropout layers and optimization dynamics, signal processing models are deterministic; they always produce the same output for the same input.

This deterministic nature makes conformal prediction less useful for models like POS and ICA. Since these methods rely on fixed transformations and statistical operations, their predictions do not change across multiple runs. This was proven in our experiments, their conformal intervals had zero-width. In such cases, the intervals do not provide meaningful insights into uncertainty because they simply confirm the model's fixed behavior rather than capturing prediction variability.

DL models, on the other hand, naturally exhibit uncertainty due to factors such as data variability, weight updates during training and inherent noise in video-based pulse estimation. Applying conformal prediction to DeepPhys and TSCAN allows us to quantify the confidence in their predictions by associating each heart rate estimate with a statistically calibrated uncertainty interval. This is particularly important in real-world applications, where factors such as motion artifacts, lighting variations and other artifacts can introduce variability in model predictions.

ST2S-rPPG was excluded because its two-stage design conflicts with CP's single-stage assumption. Since its second stage selectively filters unreliable predictions, applying

CP would either be invalid (if done before filtering) or inconsistent (if done after filtering). Given that the second stage already acts as a confidence mechanism, CP would be redundant in this case.

5.2.3 Selection of Nonconformity Measures

A key component of conformal prediction is the nonconformity measure, which determines how "unusual" a prediction is relative to past observations. The choice of nonconformity function directly influences the quality and reliability of the conformal prediction intervals. In this study, we evaluate two different nonconformity functions: a standard error-based measure and a quality-aware measure that incorporates ML_QM as an additional factor.

5.2.3.1 Error-Based Nonconformity Measure

The first approach follows the widely used absolute error-based nonconformity measure. In this case, the nonconformity score is computed as the MAE between the predicted and true heart rate values. For a given sample, the nonconformity score is defined as:

$$S_i = |Y_i - \hat{Y}_i| \tag{5.8}$$

where Y_i is the ground truth heart rate and $\hat{Y}_i$ is the predicted heart rate. This approach is computationally efficient and straightforward to implement, making it a common choice in regression-based conformal prediction studies. By directly measuring the deviation between predictions and actual values, it provides an estimate of prediction uncertainty without introducing additional assumptions. Furthermore, it has been extensively applied in prior work on CP and serves as a well-established baseline for comparison.

5.2.3.2 Quality-Aware Nonconformity Measure

While the absolute error-based approach provides a simple and effective means of quantifying nonconformity, it assumes that all samples are equally difficult to predict. However, as demonstrated in previous chapters, rPPG performance is highly dependent on video quality. Factors such as motion artifacts, illumination variability, resolution and occlusions significantly influence heart rate estimation accuracy. To address this limitation, we introduce a quality-aware nonconformity measure based on the video's quality score.

The proposed quality-aware nonconformity function is defined as:

$$S_i = Q(X_i) \tag{5.9}$$

where $Q(X_i)$ represents the ML_QM score, which quantifies the impact of various degradation factors on rPPG performance. This quality score is computed using ML_QM developed in chapter 4, which was designed specifically for rPPG applications. The metric incorporates multiple video quality attributes and was validated by demonstrating its correlation with model performance. By leveraging this metric, the quality-aware nonconformity measure ensures that videos with lower quality are assigned higher nonconformity scores, leading to wider prediction intervals.

The motivation for incorporating a quality-aware adjustments stems from the fact that rPPG models exhibit variable performance depending on input conditions. Traditional conformal prediction methods operate under the assumption that all test samples are drawn from the same distribution and share a similar level of difficulty. However, this assumption does not apply in rPPG applications, where signal extraction reliability varies significantly based on video characteristics. By integrating ML_QM into the nonconformity function, the conformal intervals adapt dynamically to account for varying levels of uncertainty. In practice, this means that high-quality videos, where rPPG models are expected to perform reliably, will yield narrower prediction intervals, whereas low-quality videos, which introduce greater ambiguity, will be assigned wider intervals. This dynamic adjustment improves both the robustness and interpretability of the predictions, ensuring that uncertainty estimates more accurately reflect the reliability of the underlying model.

By comparing these two approaches, we assess the impact of integrating video quality into conformal prediction. The absolute error-based approach serves as a strong baseline, providing a direct estimate of model deviation, while the quality-aware methods introduce an additional layer of adaptability to account for variations in input conditions.

5.2.4 Conformal Prediction Method: Split Conformal Prediction

In this study, we employ split-CP as our uncertainty estimation method. Split-CP is a computationally efficient variant of conformal prediction that provides statistical coverage guarantees while maintaining computational feasibility. Split-CP operates by reserving a separate calibration set to estimate nonconformity scores and determine the prediction interval width. This approach ensures that the computed uncertainty estimates remain well-calibrated without modifying the base model.

Split-CP is particularly advantageous for rPPG applications due to its scalability, compatibility with DL models and computational efficiency. Split-CP reduces computational overhead by only requiring the sorting of calibration errors and the computation of quantiles, making it well-suited for DL-based rPPG models.

To apply split-CP in this study, we utilize the pre-trained models from chapter 3, taken from the rPPG Toolbox [Liu et al. (2024)]. We run separate experiments with COHFACE and UBFC-rPPG. Each dataset is split into a calibration set and a test set. The calibration set is held out from training and is used to compute nonconformity scores, while the test set is used for final evaluation. Using the calibration set, we calculate the two previously mentioned types of nonconformity scores: MAE and the custom quality-aware metric. These scores capture the degree of deviation between the model's predictions and the true heart rate values.

After computing the nonconformity scores, we estimate the $(1 - \alpha)$ quantile of the distribution of these scores. This quantile serves as the threshold for determining the width of the prediction interval. We experiment with $\alpha = 0.1$ and $\alpha = 0.2$, capturing 90% certainty and 80% certainty respectively. The computed threshold is then applied to new test samples, ensuring that each prediction is accompanied by a confidence interval that accounts for the model's uncertainty.

5.2.5 Implementation Modifications for Conformal Prediction

Since CP are not supported in the rPPG Toolbox, we make several modifications to integrate CP into the framework. We modified the testing pipeline to split the dataset into calibration and test sets. We implemented custom functions to compute nonconformity scores using both MAE and quality-based metrics. We estimate the quantile for uncertainty estimation and apply split-CP and generate prediction intervals. We added functionality to track predictions per video, allowing us to analyze how interval width changes based on video characteristics and whether prediction intervals adapt dynamically to video quality.

5.3 Results

In this section, we begin by analyzing the baseline conformal prediction results using MAE as the nonconformity measure, comparing different significance levels ($\alpha = 0.1$ and $\alpha = 0.2$). Finally, we assess the impact of incorporating ML_QM as a nonconformity measure, comparing its performance to the MAE-based approach.

We must emphasize that in contrast to previous chapters, results in this section are presented as per-frame pulse predictions rather than aggregated heart rate (bpm) estimates. This is because the models operate at the frame level, predicting a continuous pulse waveform rather than a single bpm value per video. The MAE and confidence intervals are computed over the per-frame predictions, allowing for a more granular assessment of model performance and uncertainty. This approach aligns with the nature of rPPG-based methods, which estimate a time-series signal rather than discrete bpm values. As a result, the reported errors reflect deviations in the predicted pulse waveform relative to the ground truth, rather than direct bpm differences.

5.3.1 Conformal Prediction with MAE Nonconformity

5.3.1.1 COHFACE

The results are summarized in Table 5.1, which highlights the effect of different models and alpha values on interval width and coverage probability.

TABLE 5.1: Summary of Conformal Prediction Results on COHFACE

Model	Alpha	Mean Interval Width	MAE	Coverage Probability
DeepPhys	0.1	3.04	0.57	0.89
DeepPhys	0.2	1.51	0.57	0.79
TSCAN	0.1	3.01	0.57	0.89
TSCAN	0.2	1.5	0.57	0.78

The analysis focuses on key metrics, including the mean interval widths, MAE and coverage probability, across different significance levels ($\alpha = 0.1$ and $\alpha = 0.2$). This allows us to examine the reliability and effectiveness of the conformal prediction framework for physiological signal estimation.

Both DeepPhys and TSCAN exhibit nearly identical mean predicted values across all settings. Since the raw values in the dataset represent per-frame pulse signal estimates, this number is not directly interpretable in a physiological sense but the fact that both models have the nearly identical mean suggests that, on average, they are making very similar predictions across all videos. As we mentioned previously, the results are not in bpm, they represent the predicted pulse waveform per frame, which is why the mean appears close to zero. The standard deviation of predictions is also relatively small (0.14 to 0.15), suggesting that both models are stable and consistent in their outputs.

One of the expected trends is the reduction in interval width when moving from $\alpha = 0.1$ to $\alpha = 0.2$. Specifically, for DeepPhys, the mean interval width decreases from 3.04 at $\alpha = 0.1$ to 1.50 at $\alpha = 0.2$. A similar pattern is observed for TSCAN, where the interval width decreases from 3.01 to 1.49. This aligns with the theoretical expectation that a

lower α value results in more conservative (wider) confidence intervals, ensuring that the true values are captured with higher probability.

The MAE values remain nearly identical for both models, with DeepPhys achieving an MAE of 0.57446 and TSCAN achieving 0.57442. While this similarity in performance suggests that both models have comparable prediction accuracy, the confidence intervals provide an additional layer of information about prediction reliability. Specifically, TSCAN consistently produces slightly narrower confidence intervals compared to DeepPhys, which suggests that TSCAN provides slightly more precise predictions, as it captures the true values with tighter bounds. We must distinguish that with the previous statement, we do not suggest that TSCAN has higher accuracy, but that TSCAN assigns lower uncertainty to its predictions.

In Figure 5.1 we visualize the coverage probabilities achieved by DeepPhys and TSCAN on the COHFACE dataset. In this context, coverage probability refers to the proportion of test samples for which the ground truth falls within the predicted conformal interval. This visualization allows us to assess whether the uncertainty intervals behave as expected across different models and significance levels. Here, both DeepPhys and TSCAN achieve coverage probabilities close to the theoretical targets, confirming that the conformal prediction framework is functioning correctly. The slight deviations from the expected values (e.g., slightly below 90% or 80%) are normal and are attributed to finite calibration sample size. This minor discrepancy indicates that while the models are not perfectly calibrated, they remain highly reliable. Overall, the results demonstrate that conformal prediction produces trustworthy confidence intervals for rPPG predictions.

In terms of coverage probability, DeepPhys and TSCAN perform similarly, with slightly better coverage for $\alpha = 0.1$, which is expected due to wider intervals. TSCAN exhibits slightly narrower confidence intervals compared to DeepPhys, indicating that it produces more precise predictions. However, the differences are minimal, further reinforcing the observation that both models are well-calibrated under the conformal prediction framework.

Figure 5.2 presents a visualization of the ground truth pulse signal, predicted signal and CP coverage for DeepPhys and TSCAN on the COHFACE dataset, with a significance level of $\alpha = 0.1$. The blue line represents the predicted signal, while the red dashed line corresponds to the ground truth pulse signal. The shaded region illustrates the conformal prediction interval.

From the figure, we observe that both models generally track the ground truth pulse signal, but with varying degrees of confidence. DeepPhys exhibits smoother predictions but with relatively wider confidence intervals, whereas TSCAN shows slightly narrower intervals, indicating that it assigns lower uncertainty to its predictions. The

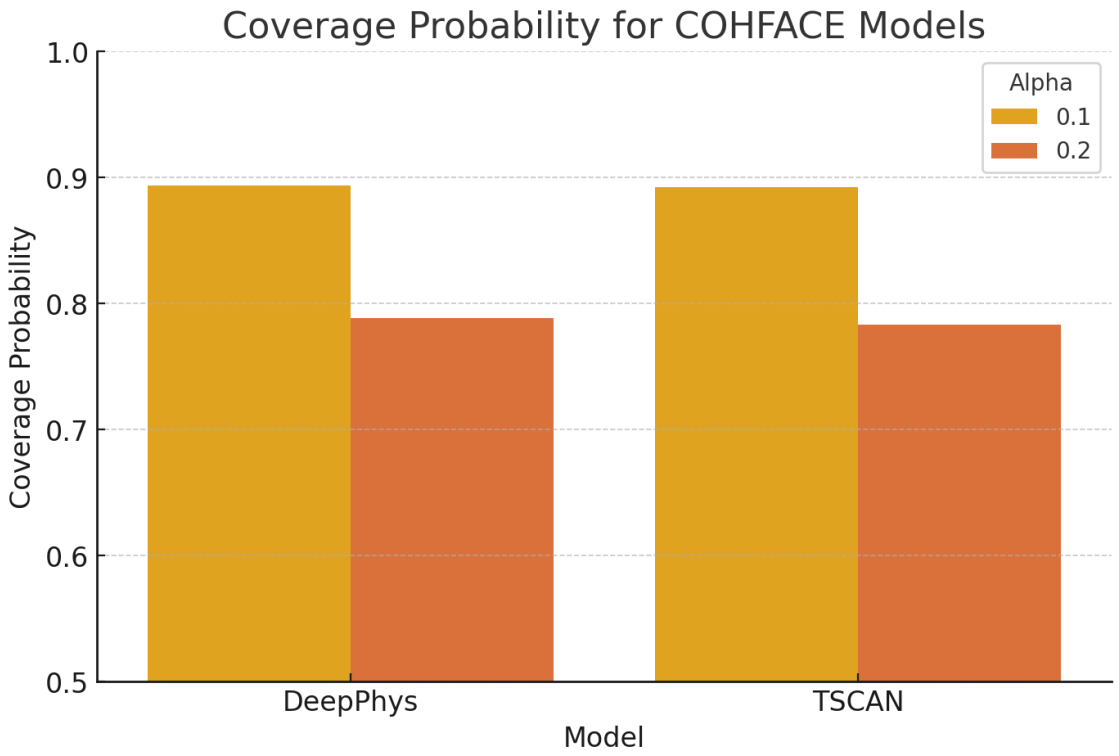


FIGURE 5.1: Coverage probability for COHFACE using MAE as a nonconformity measure for DeepPhys and TSCAN.

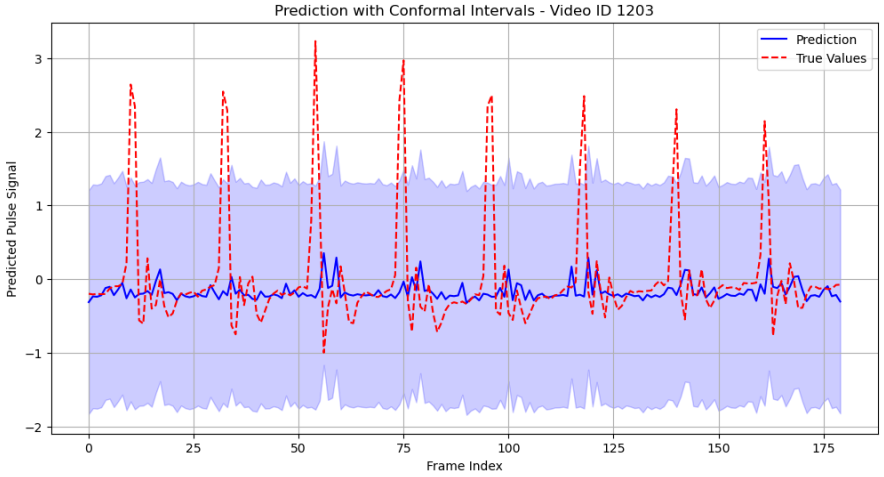
CP intervals successfully capture the ground truth signal most of the time, confirming that the uncertainty estimates are well-calibrated. Differences in model behavior highlight the impact of their respective architectures on performance and robustness to noise.

Overall, this analysis highlights that both DeepPhys and TSCAN achieve similar performance on COHFACE when evaluated using CP. The differences in interval widths and coverage probabilities are minimal, indicating that both models offer reliable uncertainty estimates.

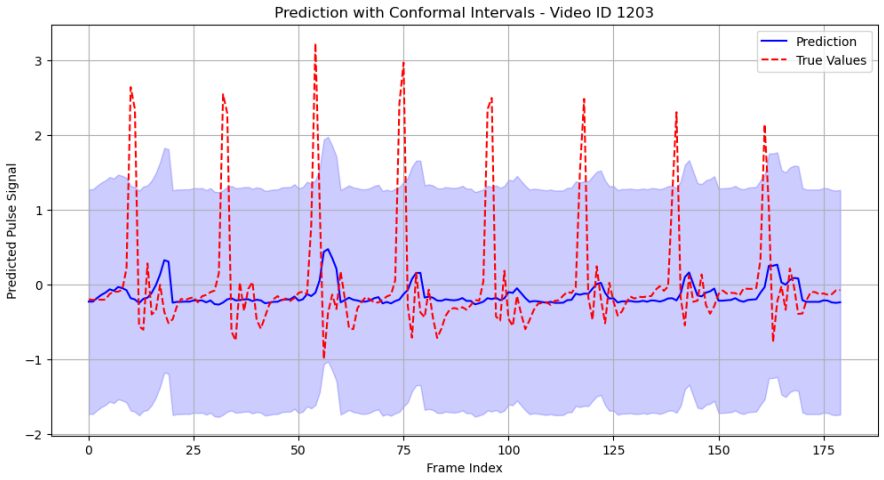
5.3.1.2 UBFC-rPPG

The results for UBFC-rPPG are summarized in Table 5.2, which highlights the effect of different models and alpha values on interval width and coverage probability.

Both DeepPhys and TSCAN exhibit comparable mean predicted values across all settings, aligning with the results observed for COHFACE. The standard deviation of predictions remains relatively small, confirming that both models maintain stable and consistent outputs.



(A) DeepPhys



(B) TSCAN

FIGURE 5.2: Visualization of ground truth pulse signal, predicted signal and CP coverage for DeepPhys and TSCAN for $\alpha = 0.1$ on the COHFACE dataset.

TABLE 5.2: Summary of Conformal Prediction Results on UBFC-rPPG

Model	Alpha	Mean Interval Width	MAE	Coverage Probability
DeepPhys	0.1	2.862	0.605	0.899
DeepPhys	0.2	2.06	0.605	0.803
TSCAN	0.1	2.587	0.543	0.903
TSCAN	0.2	1.809	0.543	0.809

The expected trend of decreasing interval width with increasing α is again observed here. For DeepPhys, the mean interval width decreases from 2.862 at $\alpha = 0.1$ to 2.060 at $\alpha = 0.2$ and a similar reduction occurs for TSCAN, where the interval width decreases from 2.587 to 1.809. Compared to COHFACE, we observe that the intervals for UBFC-rPPG are generally narrower across both models and both α values. This suggests that the models assign lower uncertainty to their predictions on UBFC, potentially due to differences in dataset quality, lighting conditions or participant variability.

The MAE values reveal a slight difference in performance between the two datasets. While DeepPhys maintains a MAE of approximately 0.605 on UBFC, it was slightly lower (0.574) on COHFACE. The same pattern is observed for TSCAN, where its MAE on UBFC-rPPG is 0.543 compared to 0.574 on COHFACE. This suggests that both models achieve better prediction accuracy on UBFC-rPPG than COHFACE. However, the relative performance ranking remains unchanged, TSCAN continues to achieve slightly lower MAE values than DeepPhys.

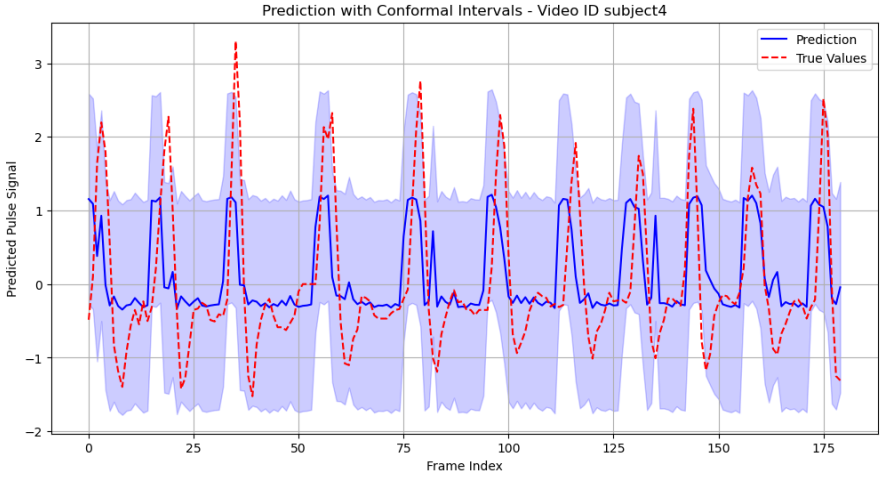
Regarding coverage probability, DeepPhys and TSCAN again perform similarly, with slightly higher coverage at $\alpha = 0.1$ due to the wider intervals. Interestingly, UBFC-rPPG coverage probabilities are slightly higher than those observed on COHFACE, meaning that the CP intervals are more likely to contain the true values in UBFC-rPPG than in COHFACE. This suggests that the CP framework is slightly more conservative on this dataset. This behavior is expected, as UBFC-rPPG is more similar to the dataset on which these models were trained. As a result, both DeepPhys and TSCAN exhibit lower MAE and narrower confidence intervals, indicating better generalization and lower overall uncertainty compared to COHFACE. TSCAN continues to exhibit slightly narrower confidence intervals compared to DeepPhys, indicating that it assigns lower uncertainty to its predictions. However, the differences remain minor, reinforcing the observation that both models are well-calibrated under the conformal prediction framework. Figure 5.3 further supports our observations.

5.3.2 Video Quality as a Nonconformity Measure

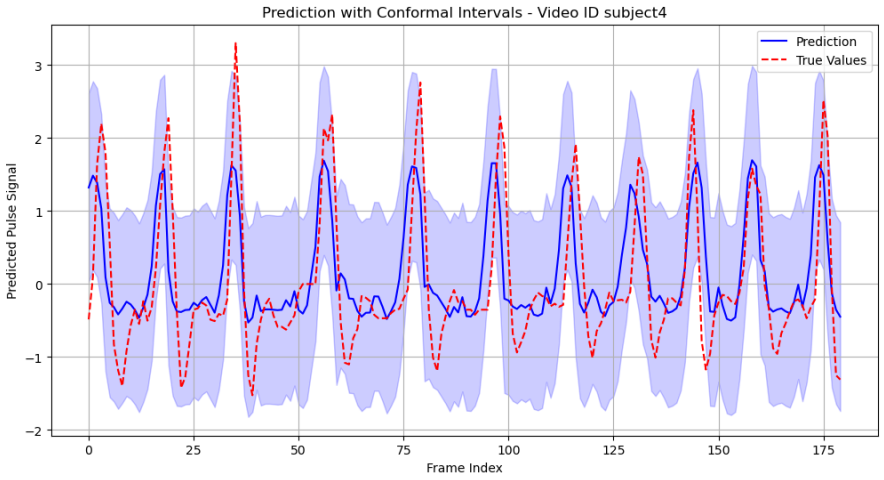
In this section, we analyze the use of ML_QM as a nonconformity measure in CP for rPPG uncertainty estimation. We evaluate its effectiveness by comparing it to MAE-based CP intervals.

ML_QM was derived using ML models trained on video quality factors (motion, illumination, resolution and occlusion among others) to predict expected rPPG error. Since different ML models learn slightly different patterns, we take the average of the predictions from all the five models (Random Forest, XGBoost, Gradient Boosting, Extra Trees and KNN). In the previous chapter, we observed variation in correlation strength between different ML models and MAE across datasets. Some models performed better on UBFC, while others were stronger on COHFACE or MMSE-HR. By averaging, we reduce overfitting to any single dataset. Averaging also ensures no single model dominates, creating a more balanced and robust metric.

In Table 5.3 we analyze both datasets' performance to ensure that ML_QM is an appropriate substitute for MAE in CP. The results for COHFACE and UBFC-rPPG datasets highlight key trends across different models and alpha values. For COHFACE, both DeepPhys and TSCAN exhibit nearly identical MAE values, with DeepPhys achieving



(A) DeepPhys



(B) TSCAN

FIGURE 5.3: Visualization of ground truth pulse signal, predicted signal and CP coverage for DeepPhys and TSCAN for $\alpha = 0.1$ on the UBFC-rPPG dataset.

an MAE of 0.57446 and TSCAN at 0.57442. These results indicate that both models have comparable accuracy in their raw predictions. As expected again, moving from $\alpha = 0.1$ to $\alpha = 0.2$ results in narrower intervals, reducing from approximately 3.04 to 1.51 for DeepPhys and from 3.01 to 1.50 for TSCAN. The coverage probability is slightly higher for $\alpha = 0.1$ compared to $\alpha = 0.2$, further reinforcing that wider intervals capture more ground truth values, albeit at the cost of reduced precision. The similarities in CP results between DeepPhys and TSCAN indicate that both models produce stable and well-calibrated uncertainty estimates.

For UBFC, DeepPhys with $\alpha = 0.1$ shows a slightly higher MAE (0.60499) compared to its performance on COHFACE. This could be attributed to the fact that we are now using ML_QM as a nonconformity measure instead of MAE. Unlike MAE, which directly quantifies prediction error, ML_QM is an indirect estimate of video suitability for rPPG performance. While ML_QM is strongly correlated with MAE, it does not always map

TABLE 5.3: Summary of Conformal Prediction Results on UBFC-rPPG and COHFACE.

Model	Dataset	Alpha	Mean Interval Width	MAE	Coverage Probability
DeepPhys	COHFACE	0.1	3.04	0.57	0.894
DeepPhys	COHFACE	0.2	1.51	0.57	0.789
TSCAN	COHFACE	0.1	3.01	0.57	0.893
TSCAN	COHFACE	0.2	1.5	0.57	0.785
DeepPhys	UBFC-rPPG	0.1	2.86	0.6	0.9
DeepPhys	UBFC-rPPG	0.2	2.06	0.6	0.803
TSCAN	UBFC-rPPG	0.1	2.59	0.54	0.903
TSCAN	UBFC-rPPG	0.2	1.809	0.54	0.809

one-to-one with prediction errors. This means that the conformal prediction framework may assign uncertainty differently based on video quality factors rather than purely on observed prediction deviations.

The interval widths on UBFC-rPPG are slightly narrower than those observed on COHFACE, suggesting that the CP framework is somewhat more confident in its predictions for this dataset. This is also reflected in the higher coverage probability for DeepPhys on UBFC-rPPG (0.8986) compared to COHFACE (0.8937). These results suggest that CP adapts well to dataset characteristics, maintaining stable coverage while reflecting subtle dataset-specific variations in prediction reliability.

The comparison between MAE-based and quality-metric-based CP intervals reveals important insights into the effectiveness of using ML_QM as a nonconformity measure. Across both datasets (COHFACE and UBFC-rPPG), the key patterns observed in interval widths, coverage probabilities and overall calibration provide evidence that ML_QM can serve as a viable substitute for MAE in CP.

The interval widths are slightly larger when using ML_QM compared to MAE, particularly for lower values of alpha. For instance, in COHFACE, DeepPhys at $\alpha = 0.1$ had a mean interval width of 3.04 using MAE, whereas it increased to 3.18 using ML_QM. A similar trend is observed for TSCAN, where the interval width increased from 3.01 (MAE) to 3.15 (ML_QM). This suggests that ML_QM results in marginally more conservative (wider) intervals.

For UBFC-rPPG, the interval width differences are more pronounced. DeepPhys at $\alpha = 0.1$ increased from 2.86 (MAE) to 3.02 (ML_QM) and TSCAN from 2.59 to 2.88. This could indicate that when using ML_QM, the CP intervals incorporate more uncertainty than those based on MAE. Given that ML_QM is an indirect measure of rPPG performance rather than a direct prediction error, the CP framework may assign broader uncertainty to accommodate cases where video quality deteriorates, even if model error does not significantly change.

Despite slightly wider intervals, the quality-metric-based CP intervals achieve comparable, if not slightly better, coverage probabilities relative to MAE. For example, in

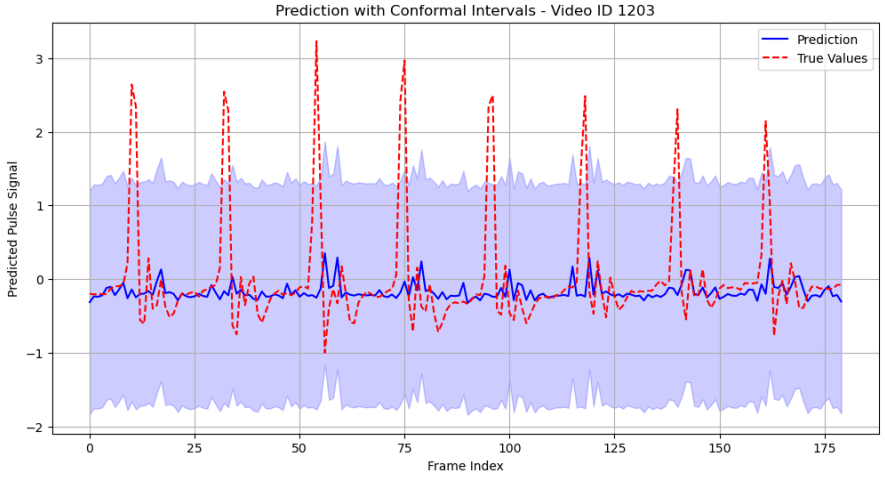
COHFACE, DeepPhys at $\alpha = 0.1$ maintains a coverage probability of 0.89 using both MAE and ML_QM, suggesting that the uncertainty estimates are well-calibrated in both cases. However, for $\alpha = 0.2$, the coverage probability using ML_QM slightly increases from 0.79 (MAE) to 0.81, indicating a small improvement in calibration.

A similar trend is observed in UBFC-rPPG, where TSCAN at $\alpha = 0.1$ increases its coverage from 0.90 (MAE) to 0.91 using ML_QM. This suggests that while the quality-metric-based CP intervals are slightly wider, they provide slightly better coverage, ensuring that more true values fall within the estimated confidence intervals.

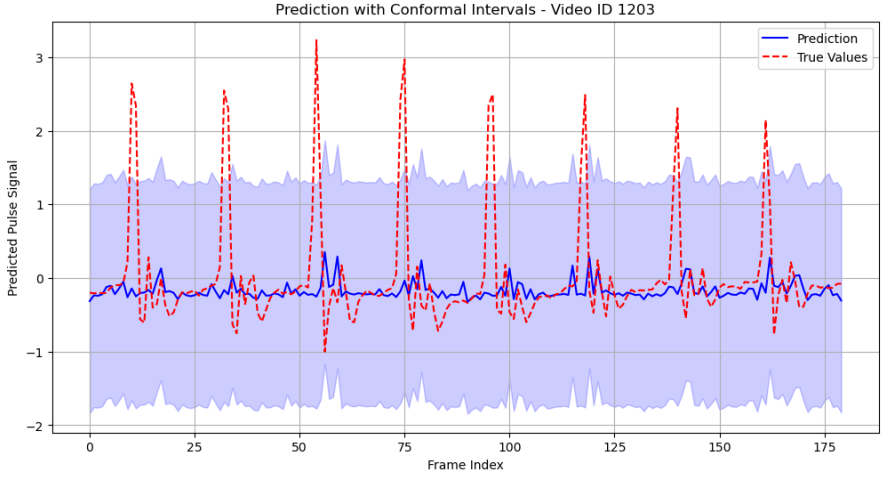
These results suggest that ML_QM, despite not being a direct error measure, effectively captures factors influencing rPPG uncertainty and can serve as a reliable nonconformity measure for CP. The slight increase in interval width indicates that CP accounts for more variability when using ML_QM, likely due to its ability to detect broader video quality degradations that affect model performance. This results in slightly more conservative but well-calibrated confidence intervals, ensuring robust coverage. This is further observed in Figure 5.4. Despite the fact that the images look identical, when we estimate the mean pixel difference between the two images is approximately 0.64. This suggests that the images are nearly identical, with only minimal variations.

Another key observation is that CP using ML_QM maintains dataset-specific trends observed with MAE. For example, UBFC-rPPG continues to exhibit narrower intervals than COHFACE, suggesting that the method adapts well to dataset characteristics regardless of the nonconformity measure. This reinforces the idea that video quality is a strong predictor of rPPG performance and can be effectively integrated into CP frameworks.

These findings validate the feasibility of replacing MAE with a domain-specific quality assessment metric in CP, providing a more interpretable and flexible approach to rPPG uncertainty estimation. We demonstrate that using ML_QM as a nonconformity measure produces prediction intervals that closely match those obtained using MAE. The fact that the coverage probabilities and CP intervals remain nearly identical reinforces the effectiveness of ML_QM in uncertainty estimation. Additionally, the slightly wider intervals observed when using ML_QM suggest a more conservative approach, ensuring robustness in cases where video quality may impact model performance. This confirms that ML_QM is a viable alternative to MAE for conformal prediction in rPPG applications, providing well-calibrated uncertainty estimates while accounting for variations in video quality.



(A) DeepPhys with MAE as a nonconformity measure.



(B) DeepPhys with *ML_QM* as a nonconformity measure.

FIGURE 5.4: Visualization of ground truth pulse signal, predicted signal and CP coverage for DeepPhys for $\alpha = 0.1$ on the COHFACE dataset using the MAE nonconformity and our custom quality metric as a nonconformity measure.

5.4 Discussion

This study presents the first application of CP in rPPG, introducing a framework for quantifying uncertainty in heart rate estimation. Given the challenges of rPPG, such as motion artifacts, lighting variations and unpredictable video conditions, ensuring that predictions are both accurate and well-calibrated is critical for real-world applications. By applying CP to DL rPPG models across datasets, we demonstrate how statistical confidence intervals enhance the interpretability and reliability of heart rate predictions.

Initial results using CP with MAE as the nonconformity measure showed that the framework effectively generated confidence intervals that adapted to different significance levels. As expected, lower significance values resulted in wider intervals with

higher coverage, while higher significance values produced narrower intervals with slightly lower coverage probabilities. Across datasets, the models exhibited similar uncertainty calibration, reinforcing the generalizability of CP for rPPG applications. Conformal predictions with MAE assigned uncertainty intervals based purely on prediction errors, without explicitly incorporating the impact of video quality on rPPG performance. This motivated the integration of a quality-aware nonconformity measure, allowing CP to dynamically adjust uncertainty intervals based on the expected difficulty of heart rate extraction from each video.

By incorporating ML_QM as a nonconformity measure, this study introduces a novel adaptation of conformal prediction that accounts for data variability at a more granular level. The results show that ML_QM maintains reliable uncertainty calibration, producing confidence intervals and coverage probabilities comparable to those obtained with MAE. However, ML_QM leads to slightly wider intervals, reflecting a more conservative estimation of uncertainty that better accounts for variations in video quality. While split-CP assigns a uniform width across test samples, the overall distribution of interval widths expands when using ML_QM, indicating a cautious adjustment to the increased variability in input data. This adaptation ensures that predictions remain robust and trustworthy, especially in more challenging video conditions.

Unlike MAE, which quantifies absolute prediction errors, ML_QM incorporates structured information about video characteristics such as motion artifacts, lighting variations and resolution that influence model confidence. This ensures that CP intervals are not just reflective of statistical error but also provide a meaningful way to assess the conditions under which an rPPG model is more or less confident in its predictions.

From a practical perspective, integrating video quality into CP enables adaptive uncertainty estimation, which is particularly relevant for telemedicine and mobile health applications. rPPG models deployed in real-world environments must handle highly variable conditions, including fluctuations in lighting, camera quality and user movement. By dynamically adjusting confidence intervals based on video quality, this approach ensures that clinicians and end-users receive uncertainty estimates that reflect the reliability of the input data. This can help prevent overconfidence in unreliable predictions, supporting more informed decision-making in remote health monitoring.

While this study focuses on rPPG, the proposed approach has broader implications. Many ML applications rely on video-based predictions where input quality directly impacts model performance. The quality-aware CP framework introduced here could be extended to other domains, such as facial recognition, medical imaging and video-based biometric analysis. In these fields, incorporating data quality into uncertainty estimation could enhance model reliability and transparency, ensuring that confidence intervals adjust dynamically based on real-world conditions rather than static error assumptions.

Overall, this study validates the use of a video quality metric as a nonconformity measure in CP, offering a new method for uncertainty estimation in rPPG. The ability to replace MAE without compromising calibration represents a strong contribution to both conformal prediction research and video-based physiological signal analysis. By improving robustness, interpretability and real-world applicability, this work lays the foundation for future uncertainty-aware video processing techniques, opening new possibilities for adaptive and reliable ML systems in health monitoring and beyond.

Chapter 6

Deployment - Conclusions and Future Work

The final chapter bridges research and application, transforming theoretical advancements into real-world impact.

Remote photoplethysmography has the potential to revolutionize health monitoring by enabling non-contact heart rate measurement using everyday cameras, transforming how we track and understand physiological signals. Unlike traditional contact-based methods like ECG and PPG, which require physical sensors attached to the skin, rPPG operates by detecting subtle variations in skin color caused by blood flow. This ability to extract vital information from a simple video recording paves the way for more seamless, accessible and widespread health monitoring.

One of the most promising applications of rPPG is in telemedicine, where remote consultations and diagnostics are becoming increasingly common. Clinicians can assess patients' heart rates and potential arrhythmias without requiring specialized equipment. It also has the potential to improve elderly care and post-surgical monitoring, allowing patients to recover at home while still being continuously observed for any concerning changes in their vital signs.

Beyond clinical settings, consumer health and fitness applications are rapidly adopting rPPG technology. From smartwatch cameras to mobile health apps, non-contact heart rate tracking can enhance wellness monitoring, stress detection and personalized fitness insights.

rPPG also introduces new possibilities for mental health assessment and human-computer interaction. By analyzing heart rate variability and other physiological signals, rPPG could be integrated into workplace wellness programs, stress monitoring tools or even adaptive learning environments that adjust based on user engagement levels. Non-contact monitoring could help detect signs of driver fatigue or stress, reducing the risk of accidents. Additionally, the integration of rPPG in AR/VR systems could enable more immersive experiences by dynamically adjusting virtual environments based on users' physiological states.

As technology advances, rPPG is set to play a pivotal role in shaping the future of healthcare, promising a world where health monitoring is effortless and truly personalized. While challenges remain in ensuring robustness across diverse populations, improving accuracy under varied conditions and addressing privacy concerns, ongoing research and innovation continue to push the boundaries of what is possible.

6.1 Summary of Findings

This thesis explores multiple aspects of rPPG, from fundamental signal processing techniques to machine learning advancements, video quality considerations and uncertainty quantification. Each chapter contributes to the broader goal of making rPPG more reliable and interpretable.

In chapter 2, we investigated traditional signal processing methods for rPPG, establishing a baseline for pulse estimation. We enhanced performance by applying a particle-based feature tracking approach, improving robustness against motion artifacts. We explored feature clustering with K-Means to optimize computational efficiency while maintaining accuracy. Our findings underscore the inherent challenges of traditional signal processing techniques and set the stage for the adoption of more adaptable, learning-based approaches in later chapters.

In chapter 3, we introduced a novel framework that combines video stabilization, spatiotemporal feature extraction and a two-stage learning approach. This method improved robustness by leveraging machine learning while retaining interpretability, addressing key limitations observed in signal processing methods. By structuring learning in two stages (first filtering unreliable frames and then refining pulse estimation) we demonstrated a significant improvement in accuracy. This approach bridges the gap between traditional methods and deep learning.

In chapter 4, we quantified the impact of video quality factors on rPPG accuracy, introducing tailored metrics to assess video suitability for pulse estimation. This provided

insights into how factors such as motion, resolution and lighting influence model performance. Our findings emphasize the necessity of incorporating video quality considerations into rPPG pipelines to ensure reliable pulse estimation across diverse real-world scenarios.

In chapter 5, we applied conformal predictions to rPPG, generating confidence intervals to assess model reliability. By comparing MAE-based and quality-aware nonconformity measures, we highlighted the impact of different uncertainty estimation approaches on rPPG reliability. By leveraging conformal predictions, we ensured that model predictions were not only accurate but also accompanied by well-calibrated confidence intervals, improving trust and usability in practical deployments.

The findings of this thesis have direct implications for real-world applications, making rPPG more practical for deployment in various settings. The insights gained from signal processing approaches highlight the feasibility of rPPG in low-compute environments, such as mobile devices and embedded systems. Traditional methods provide an efficient alternative to deep learning-based solutions, making rPPG more accessible for remote monitoring. The proposed ST2S-rPPG framework enhances robustness against motion artifacts, making it suitable for applications in telemedicine, where patient movement can degrade signal quality. By incorporating adaptive learning, this approach ensures better generalization across diverse populations. Video quality assessment is critical for real-world deployment, particularly in telehealth and consumer health tracking. The ML-QM metric can be used to filter out low-quality videos, reducing false readings and ensuring more reliable heart rate estimation in everyday applications. By providing confidence intervals for heart rate predictions, conformal predictions enable more informed decision-making in clinical settings. This is particularly beneficial for automated health monitoring systems, where uncertainty quantification can prevent incorrect diagnoses based on unreliable data.

6.2 Limitations and Real-World Deployment

While this thesis presents advancements in rPPG, several challenges remain. Signal processing methods, despite their efficiency and interpretability, struggle significantly with motion artifacts, requiring extensive pre-processing and filtering to achieve stable performance. Additionally, they do not adapt well to varying lighting conditions, skin tones and camera settings, which limits their generalizability. The proposed spatiotemporal deep learning framework improves robustness by leveraging data-driven representations, but its effectiveness is contingent on access to diverse, large-scale datasets for training. Without sufficient variation in training data, the model may fail to generalize across different demographic groups or real-world scenarios. Video quality

metrics, such as WS-QM and ML-QM, provide a structured way to assess input reliability and adapt model behavior accordingly, yet they do not entirely eliminate performance degradation in extreme conditions, such as excessive motion blur, occlusions or low-light environments. Finally, conformal predictions offer a valuable mechanism for quantifying uncertainty, in practical applications, however, real-world conditions introduce dynamic, unpredictable variations potentially leading to miscalibrated confidence intervals. These limitations indicate the need for further refinement of both methodological and practical implementations to ensure robustness in diverse deployment scenarios.

6.2.1 Dataset Quality and Diversity

Dataset diversity is a key determinant of how reliably rPPG models generalise to real-world use. The datasets evaluated in this thesis differ in illumination, motion, camera resolution and acquisition setup, providing a practical benchmark for cross-condition robustness. Evaluating models across multiple datasets without overlapping subjects helps estimate out-of-domain performance, revealing how well methods handle unseen scenarios.

However, most publicly available datasets remain limited in demographic and environmental diversity. Many are collected under controlled lighting, consistent camera settings and with a narrow range of skin tones or age groups. These constraints reduce the validity of rPPG models when deployed in unconstrained environments such as telehealth platforms, smartphones or workplace monitoring systems. To achieve true generalisability, future research must prioritise data inclusivity and standardisation. This can be achieved by curating datasets that span diverse lighting conditions, camera devices, demographics and activity levels. Real-world deployment also requires continuous data auditing, where performance is monitored across subgroups (e.g., by skin tone, gender or motion intensity) to identify and mitigate potential biases before large-scale use.

6.2.2 Real-Time Deployment Feasibility

Although this thesis primarily evaluates accuracy and robustness, the proposed methods are designed with real-time deployment in mind. Each model operates on short temporal segments, enabling predictions to be updated continuously using a rolling window as new frames arrive. This design allows the system to process incoming video streams with minimal latency, suitable for live applications such as remote consultations or fitness tracking.

In practical deployments, achieving real-time performance depends on both algorithmic efficiency and infrastructure. The image-based approaches proposed in this thesis require significantly less computation than end-to-end deep video models, making them feasible for on-device processing where privacy and low latency are critical. Integration into existing pipelines could continuously buffer short segments, perform rPPG estimation and update heart rate and confidence intervals every few seconds. Future engineering work should focus on optimising inference speed, adopting lightweight architectures and exploiting GPU or neural accelerator hardware for embedded deployment.

6.2.3 Practical Deployment Considerations

Translating rPPG research into deployed systems requires addressing a range of operational and design challenges.

- **Quality Assurance:** Integrate the proposed video quality metric as a real-time input filter. The system should flag or reject frames with poor motion, illumination, or blur conditions and prompt users to reposition or improve lighting before analysis.
- **Transparency and Feedback:** Provide users or clinicians with both heart rate estimates and uncertainty intervals from conformal predictions. Flagging outputs with high uncertainty can prevent false readings and improve trust.
- **Privacy and Security:** Perform processing locally whenever possible to avoid unnecessary transfer of raw video. When cloud processing is required, apply strong encryption and access control policies.
- **Hardware and Environment:** Define operational guidelines for camera specifications, frame rates and distance ranges to ensure consistency between devices. Implement adaptive frame sampling to balance latency and power use.
- **Human Oversight:** For clinical or high-stakes applications, incorporate human-in-the-loop review. Continuous performance monitoring across demographic or device categories should form part of the maintenance pipeline.

These practical steps transform the methods proposed in this thesis into a deployable framework, aligning with responsible AI and regulatory standards.

6.2.4 Ethical Implications and Bias

As rPPG technologies move closer to real-world applications, it becomes essential to address the ethical implications associated with their use. While this thesis focuses

on improving the accuracy, quality and robustness of pulse estimation from video, the deployment of such systems inevitably raises questions about privacy, fairness and responsible use. These aspects are fundamental to ensuring that advancements in remote sensing are aligned with broader principles of trust and societal benefit.

rPPG systems rely on facial video data which inherently contain identifiable visual features and sensitive physiological information. When such data are captured or stored, there is potential for privacy violations if appropriate safeguards are not implemented. Ethical deployment therefore requires informed consent, secure data handling, and clear communication regarding how and where data are used and stored. In practical applications, privacy risks can be mitigated through methods such as on-device processing, anonymisation of video frames or feature extraction without storing raw data. Adopting approaches that preserve privacy ensures that user trust is maintained while still allowing meaningful physiological analysis.

However, privacy in the context of rPPG extends beyond the protection of visual identity; it encompasses the safeguarding of biometric and physiological information that can reveal sensitive health or emotional states. Unlike ordinary video data, recordings used for rPPG analysis implicitly contain health-related signals such as heart rate or stress level, which fall under special category data within regulations such as the General Data Protection Regulation (GDPR). As a result, even seemingly simple recordings of faces can expose private medical information if not handled responsibly. Ethical deployment therefore requires a strong privacy framework at every stage of the pipeline, from data collection and storage to model training and inference.

Security measures are equally important. Encryption during storage and transmission, access control and transparent data governance policies are critical to ensuring that sensitive information cannot be misused or identified. Beyond compliance, users must be clearly informed about what is being measured, why it is being measured and how their data are protected. Providing such transparency is central to maintaining user trust, particularly when rPPG systems are deployed in uncontrolled environments such as telehealth platforms, online classrooms or workplace applications.

Another major consideration concerns bias, which can arise from uneven performance across demographic or environmental conditions. rPPG accuracy can be influenced by factors such as skin tone, lighting, camera quality and motion patterns, leading to potential disparities in outcomes. Such biases not only reduce reliability but may also amplify existing inequities when systems are deployed at scale. Addressing this requires the use of diverse and representative datasets, transparent performance reporting and ongoing bias auditing as models evolve. In this thesis, the emphasis on evaluating performance under varied video quality conditions provides a foundation for identifying and mitigating such disparities.

Ultimately, the ethical deployment of rPPG technology depends on transparency, accountability and a commitment to user welfare. Developers and organisations should ensure that these systems are used in appropriate contexts, that limitations are clearly communicated, and that users retain control over when and how their data are processed. By combining technical reliability with ethical awareness, the deployment of rPPG can contribute positively to healthcare, research, and everyday well-being — without compromising fairness or privacy.

6.3 Future Work

Future research can explore more adaptive filtering techniques to mitigate motion artifacts and improve real-time implementation of signal processing-based rPPG. Traditional bandpass filtering and PCA-based denoising could be expanded with deep filtering techniques that dynamically adapt to motion patterns rather than applying static thresholds. These techniques could reduce false heart rate estimations by better preserving physiological signals in noisy conditions. Integrating signal processing methods with lightweight machine learning models could provide a hybrid approach that balances efficiency and robustness. Instead of relying solely on handcrafted feature extraction, a hybrid approach like ST2S-rPPG could first apply traditional signal decomposition techniques (e.g., ICA, CHROM or POS) to extract preliminary pulse signals, which are then refined by lightweight machine learning models. These models could leverage recurrent networks or attention-based mechanisms to capture temporal dependencies, enhancing signal stability without significantly increasing computational cost. The combination of these techniques represents a promising direction, as it would retain the interpretability and efficiency of signal processing while leveraging the adaptability and feature extraction power of machine learning.

In terms of video quality assessment, extending the ML-QM metric by incorporating additional degradation factors such as color distortions, noise levels and compression artifacts could enhance its ability to predict rPPG performance under diverse conditions. Real-world validation across multiple demographic groups and device types would be essential to ensure its reliability. Regarding uncertainty quantification, adaptive conformal prediction methods could dynamically adjust confidence intervals based on real-time feedback from the model's own predictions. For example, instead of using a fixed significance level for all predictions, an adaptive conformal predictions framework could adjust interval widths based on contextual factors such as motion levels or occlusion severity. This could be implemented using meta-learning strategies where the model learns to adjust its uncertainty estimates based on past prediction errors. Future work could also further explore the direct relationship between video quality factors and confidence interval behavior, offering additional insights into the mechanisms driving model uncertainty in rPPG.

Looking further ahead, multi-modal approaches could present a promising avenue for enhancing rPPG reliability. Integrating rPPG with audio analysis could provide richer physiological insights and improve measurement robustness. Audio analysis could capture breathing rate and correlate it with heart rate variability. Large Language Models (LLMs) could provide an additional layer of patient understanding during pulse measurement by analyzing spoken words in conjunction with physiological signals. By integrating real-time speech analysis with rPPG, LLMs could assess emotional state, stress levels or cognitive load, offering deeper insights into a patient's well-being. This multi-modal approach could enhance telemedicine consultations, mental health assessments and personalized healthcare interventions by contextualizing physiological data with verbal expressions. Generative models and synthetic data augmentation have much potential in addressing dataset limitations, allowing for better model generalization without excessive reliance on costly real-world data collection.

As technology advances, rPPG holds immense potential to transform the way we monitor health, bridging the gap between convenience and medical-grade accuracy. Moving forward, critical areas for future research include developing adaptive uncertainty estimation frameworks that dynamically adjust prediction intervals in real time, creating hybrid models that combine traditional signal processing with lightweight deep learning for improved robustness and advancing multi-modal techniques to enhance physiological interpretation. Furthermore, addressing fairness through diverse, large-scale datasets and exploring synthetic data generation with generative models will be essential for achieving generalizable, inclusive rPPG systems.

From enabling remote patient monitoring to enhancing AI-driven diagnostics, rPPG stands at the forefront of a shift towards non-invasive, intelligent healthcare solutions. Continued research into improving robustness, interpretability and deployment readiness will ensure that rPPG reaches its full potential, empowering individuals and medical professionals. With further innovation, rPPG can pave the way for a future where vital sign monitoring is seamless, accessible and more inclusive than ever before.

References

Luqman Qader Abdulrahaman. Two-stage motion artifact reduction algorithm for rPPG signals obtained from facial video recordings. *Arabian Journal for Science and Engineering*, 49(3):2925–2933, 2024.

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3430–3437, 2013.

Fabrizio Bigotti and David Taylor. The pulsilogium of santorio: New light on technology and measurement in early modern medicine. *Societate si politica*, 11(2):53, 2017.

Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9(20):4364, 2019.

Christian Cajavilca and Joseph Varon. Willem einthoven: The development of the human electrocardiogram. *Resuscitation*, 76(3):325–328, 2008.

Rui Cao, Iman Azimi, Fatemeh Sarhaddi, Hannakaisa Niela-Vilen, Anna Axelin, Pasi Liljeberg, and Amir M Rahmani. Accuracy assessment of oura ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: comprehensive analysis. *Journal of Medical Internet Research*, 24 (1):e27487, 2022.

Matthew Charlton, Sophie A Stanley, Zoë Whitman, Victoria Wenn, Timothy J Coats, Mark Sims, and Jonathan P Thompson. The effect of constitutive pigmentation on the measured emissivity of human skin. *Plos one*, 15(11):e0241843, 2020.

Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018.

Elizabeth Craik. *The'Hippocratic'corpus: Content and context*. Routledge, 2014.

Lucas José Sá da Fonseca, Marco Ant, Luíza A Rabelo, et al. Radial applanation tonometry as an adjuvant tool in the noninvasive arterial stiffness and blood pressure assessment. *World Journal of Cardiovascular Diseases*, 2014, 2014.

Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013.

John P DiMarco and John T Philbrick. Use of ambulatory electrocardiographic (holter) monitoring. *Annals of internal medicine*, 113(1):53–68, 1990.

Robert Ellis Dudgeon. *The sphygmograph: its history and use as an aid to diagnosis in ordinary practice*. Baillière, Tindall, and Cox, 1882.

Charles A Elsberg. The edwin smith surgical papyrus: and the diagnosis and treatment of injuries to the skull and spine 5000 years ago. *Annals of Medical History*, 3(3):271, 1931.

Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, and Ruiyi Ma. Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):879–891, 2014.

Litong Feng, Lai-Man Po, Xuyuan Xu, Yuming Li, Chun-Ho Cheung, Kwok-Wai Cheung, and Fang Yuan. Dynamic roi based on k-means for remote photoplethysmography. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1310–1314. IEEE, 2015.

Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

Marius Gade, Kevin Mekhaphan Nguyen, Sol Gedde, and Alvaro Fernandez-Quilez. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. *Insights into Imaging*, 15(1):286, 2024.

Anup Kumar Gupta, Rupesh Kumar, Lokendra Birla, and Puneet Gupta. Radiant: Better rPPG estimation using signal embeddings and transformer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4976–4986, 2023.

Sebastian Hanfland and Michael Paul. Video format dependency of ppgi signals. In *Proceedings of the International Conference on Electrical Engineering*, volume 1, page 2, 2016.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022.

Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

Kokila Bharti Jaiswal and Toshanlal Meenpal. Heart rate estimation network from facial videos using spatiotemporal feature image. *Computers in Biology and Medicine*, 151: 106307, 2022.

Jiaqi Kang, Su Yang, and Weishan Zhang. Transppg: Two-stream transformer for remote heart rate estimate. *CCF Transactions on Pervasive Computing and Interaction*, pages 1–10, 2024.

Adam Kiddle, Helen Barham, Simon Wegerif, and Connie Petronzio. Dynamic region of interest selection in remote photoplethysmography: Proof-of-concept study. *JMIR Formative Research*, 7:e44575, 2023.

Dae-Yeol Kim, Kwangkee Lee, and Chae-Bong Sohn. Assessment of roi selection for facial video-based rPPG. *Sensors*, 21(23):7923, 2021.

Jihoon Kim, Sung-A Chang, and Seung Woo Park. First-in-human study for evaluating the accuracy of smart ring based cuffless blood pressure measurement. *Journal of Korean medical science*, 39(2), 2024.

Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5): 1565–1588, 2015.

Sungjun Kwon, Hyunseok Kim, and Kwang Suk Park. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. In *2012 annual international conference of the IEEE engineering in medicine and biology society*, pages 2174–2177. IEEE, 2012.

Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rPPG: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020.

Georg Lempe, Sebastian Zaunseder, Tom Wirthgen, Stephan Zipser, and Hagen Malberg. Roi selection for remote photoplethysmography. In *Bildverarbeitung für die Medizin 2013: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 3. bis 5. März 2013 in Heidelberg*, pages 99–103. Springer, 2013.

Shuo Li, Mohamed Elgendi, and Carlo Menon. Optimal facial regions for remote heart rate measurement during physical and cognitive activities. *npj Cardiovascular Health*, 1(1):33, 2024.

Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.

Zhihua Li and Lijun Yin. Contactless pulse estimation leveraging pseudo labels and self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20588–20597, 2023.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

Xin Liu, Brian L Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. *arXiv preprint arXiv:2110.04447*, 2021a.

Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. Metaphys: few-shot adaptation for non-contact physiological measurement. In *Proceedings of the conference on health, inference, and learning*, pages 154–163, 2021b.

Xin Liu, Yuntao Wang, Sinan Xie, Xiaoyu Zhang, Zixian Ma, Daniel McDuff, and Shwetak Patel. Mobilephys: Personalized mobile camera-based contactless physiological sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–23, 2022.

Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rPPG-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems*, 36, 2024.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016, 2022.

Magdalena Madej, Jacek Rumiński, Tomasz Kocejko, and Jkedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. 2011.

Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61(10):2593–2601, 2014.

Daniel McDuff, Xin Liu, Javier Hernandez, Erroll Wood, and Tadas Baltrusaitis. Synthetic data for multi-parameter camera-based physiological sensing. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3742–3748. IEEE, 2021.

Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 63–70. IEEE, 2017.

Hamed Monkaresi, Rafael A Calvo, and Hong Yan. A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE journal of biomedical and health informatics*, 18(4):1153–1160, 2013.

Rohini Nanthakumar and Nivethika Sivakumaran. Role of biomedical engineering for diagnose and treatment. *Role of Biomedical Engineering for Diagnose and Treatment.*, 4 (11):94–112, 2018.

Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585. IEEE, 2018.

Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019a.

Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019b.

Vladislav Ostankovich, Geesara Prathap, and Ilya Afanasyev. Towards human pulse rate estimation from face video: automatic component selection and comparison of blind source separation methods. In *2018 International Conference on Intelligent Systems (IS)*, pages 183–189. IEEE, 2018.

Harris Papadopoulos, Efthyvoulos Kyriacou, and Andrew Nicolaides. Unbiased confidence measures for stroke risk estimation based on ultrasound carotid image analysis. *Neural Computing and Applications*, 28:1209–1223, 2017.

Seung-Min Park, Jun-Yeup Kim, Kwang-Eun Ko, In-Hun Jang, and Kwee-Bo Sim. Real-time heart rate monitoring system based on ring-type pulse oximeter sensor. *Journal of Electrical Engineering and Technology*, 8(2):376–384, 2013.

Dung Phan, Lee Yee Siong, Pubudu N Pathirana, and Aruna Seneviratne. Smartwatch: Performance evaluation for long-term heart rate monitoring. In *2015 International symposium on bioelectronics and bioinformatics (ISBB)*, pages 144–147. IEEE, 2015.

Lai-Man Po, Litong Feng, Yuming Li, Xuyuan Xu, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Block-based adaptive roi for remote photoplethysmography. *Multimedia Tools and Applications*, 77:6503–6529, 2018.

Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010a.

Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010b.

Blaine Reeder and Alexandria David. Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of biomedical informatics*, 63:269–276, 2016.

Rene Alejandro Rejon Pina and Chenglong Ma. Classification algorithm for skin color (casco): A new tool to measure skin color in social science research. *Social Science Quarterly*, n/a(n/a). . URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ssqu.13242.

Domenico Ribatti. William harvey and the discovery of the circulation of the blood. *Journal of angiogenesis research*, 1:1–2, 2009.

Ariel Roguin. Rene theophile hyacinthe laënnec (1781–1826): the man behind the stethoscope. *Clinical medicine & research*, 4(3):230–235, 2006.

Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2021.

Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80:72–91, 2008.

Fatemeh Sarhaddi, Kianoosh Kazemi, Iman Azimi, Rui Cao, Hannakaisa Niela-Vilén, Anna Axelin, Pasi Liljeberg, and Amir M Rahmani. A comprehensive accuracy assessment of samsung smartwatch heart rate and heart rate variability. *PloS one*, 17 (12):e0268361, 2022.

Marko Savic and Guoying Zhao. Physu-net: Long temporal context transformer for rPPG with self-supervised pre-training. In *International Conference on Pattern Recognition*, pages 228–243. Springer, 2024.

John W Severinghaus. Takuo aoyagi: discovery of pulse oximetry. *Anesthesia & Analgesia*, 105(6):S1–S4, 2007.

Hang Shao, Lei Luo, Jianjun Qian, Shuo Chen, Chuanfei Hu, and Jian Yang. Tranphys: Spatiotemporal masked transformer steered remote photoplethysmography estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994.

Barry D Silverman. Jean baptiste bouillaud. *Clinical Cardiology*, 19(10):836–837, 1996.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

Rencheng Song, Senle Zhang, Chang Li, Yunfei Zhang, Juan Cheng, and Xun Chen. Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 69(10):7411–7421, 2020.

Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021.

Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network.

Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in neural information processing systems*, 34:6216–6228, 2021.

Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014.

Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

Anson Chui Yan Tang. Review of traditional chinese medicine pulse diagnosis quantification. *Complementary therapies for the contemporary healthcare*, pages 61–80, 2012.

H Emrah Tasli, Amogh Gudi, and Marten Den Uyl. Remote ppg based vital sign measurement using adaptive facial regions. In *2014 IEEE international conference on image processing (ICIP)*, pages 1410–1414. IEEE, 2014.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

Carlo Tomasi and Takeo Kanade. Detection and tracking of point. *Int J Comput Vis*, 9: 137–154, 1991.

Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos

under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Rosemarie Velik. An objective review of the technological developments for radial pulse diagnosis in traditional chinese medicine. *European Journal of Integrative Medicine*, 7(4):321–331, 2015.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

H Kenneth Walker, W Dallas Hall, and J Willis Hurst. Clinical methods: the history, physical, and laboratory examinations. 1990.

Faith Wallis. Signs and senses: diagnosis and prognosis in early medieval pulse and urine texts. *Social history of medicine*, 13(2):265–278, 2000.

Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rPPG. *IEEE transactions on Biomedical Engineering*, 62 (2):415–425, 2014.

Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

Wenchuan Wei, Korosh Vatanparvar, Li Zhu, Jilong Kuang, and Alex Gao. Remote photoplethysmography and heart rate estimation by dynamic region of interest tracking. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3243–3248. IEEE, 2022.

Kwan Long Wong, Jing Wei Chin, Tsz Tai Chan, Ismoil Odinaev, Kristian Suhartono, Kang Tianqu, and Richard HY So. Optimising rPPG signal extraction by exploiting facial surface orientation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2171, 2022.

Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Ming Xu, Guang Zeng, Yongjun Song, Yue Cao, Zeyi Liu, and Xiao He. Ivrr-ppg: An illumination variation robust remote-ppg algorithm for monitoring heart rate of drivers. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10, 2023.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019a.

Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 151–160, 2019b.

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

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023.

Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.