Wearable Multi-wavelength Photoplethysmography Deep Learning Heart Rate Estimation

DANIEL RAY

PhD 2024

Wearable Multi-wavelength Photoplethysmography Deep Learning Heart Rate Estimation

DANIEL RAY

A thesis submitted in fulfilment of the requirements of
Manchester Metropolitan University
for the degree of Doctor of Philosophy

Department of Engineering
Manchester Metropolitan University

2024

Declaration of Authorship

I, DANIEL RAY, declare that this thesis titled, 'Wearable Multi-wavelength Photoplethysmography Deep Learning Heart Rate Estimation' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- 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.
- Where I have consulted the published work of others, this is always clearly attributed.
- 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.
- I have acknowledged all main sources of help.
- 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.

Signed:

Date:

Abstract

Wrist-worn photoplethysmography (PPG) has become a popular method for continuous and remote heart rate monitoring, but single-wavelength PPG faces limitations in accuracy, robustness, and generalisability. This study explores multi-wavelength PPG sensing to enhance heart rate estimation accuracy, robustness, and fairness across diverse populations, particularly for healthcare applications.

A novel dataset comprising 26,442 samples from 20 participants with diverse skin types (Fitzpatrick I-VI) and varying heart rates and motion types was introduced, including blue, green, red, and infrared PPG wavelengths. Additionally, an uncertainty-aware deep learning method was developed for wrist-worn PPG heart rate estimation, optimised for single- and multi-wavelength PPG, using sensor fusion and LOSO cross-validation.

The pilot study analysed the impact of skin melanin, biological sex, and wavelength on PPG heart rate estimation. The blue-green-red-IR combination proved most effective. Significant differences in error distributions across wavelengths were observed for skin melanin and biological sex. High melanin content was associated with higher MAE (8.4 ± 2.1 BPM) compared to low melanin (6.1 ± 2.2 BPM). An uncertainty-aware post-processing method demonstrated competitive performance, mitigating the effects of skin melanin content by equalising the MAE to 3.3 ± 0.9 BPM for high melanin and 3.3 ± 1.3 BPM for low melanin. The method recorded lowest MAE values on three existing single-wavelength datasets—1.3 ± 0.6 BPM on IEEE Train, 1.2 ± 0.4 BPM on BAMI 2, and 2.5 ± 0.9 BPM on PPG DaLiA-compared to existing deep learning methods. For the newly collected multi-wavelength dataset, the method achieved a MAE of 3.3 ± 1.1 BPM.

The pilot study improved reliability through selective rejection of uncertain samples, despite lower retention rates. By investigating multi-wavelength PPG and introducing reliability indicators, this research aims to enhance accuracy and reliability of wrist-worn PPG heart rate monitoring across diverse populations, addressing disparities and improving healthcare applicability. These findings lay groundwork for further research advancing more inclusive and reliable wrist-worn PPG heart rate estimation methods.

Acknowledgements

I extend my deep appreciation to my supervisors, Dr. Tim Collins and Dr. Prasad Ponnapalli, for their guidance throughout my doctoral studies. Their expertise has been instrumental in my academic development. I consider myself privileged to have had them as mentors and will always value their significant role in shaping my research acumen.

My gratitude goes to my co-author, Dr. Sandra Woolley, for her valuable advice and contributions to this research. I also thank Dr. Tiago Pecanha for his expertise when designing the data collection protocol.

I am grateful to my examiners, Professor John Allen and Dr. Huw Lloyd, for their time and insights in reviewing this thesis.

I thank the technical and reception staff at the Institute of Sport for facilitating the laboratory work. Their assistance was key for efficient data collection.

I also want to thank Google Cloud Platform for accepting me into their Google Cloud research credits program, which provided access to Graphics Processing Units essential for this research.

Finally, I extend my appreciation to my friends and family for their unwavering support throughout this journey. I am especially grateful to my parents for their constant encouragement, which has been a cornerstone of my academic and personal growth.

Contents

List of Figures

List of Tables

Abbreviations

AAMI	Association for the Advancement of Medical Instrumentation
AC	Alternating Current
Adam	Adaptive Moment Estimation
AE	Absolute Error
AFE	Analog Front End
BPM	Beats Per Minute
BMI	Body Mass Index
CNN	Convolutional Neural Network
CV	Cross Validation
CVD	Cardiovascular Disease
DC	Direct Current
DNN	Deep Neural Network
ECG	Electrocardiography
HR	Heart Rate
HRV	Heart Rate Variability
IBI	Inter Beat Interval
IR	Infrared
IQR	Interquartile Range
LED	Light Emitting Diode
LOSO	Leave One Subject Out
LRCN	Long-term Recurrent Convolutional Network
LSTM	Long Short Term Memory
NLL	Negative Log Likelihood
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC Dropout	Monte Carlo Dropout
MHR	Maximal Heart Rate
MW PPG HR	Multi-wavelength Photoplethysmography Heart Rate Dataset
PCG	Phonocardiography
PD	Photodiode
PPG	Photoplethysmography
RNN	Recurrent Neural Network
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
SQI	Signal Quality Index
TCN	Temporal Convolutional Network
UV	Ultraviolet

Chapter 1

Introduction

1.1 Monitoring the Cardiovascular System

The cardiovascular system, part of the broader circulatory system, consists of the heart, blood vessels, and blood. Its primary functions are to deliver oxygen, nutrients, and hormones to cells throughout the body while simultaneously removing metabolic waste products, as illustrated in Figure 1.1. The heart, a muscular organ comprised of chambers and valves, pumps blood through various circuits of blood vessels in cycles [1]. Each cycle called a cardiac cycle, consists of two main phases: systole and diastole. During systole, blood is ejected into the arteries from the heart. Conversely, diastole is when blood is returned to the heart in preparation for the next systolic period [2].

The cardiovascular system's function can be assessed through various metrics. Heart rate measures cardiac cycles per minute, while pulse rate, though similar, assesses blood pulses in vessels. Both are counted in beats per minute (BPM). Blood pressure is another key metric, reflecting the force blood exerts on arterial walls during systole and diastole. Oxygen saturation denotes the percentage of oxygen-filled haemoglobin relative to its total capacity. Additional metrics such as stroke volume and cardiac output are critical in assessing cardiovascular functionality [1].

Cardiovascular diseases (CVDs), are a group of disorders affecting the heart and blood vessels. These include conditions like coronary heart disease and stroke, which impair the functionality of the cardiovascular system. Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for nearly one-third of all deaths in 2021 [3]. In England and Wales in 2020, CVDs were responsible for approximately 20% of preventable deaths and half of all treatable deaths. Notably, research suggests that up to 80% of premature CVD-related deaths could be prevented. Furthermore, the economic burden of CVDs is substantial, with an estimated annual cost of £15.8 billion [4].

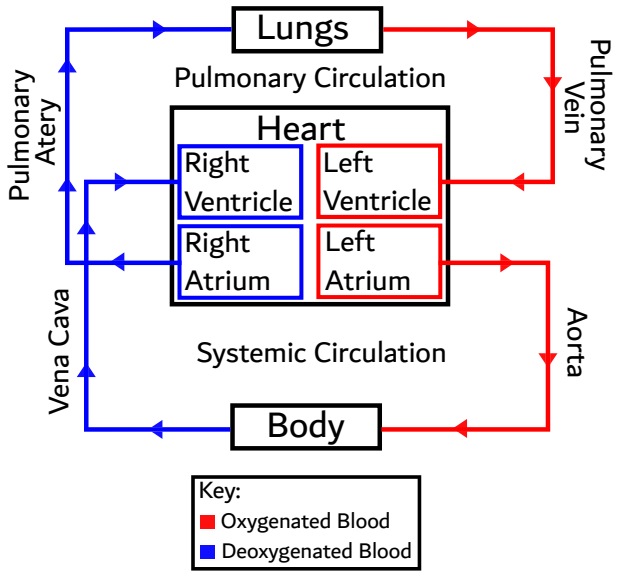


FIGURE 1.1: Block Diagram of Basic Cardiovascular System Functionality. The diagram illustrates the flow of blood through the cardiovascular system, with blue representing de-oxygenated blood and red indicating oxygenated blood. It includes the systemic and pulmonary circuits, depicting the heart, lungs, and body. Blood flows from the body to the heart, then to the lungs for oxygenation, and back to the heart before being pumped throughout the body.

CVDs manifest variably across different populations, influenced significantly by health disparities from behavioural, socioeconomic, psychological, and biological factors [4–6]. In England, South Asian and Black individuals are more vulnerable to CVDs than white individuals [4]. In the USA, racial and ethnic minorities, especially black individuals, confront heightened CVD challenges [5, 6]. They typically face delays in diagnosis and receive inferior care, leading to worse health outcomes than white individuals. Notably, while stroke rates have dropped for white individuals, black individuals are about twice as likely to experience a first stroke, as well as have increased heart failure hospitalisation rates [6].

1.2 Remote, Continuous and Non-invasive Heart Rate Monitoring

The current paradigm of passive treatments at a late stage is advancing towards proactive preventative measures, such as cost-effective, non-invasive and continuous monitoring tools aimed at enabling early and reliable diagnosis and treatment of CVDs and improving patients' quality of life [7–9]. Heart rate monitoring is well-established as an indicator of fitness levels and a training aid for various sports [10]. In healthcare, elevated resting heart rates are an independent marker for mortality and morbidity in

individuals with and without CVDs [11–14], as well as low resting heart rates being associated with CVDs [15].

Electrocardiography (ECG) is considered the 'gold standard' for continuous, non-invasive cardiovascular monitoring. While the single-lead ECG configuration is commonly used for heart rate monitoring, 12-lead ECG remains the cornerstone for comprehensive cardiac assessment, including arrhythmia detection and structural heart disease evaluation. ECG records the depolarisation of the heart's conductive pathway and the related cardiac muscle tissues during each cardiac cycle. Despite its accuracy, conventional multi-lead ECG is not ideal for continuous monitoring due to its lack of portability and convenience. The bio-electrodes used are obtrusive, can't be exposed to water, and require precise placement on the body, connecting to a recording device [7,16]. Additionally, studies indicate that only 50% of nurses and less than 20% of cardiologists correctly place leads V1 and V2, which can result in false diagnoses of myocardial infarction [17].

Various techniques are available for continuous, non-invasive remote heart rate monitoring. In the UK's health care system, it's common to equip at-risk CVD patients with a 3-lead ECG Holter monitor. Though it's more portable than its 12-lead counterpart, it still lacks convenience. While single-lead ECG chest straps used in sports science are more convenient, eliminating the need to administer electrodes, they remain too obtrusive for continuous everyday use. Wearable phonocardiogram (PCG) sensors capture the heart's acoustics and are usually patches worn on the chest. They tackle the issues of convenience and obtrusiveness associated with ECG. However, the PCG signal is often weak and prone to noise interference. Additionally, sensors worn as patches need frequent reapplication [8].

Wrist-worn photoplethysmography (PPG) has emerged as a popular method for continuous, non-invasive heart rate monitoring, driven by the proliferation of smart watches and fitness trackers over the past decade [9,18,19]. PPG is an optical technique that measures blood volume changes in the measurement site's micro-vascular bed [7]. Despite its susceptibility to noise and interference [7], PPG's simplicity and cost-effectiveness — requiring only an LED and photodiode [7] — contribute to its popularity. Additionally, the convenience and unobtrusiveness of wrist-worn sensors, combined with their historical use in timekeeping devices, further enhance their appeal and widespread adoption [18].

1.3 Wrist-worn Photoplethysmography

Commercial smartwatches and wrist-worn fitness trackers, often equipped with PPG sensors, have gained popularity in recent years, with 21% of Americans using commercial smart watches [9, 19] and a projected market value of $96.31 billion by 2027 [19]. Their influence isn't limited to the fitness sector; healthcare has utilised them, too. In fact, over 600 clinical trials involving Fitbit fitness trackers alone are registered on clinicaltrials.gov [20]. Furthermore, evidence suggests that CVD patients who use fitness trackers increase their physical activity [21], and nearly two-thirds more of them meet their desired blood pressure targets [22].

However, known sources of interference and noise affect the accuracy of PPG sensing. Alarmingly, a large body of research reveals that PPG sensing tends to be less accurate for individuals with specific demographic attributes, such as higher concentrations of skin melanin, being biologically female, and having a higher body mass index (BMI) [7, 9, 19, 23, 24]. Additionally, these validation studies often lack a representative sample of individuals with these attributes [7, 9, 25]. Furthermore, the algorithms employed for estimating physiological parameters do not possess the functionality to indicate their failure in providing reliable estimates, nor do they clarify how these parameters are calculated [26, 27]. This lack of transparency and reliability undermines the credibility of these methods in medical settings.

Paradoxically, those most susceptible to CVDs and lacking adequate health care – who would benefit most from such technology – are the ones for whom wrist-worn PPG sensing might be least accurate. Furthermore, wearable fitness trackers and other digital health solutions are under-utilised in low-income and minority communities, with cost and lack of education being significant barriers [9]. This raises concerns about the fairness and reliability of wrist-worn PPG sensing in serving those most in need.

1.4 Aim and Objectives

This thesis aims to develop an accurate, robust and reliable heart rate estimation deep learning method from wrist-worn PPG sensing for a diverse cohort. Therefore, to achieve this aim, the objectives of this thesis are set as follows:

1. **Comprehensive Literature Review on PPG Sensing and PPG heart rate Monitoring:** Conduct a comprehensive literature review on PPG sensing principles and applications, focusing on wearable multi-wavelength PPG sensing and wrist-worn PPG heart rate estimation methods.

2. **Design and Data Collection of Multi-wavelength PPG Dataset:** Based on the findings from Objective 1; develop and acquire a comprehensive multi-wavelength wrist-worn PPG heart rate monitoring dataset. This novel dataset should encompass a diverse participant cohort, considering age, biological sex, BMI, and skin melanin content. It should capture various motion types and intensities as well as include variable heart rate profiles.
3. **Quality Assessment of PPG Signals:** Based on the findings from Objective 1, develop and compare various methods for quantifying and subsequently assessing the quality of the collected PPG signals across various activities and wavelengths.
4. **Development of CNN for PPG Heart Rate Estimation:** Develop a convolutional neural network method for wrist-worn PPG heart rate estimation assessing the performance of existing and collected datasets in generalisability and robustness.
5. **Influence of Wavelength Selection on PPG heart rate Estimation:** Following objective 4, investigate the influence of wavelength selection on the accuracy and robustness of the proposed methodology compared to the conventional green PPG sensing.
6. **Impact of Skin Melanin and Biological Sex on PPG Heart Rate Estimation:** Following objective 4; investigate the influence of skin melanin content and biological sex on the performance of the proposed heart rate estimation method.
7. **Evaluation of Uncertainty Methods in Deep Learning:** Compare and evaluate aleatoric and epistemic uncertainty methods in deep learning, focusing on calibration, their distinctness or entanglement, and their relation to error rates and signal quality.
8. **Development of Post-processing Methods for PPG Heart Rate Estimations:** Following objective 7; develop threshold-based post-processing methods, comparing uncertainty-aware and assumption-based approaches, evaluating the effect on accuracy, robustness, and mitigating the influence of skin melanin content and biological sex.
9. **Comparative Evaluation of PPG Heart Rate Estimation Methods:** Following objectives 4, 7, and 8; compare and evaluate the accuracy and robustness of the proposed methodologies against existing conventional and deep learning PPG heart rate estimation methods.

Building upon the outlined objectives, this thesis will further examine a series of research questions to deepen the understanding and exploration of wrist-worn PPG heart rate estimation techniques:

1. How does the robustness and generalisability of the proposed wrist-worn PPG heart rate estimation method differ across various wavelengths and wavelength combinations, compared to the conventional green light used in consumer wrist-worn smart watches?
2. What is the impact on heart rate estimation performance based on skin melanin content and biological sex in deep learning methods for wrist-worn PPG heart rate estimation?
3. In wrist-worn PPG heart rate estimation, does deep learning demonstrate superior performance compared to conventional methods?
4. What are the most effective methods for estimating uncertainty in deep learning methods for wrist-worn PPG heart rate estimation?
5. How does incorporating uncertainty in post-processing improve the reliability of the proposed wrist-worn PPG heart rate estimation methodology?

1.5 Contributions

The main contributions of this thesis are summarised as follows:

1. A comprehensive literature review on multi-wavelength wearable PPG sensing, encompassing theoretical foundations of PPG principles and skin optics, sources of interference, hardware design considerations, and motion artefact reduction techniques. The review explores various PPG applications and wavelength selection criteria, followed by an in-depth examination of wrist-worn PPG heart rate estimation methods. This includes an analysis of available datasets, signal quality assessment methods, conventional beat detector and heart rate estimation algorithms, and emerging deep learning approaches for wrist-worn PPG heart rate monitoring.
2. A multi-wavelength wrist-worn PPG heart rate monitoring dataset that is comprised of data from 20 participants (13 female, 7 male), aged 26 ± 8 years, with proportionate representation of Fitzpatrick skin types I-VI. It contains 26,442 samples of 8-second windows with 2-second slides, representing nearly 15 hours of data. The dataset features the largest representation of high heart rates (160-180 BPM) among similar available datasets, with a fifth of the dataset indicating physical effort rates of 60% or higher. It includes the most comprehensive collection of PPG wavelengths, with two channels each for blue, green, red, and IR. The data collection protocol incorporates erratic wrist movements, cross-over effects,

motion-free periods, and increased heart rates with minimal motion, providing a robust foundation for evaluating wrist-worn PPG heart rate estimation methods.

3. An uncertainty-aware convolutional neural network for wrist-worn PPG heart rate estimation, optimised for both single- and multi-wavelength PPG sensing, using a sensor fusion architecture with LOSO cross-validation. Aleatoric uncertainty, quantified through distributional predictions strategy, captured data-related uncertainty but remained intertwined with epistemic uncertainty. Three epistemic uncertainty quantification methods were also evaluated, finding Concrete dropout to be the most effective, improving MAE and providing well-calibrated uncertainty estimates across all utilised datasets. Concrete dropout also showed a strong correlation with absolute error and ECG-derived signal-to-noise ratio (SNR) across utilised datasets, enhancing the method's reliability in variable conditions.

4. A comprehensive analysis of the impact of skin melanin content, biological sex, and wavelength selection on wrist-worn PPG heart rate estimation. It identified the blue-green-red-IR wavelength combination as the most effective, reducing MAE by 0.4 BPM compared to green light and improving accuracy by 1.3 BPM during motion-based activities like running. The study revealed significant differences in absolute error distributions across most wavelengths and combinations for both skin melanin content and biological sex. For the most accurate, blue-green-red-IR, wavelength combination high skin melanin content was associated with a MAE of 8.4 ± 2.1 BPM, compared to a MAE of 6.1 ± 2.2 BPM for low skin melanin content— a statistically significant difference.

5. An uncertainty-aware post-processing method demonstrated superior performance, achieving the lowest MAE on three existing single-wavelength wrist-worn PPG heart rate estimation datasets compared to other deep learning methods. It also mitigated the effects of skin melanin content and biological sex, equalising the MAE to 3.3 ± 0.9 BPM for high melanin and 3.3 ± 1.3 BPM for low melanin. The method recorded low MAE values on the existing datasets—1.3 ± 0.6 BPM on IEEE Train, 1.2 ± 0.4 BPM on BAMI 2, and 2.5 ± 0.9 BPM on PPG DaLiA. However, it was less effective on IEEE Test and BAMI 2, with MAE values of 6.6 ± 8.3 BPM and 2.3 ± 1.1 BPM, compared to other deep learning approaches. For the newly collected multi-wavelength dataset, the method achieved a MAE of 3.3 ± 1.1 BPM. By selectively rejecting uncertain samples during post-processing, the method improved reliability but at the cost of lower heart rate estimation retention rates.

Collectively, these contributions represent a meaningful step in the right direction, addressing key challenges and introducing novel methodologies that enhance accuracy, reliability, and fairness in wrist-worn PPG heart rate estimation methodologies.

1.6 Thesis Outline

The thesis is structured as follows: Chapter 2 reviews PPG heart rate estimation, focusing on multi-wavelength and deep learning. Chapter 3 outlines research design and methodology. Chapter 4 discusses the design and collection of the multi-wavelength wrist-worn PPG heart rate estimation dataset. Chapter 5 analyses this dataset for its efficacy and critical insights. Chapter 6 details a convolutional neural network's design, implementation, and heart rate estimation performance analysis. Chapter 7 covers uncertainty quantification and post-processing methods. Chapter 8 concludes the thesis, summarising key findings, limitations, and future research directions.

Chapter 2

Key Concepts and Related Works

This chapter includes a modified version of 'Ray, D., Collins, T., Woolley, S., & Ponnapalli, P. (2023). A Review of Wearable Multi-Wavelength Photoplethysmography. IEEE Reviews in Biomedical Engineering, 16, 136–151. https://doi.org/10.1109/RBME.2021.3121476'

The preceding chapter established the necessity and methodologies for remote, continuous, non-invasive heart rate monitoring, positioning wrist-worn photoplethysmography (PPG) sensing as a promising technique while acknowledging its limitations. This chapter addresses objective 1 of the thesis by offering an exhaustive review of PPG sensing, with an emphasis on multi-wavelength PPG and wrist-worn PPG heart rate monitoring. It begins by detailing the theoretical underpinnings of PPG sensing, including optical interactions and principles. The chapter then details interference sources affecting signal quality and discusses key hardware considerations like sensor geometry, measurement site, and contact force. The section culminates with exploring motion artefacts, mitigation strategies, and diverse applications of PPG sensing, including wavelength selection.

The latter section covers computational methods for wrist-worn PPG heart rate monitoring, examining various conventional approaches. It highlights the need for diverse datasets regarding cohort characteristics and motion types/intensities for method validation. The chapter also discusses the key aspect of signal quality indicators, underlining their importance in assessing the robustness of the developed downstream methods. The chapter concludes with a comprehensive review of various deep learning approaches to PPG heart rate estimation, identifying research gaps and potential areas for investigation, thereby setting the stage for this research.

2.1 Wearable Multi-wavelength Photoplethysmography

Wearable PPG sensing has increased in popularity over recent years as a simple and unobtrusive method to monitor various physiological parameters remotely. However,

showing promise as a tool to advance a proactive approach to healthcare and lifestyle choices, various intricacies and considerations need to be addressed. This section details the principles of PPG sensing, the complex interactions of skin and light, the numerous sources of interferences and the various aspects of signal acquisition for wearable PPG sensing. The section then covers motion artefacts reduction techniques, the selection of the wavelength and the multitude of applications PPG sensing offers.

2.1.1 The Principles and Origin of Photoplethysmography Sensing

PPG is a low-cost, simple and unobtrusive method consisting of a light source and photo-detector. Light is emitted into the skin, and the intensity of light transmitted into the photo-detector will vary depending on the volume of blood in the vascular bed of the measurement site, taking advantage of blood's absorbent qualities to visible and infrared (IR) light.

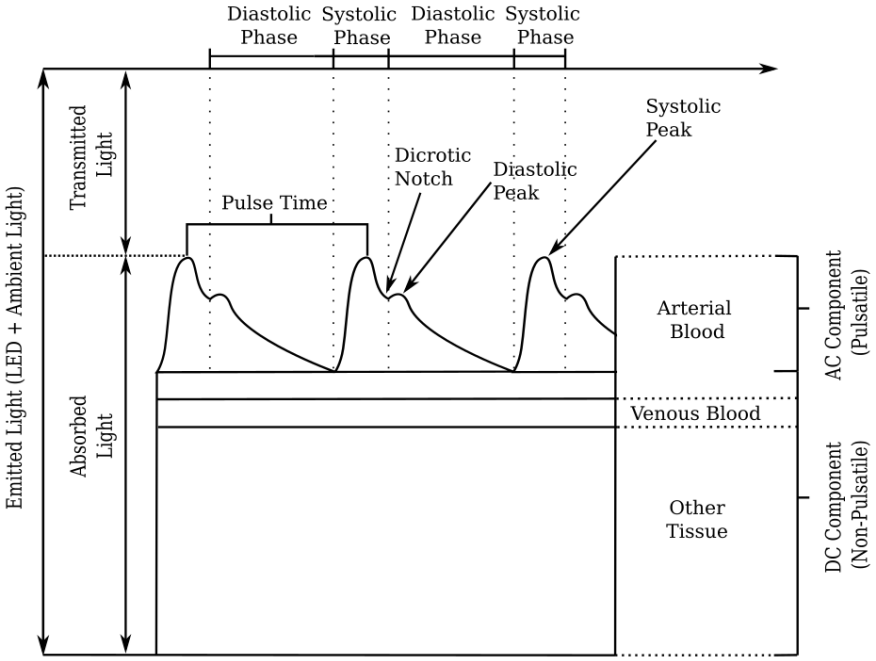


FIGURE 2.1: A typical PPG waveform adapted from Lemay et al. [29, Chapter 2.3]. The PPG waveform is divided into systolic and diastolic phases, showing blood volume changes within vessels. Key waveform features such as the systolic peak, the diastolic peak and the dicrotic notch are shown. The diagram distinguishes between the AC component (pulsatile arterial blood) and the DC component (venous blood and other tissue). Absorbed and transmitted light reflects blood volume changes, with emitted light being a combination of LED and ambient light.

During the contraction of the left ventricle, blood is ejected out of the heart. It propagates along the circulatory system, corresponding to the initial positive slope of a PPG pulse (Figure 2.1). The systolic peak marks the maximum amount of blood in the vascular bed

at the measurement site. The pulse waveform then decreases in amplitude until a local minimum where it transitions into the diastolic phase. The local minimum or dicrotic notch has been traditionally attributed to the closure of the aortic valves [28]. However, an alternative theory suggests it may be related to reflected wave [28]. The mechanism underlying the dicrotic notch remains an active area of research [28]. The end of the diastolic phase marks the closure of the mitral valve and the completion of a cardiac cycle [29]. As well as the AC (Alternating Current) or pulsatile component of the signal, PPG sensing also collects the DC (Direct Current) or non-pulsatile component, which is shaped by respiration, sympathetic nervous system activity, blood pressure control and thermoregulation [16, 28].

There are two modes of PPG sensing with different measurement sites (Figure 2.2). Transmission PPG sensors are usually sited on the fingertip or earlobe, where the light source and detector are separated by tissue. Reflectance PPG sensors, which have both components positioned alongside each other on the same side of the tissue, are commonly sited on the wrist, forehead or torso [16].

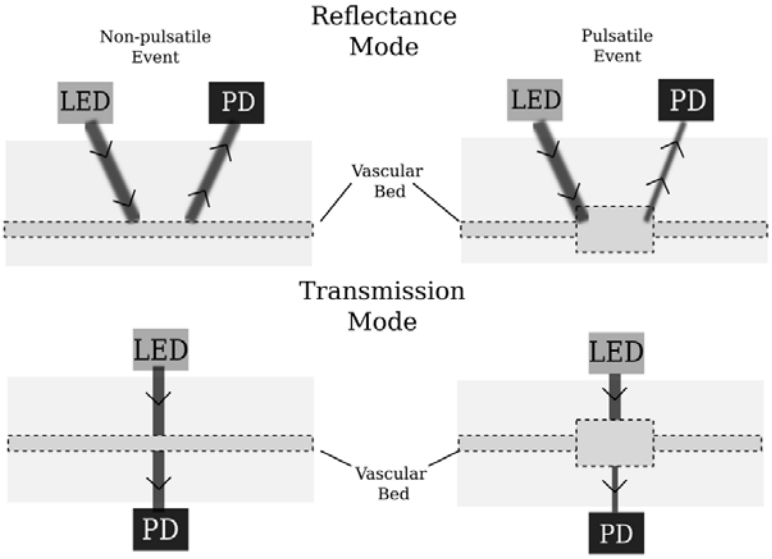


FIGURE 2.2: The two modes of PPG sensing. The diagram illustrates the reflectance (top) and transmission (bottom) modes of PPG sensing. The left column shows minimal blood in the vessel, while the right column shows maximal blood in the vessel. In reflectance mode, the LED emits light that is reflected back to the photodetector (PD) through the skin and blood vessels. In transmission mode, the LED emits light that passes through the skin and blood vessels and is detected by the PD on the opposite side. When there is maximal blood in the vessel, less light is transmitted back to the PD in both modes, compared to when there is minimal blood in the vessel.

A sensing method similar to PPG sensing was first devised in 1936 by two American research groups [31], but Alrick Hertzman established PPG sensing in 1937 [32]. Since

then, with the advancement of semiconductor technologies, transmission mode PPG sensing has been widely adopted in clinical settings for pulse oximetry measurements [24]. Reflectance mode PPG and PPG sensing for other physiological measurements have been gaining popularity in recent years in both commercial and research settings but have not been widely adopted in clinical practice.

2.1.2 Photoplethysmography Skin Optics

Human skin is a complex heterogeneous medium consisting of three main layers: epidermis, dermis and hypodermis (or subcutaneous tissue). The thickness of these layers varies based on the specific body location, adhering to a general pattern [33, 34]. The outermost layer, the epidermis, is composed of multiple sub-layers of both living and non-living cells, with minimal to no blood circulation. The stratum corneum, the nonliving part of the epidermis, is usually about 20 µm in thickness and is made up solely of dead squamous cells [34]. Directly below, the living epidermis has an average thickness of 100 µm and contains the majority of skin pigment compounds, including pheomelanin and eumelanin, collectively known as melanin [33–36].

Located beneath the epidermis is the dermis, which is divided into two primary layers: the papillary dermis, usually about 150 µm thick, and the reticular dermis, with a thickness that typically varies between 1-4 mm based on the region of the body [34]. The papillary dermis is composed of loose connective tissue, which is vascularised by a network of capillaries and small blood vessels typically ranging from 1 to 8 µm in diameter [37]. These vessels exchange materials, such as oxygen and carbon dioxide, between blood and tissue. The reticular dermis is made up of dense connective tissue housing structures such as nerves, glands and hair follicles. Additionally, the reticular dermis contains arterioles and venules, which are slightly larger blood vessels, typically ranging from 2-30 µm in diameter [37], that connect the capillaries to the arteries and veins [33].

The deepest layer of the skin is the hypodermis, which connects the skin to the underlying bones and muscles. Its thickness generally varies from 1-6 mm, contingent on the specific body location [34]. The hypodermis contains larger blood vessels, arteries and veins, typically ranging from 500-5000 µm in diameter [37], which transport blood around the body. The hypodermis is mainly used to store fat and primarily consists of loose connective tissue [33].

Due to the inhomogeneous distribution of blood, cells and pigments in the skin, measuring the optical properties is challenging. Usually, the main optical properties of skin are described as absorption, scattering and penetration depth along with reflection,

transmission and fluorescence [30,31,34,35,38–41]. Researchers have employed several methods to model these properties, such as the radiative transport equation, the Beer-Lambert law, stochastic models like Monte Carlo simulation and random walk, and the adding-doubling method, all with varying results [40,42].

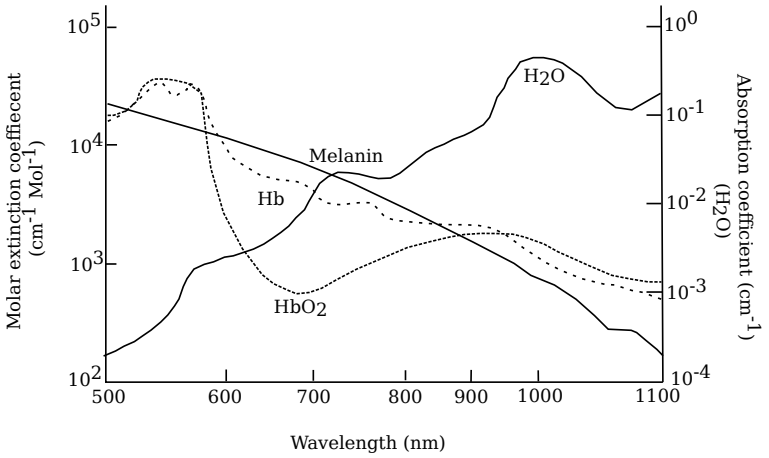


FIGURE 2.3: The light absorption coefficients of biological compounds present in the epidermis-hypodermis layers of skin adapted from Lemay et al. [30, Chapter 2.3]. (Hb - haemoglobin, HbO^2 - oxygenated haemoglobin).

The main light-absorbing components within the skin are water, haemoglobin and melanin; however, each absorbs light differently depending on the wavelength of light and chemical bonding (Figure 2.3). Water, the main component of skin, strongly absorbs IR light (900-1100 nm) but exhibits minimal absorption in the visible light spectrum (390-780 nm) [30,31,34,41,43]. Melanin protects the skin against the sun's harmful ultraviolet (UV) radiation [35]. Its absorption capacity intensifies with decreasing light wavelengths, making it particularly effective at absorbing shorter wavelengths ranging from UV to yellow light (200-600 nm) [16,30,31,34–36,38,39]. Similarly, haemoglobin's absorbing qualities decrease as the wavelength of light increases. However, when chemically bonded with oxygen, its absorbing qualities dramatically reduce when exposed to light in the range of 570-700 nm and is more absorbent to longer wavelengths such as IR when compared to non-oxygenated haemoglobin [30,31,34,35,38,39,41,43].

Scattering in the skin can manifest in two primary ways: as a surface phenomenon like reflection and refraction or as an interaction with skin components that have distinct optical properties. The skin's surface is estimated to reflect about 4-7% of light, regardless of its wavelength [38]. Generally, as the light's wavelength increases, the scattering coefficients within the skin decrease [34,38–41]. Large melanosomes exhibit mainly forward scattering in the epidermis, whilst small "melanin dust" has an isotropic scattering profile. In the dermis, the scattering profile is primarily determined by the fibrous structures of collagen. Meanwhile, in the hypodermis, the primary scatterers

are spherical lipid droplets [34]. Research also indicates that scattering effects are more pronounced in areas like the breast, abdomen, and forehead compared to the arm [41].

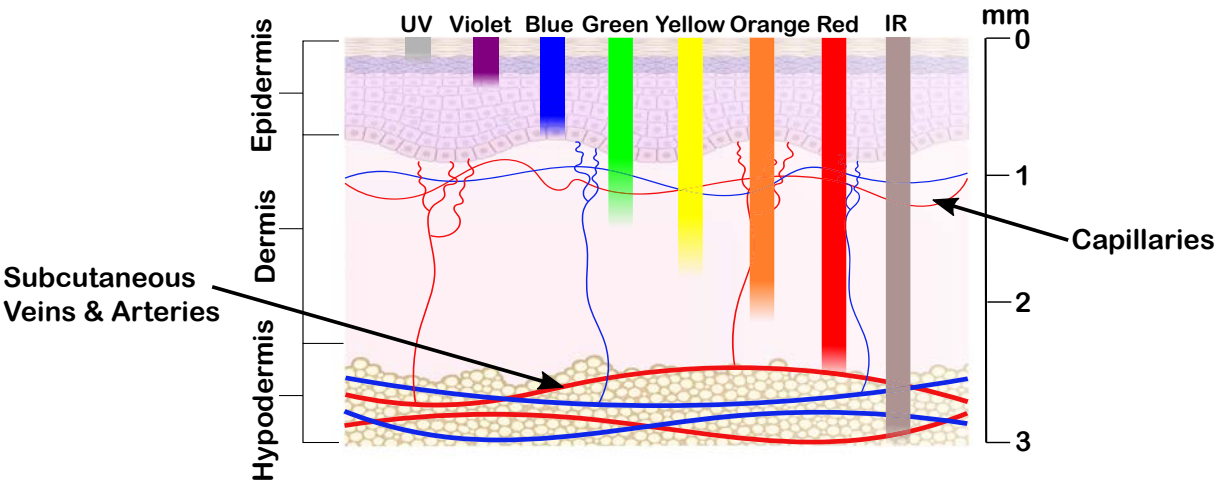


FIGURE 2.4: Approximate maximum penetration depth of each wavelength of light in the skin using reflectance mode sensing. The figure shows the approximate maximum penetration depths of different wavelengths of light in human skin using reflectance mode sensing. The light spectrum ranges from ultraviolet (UV) to infrared (IR). UV light penetrates the shallowest, mainly within the epidermis, while IR light penetrates the deepest, reaching the subcutaneous tissue. The depth scale on the right indicates the penetration in millimetres, illustrating how each wavelength interacts with the skin's layers, from the stratum corneum to the subcutaneous tissue.

In reflectance mode PPG sensing, the trajectory of light within the skin is theorised to follow a "banana-like" shape [44]. The penetration depth, governed by the light's absorption and scattering coefficients in the tissue, is defined as the depth at which the light intensity diminishes to $1/e$ (approximately 37%) of its original surface intensity [40]. Conversely, in transmission mode PPG sensing, the light's path moves directly through the skin, from the Light Emitting Diode (LED) source to the photodiode. Generally, the penetration depth for reflectance mode sensing increases as the wavelength of light increases in the range of visible and near-IR light (Figure 2.4) [16,30,31,34,39,41,43,45–47] with the maximal penetration depth being 3-4mm for IR light (800-1100 nm) [34,41,46,48]. However, when the light's wavelength extends beyond 1250-1400 nm, penetration depth shows a notable decline [34, 41, 48]. The penetration depth in reflectance mode sensing can also vary based on the measurement location. For instance, the breast and abdomen tend to have deeper penetration compared to regions like the arm and forehead [41].

2.1.3 Sources of Interference

Several factors can affect the collected PPG signal's intensity, morphology and noise level, consequently interfering with the measurement of physiological data. These

sources can be categorised as biological, physiological and external, as summarised in Figure 2.5. Beyond these sources of interference, sensor design and configuration can affect the quality of the collected PPG signal. These intricacies and implications are comprehensively discussed in this section and subsequent sections.

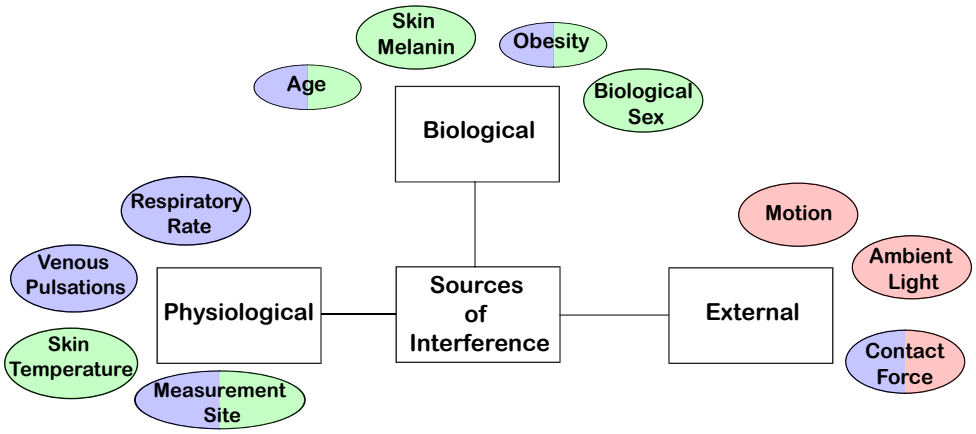


FIGURE 2.5: Sources of Interference in PPG Sensing and Their Impact on the Signal [23]. This figure categories the factors that interfere with PPG sensing into external, biological, and physiological sources, highlighting their specific effects: green indicates changes in signal intensity, blue denotes alterations in signal morphology, and red represents increased noise levels. Note that hardware-related factors, which can also affect PPG sensing, are discussed in Section 2.1.4.

Age, Biological Sex and Obesity

Ageing causes anatomical and physiological shifts, especially in vascular structures, such as arterial thickening and increased stiffness, and non-cardiovascular changes, such as reduced skin thickness. These factors can modify the PPG waveform morphology, reducing the clarity of the cardiac information within the signal [23].

Biological sex has been theorised to contribute to changes in the morphology of the PPG waveform [23]. Differences in heart mass, arterial diameters and stiffness between sexes can alter the amount and pressure of blood flowing through the vascular bed. Beyond cardiovascular distinctions, studies indicate variations in skin thickness between sexes, which may impact the amount of cardiac information in the signal [23]. Shcherbina et al. found that biological male subjects have significantly higher error rates than biological female subjects using several commercial wrist-worn reflectance mode PPG sensing [49]. The impact of biological sex on PPG sensing remains unclear due to limited research.

Obesity, linked to a higher BMI, brings about physiological changes that influence the PPG signal's intensity and quality. Factors such as increased skin thickness, variations in blood flow, oxygen saturation, and capillary density all contribute to the alterations

in the PPG signal. These changes, combined with the individual's metabolic state, body location, gender, skin tone and age, can lead to significant reductions in the quality of the collected PPG signals [23].

Skin Tone

The accuracy and reliability of PPG sensing have been observed to vary across different ethnic groups, with initial studies focusing on the influence of skin tone [7,23,24,49–55]. Pulse oximetry studies first highlighted potential inaccuracies for people with darker skin tones [53]. Patients darker skin tones and low blood oxygen showed up to 10% variation in oximetry readings [56]. In hypoxia, their oxygen saturation was often overestimated [54]. A study of 1609 subjects found black patients had nearly triple the rate of occult hypoxemia compared to white patients [55]. A comprehensive review confirmed this trend, finding that most studies showed decreased oximetry accuracy in patients with darker skin tones [24]. Yet, some studies found no impact on oximetry performance from skin tone [57,58].

Preejith et al. found skin tone significantly impacts green light reflectance mode wrist-worn PPG sensing in non-clinical active settings. Analysing 256 subjects, they found a mean absolute error of 1.04 BPM for subjects with lighter skin tones compared to 10.90 BPM for subjects with darker skin tones when computing heart rate estimations [51]. Shcherbina et al. discovered factors such as darker skin tone, greater wrist size, and elevated BMI were associated with higher heart rate error rates in commercially available reflectance mode wrist-worn PPG devices [49]. However, Bent et al. found no statistically significant differences in heart rate estimation accuracy across skin tones for commercially available wrist-worn reflectance mode PPG devices [59].

The use of green light in many non-clinical PPG systems is motivated by haemoglobin's high absorption spectrum in this range [23]. Fallow et al. examined blue (470 nm), green (520 nm), red (630 nm), and IR (880 nm) reflectance mode wrist-worn PPG sensing, finding green light produces the highest mean modulation at rest for all skin tones but saw a trend towards decreasing mean modulation when increasing skin melanin content. During exercise, they found blue and green wavelengths had higher signal-to-noise (SNR) ratios compared to red or IR [52]. Yen et al. found similar results of green light producing the highest mean modulation across all skin tones using a palm-worn reflectance mode PPG sensor [60]. Contrastingly, Mohapatra et al. found orange (590 nm) PPG to produce increased perfusion index, pulsatile strength, and SNR across all skin tones compared to green (520 nm) wrist-worn reflectance mode PPG sensing. The improved performance was especially noticeable for subjects with darker skin tones,

suggesting that specific wavelengths might be more effective for certain skin tones, especially during physical activity [50].

Skin Temperature

Reduced skin temperature is associated with lower perfusion rates in the vascular bed, a response linked to the Autonomic Nervous System constricting blood vessels in the dermis to conserve body heat [61, 62]. Reductions in skin temperature typically affect the peripheral circulation more than the central areas of the body; for example, when the body is exposed to 10°C ambient temperature, the blood flow in the hand decreases to less than 1 ml/min [62]. All studies exploring temperature and PPG sensing agree that temperature influences the signal [43, 61–66] but to differing degrees.

Ralston et al. posited that skin temperature variations might not cause clinically significant errors in transmission mode PPG sensing [63]. Conversely, Budidha et al. observed that cold exposure significantly reduced the amplitude of the PPG signal in some volunteers, rendering it ineffective for ear-worn reflectance mode PPG sensing [62]. Maeda et al. determined that at temperatures below 15°C, green light (525 nm) PPG heart rate estimates correlated better with ECG heart rate estimates than IR light (880 nm) [65]. In another study, Maeda et al. found that cold exposure reduced the pulsatile component of both green and IR signals. In contrast, hot exposure increased both the pulsatile and non-pulsatile components of the IR signal due to increased blood in peripheral vascular bed [66].

Respiratory Rate and Venous Pulsations

Respiration significantly impacts the non-pulsatile (DC) component of the PPG signal, introducing variations that can affect heart rate measurements [23]. An increase in respiration rate is closely linked to changes in heart rate variability. The PPG signal reflects this by combining cardiac cycle signals with lower-frequency waveforms related to respiration, which primarily originate from the venous system [67].

The venous system's contributions to the PPG signal are often seen as interference but represent a distinct waveform influenced by cardiac, respiratory, and autonomic functions [23]. To reduce venous influences, pressure is sometimes applied at the measurement site, though this can alter the PPG waveform itself [23].

Studies have identified three key types of respiratory-induced PPG variations: intensity, amplitude, and frequency [23, 67]. These variations can modulate the baseline, alter peak amplitudes, and induce phase shifts in the PPG signal. Higher respiration rates tend to reduce these fluctuations [23, 67].

Techniques such as filtering and advanced algorithms are being developed to separate respiratory influences from the PPG signal [23, 67]. Although venous contributions are often considered noise, they can still provide valuable physiological information, but require careful management to avoid distorting the PPG waveform [23].

2.1.4 PPG Hardware Design and Considerations

Over the past decade, there have been significant advancements in multi-wavelength PPG sensing hardware in research settings. The early stages of this technology heavily depended on fibre optics [45, 68]. This then progressed into Optical Electronic Patch Sensor (OEPS) development [60, 69] due to its low cost and simple form factor, with researchers also exploring ear-worn, finger-worn, forehead-worn and wrist-worn PPG sensors [62, 70–72]. The most recent innovation in multi-wavelength PPG sensing hardware is the integration of an on-chip spectrometer, utilising plasmonic filters [73]. This approach has been refined to produce an all-wavelength PPG sensing device [74].

The measurement site of PPG sensing is key due to variations in tissue thickness, skin melanin concentrations, vascular network blood flow, and potential movement at the site [75–80]. Researchers evaluated 52 different measurement locations across the body in a comprehensive study. They determined that the fingers, palms, face, and ears yielded higher amplitude readings for the pulsatile component of the signal in comparison to other sites [77], aligning with further research [75, 80]. However, when examining the effects of motion at various measurement sites, it was found that motion significantly affected the blood distribution in the vascular bed at peripheral measurement sites such as fingers and wrist [75, 78].

Due to the preexisting widespread adoption of wrist-worn devices [18] and their unobtrusive nature, the wrist is the most common measurement site for consumer-grade PPG sensing devices. However, studies indicate that the wrist is not optimal for capturing HR, pulse oximetry, and respiration rate during rest and activity [80], highlighting the need for a more robust methodology. Additionally, researchers have challenged the typical measurement site for wrist-worn PPG sensing devices, suggesting the radial zone, side of the wrist with the thumb, may produce improved signal quality dependent on light wavelengths selected when compared to the central zone of the dorsal surface of the wrist [81, 82].

In the commercial setting, Polar Unite, Grit X and Vantage V2 are the only devices that currently use four wavelengths [83] whilst the other commercial devices have at most three, typically using green light for heart rate measurements and red and IR light for pulse oximetry measurements. While ‘research-grade’ wrist-worn PPG devices like

Empatica E4 and Biovotion Everion (now under Biofourmis Biovitals) offer raw data streams, their heart rate accuracy is reportedly lower than consumer-grade devices. Specifically, Empatica E4 has a mean absolute error of 11.3 BPM at rest and 12.8 BPM during activity. Biovotion Everion's mean absolute error is 16.5 BPM at rest and 19.8 BPM during activity, whereas the Apple Watch has a mean absolute error of 4.4 BPM at rest and 4.6 BPM during activity [59]. This aligns with the findings of Rukasha et al., which reported Empatica E4's heart rate estimate mean absolute percentage errors (MAPE) ranging from 7.2% to 29.2% on a treadmill and 5.3% to 13.5% during 12-hour continuous monitoring [84]. The significant difference in heart rate accuracy between these devices may be attributed to both hardware and software factors. While hardware limitations such as sensor quality and the number of wavelengths likely impact accuracy, it is hypothesised that the methods for processing PPG signals play a key role. It is plausible that Apple's larger user base provides extensive data, allowing for more refined and accurate algorithmic models. Therefore, it is theorised that the superior heart rate accuracy of the Apple Watch is largely due to its advanced data-driven algorithmic processing.

Designing multi-wavelength PPG devices involves multiple considerations, including the number and positioning of LEDs and photo-detectors (PD), LED light intensity, sample rate, contact force, and measures to counter ambient light and electrical noise. Table 2.1 summarises various integrated multi-wavelength PPG sensors that have been developed to bypass these design choices. However, these sensors often lack the flexibility needed for specific research scenarios. Analog Front Ends offers a solution by allowing the creation of a custom sensor module tailored to particular requirements. A summary of these multi-wavelength PPG Analog Front Ends (AFE) can be found in Table 2.2.

In PPG sensing with an AFE, the placement of LEDs and PDs is key for optimal signal strength. For the highest AC/DC ratio, green LEDs should be 1.85 mm from the PD, while red and IR LEDs should be 2.35 mm and 2.75 mm away, respectively [85]. At a 9.75 mm separation, no pulsatile waveform is detected at any wavelength [46], and it was found that nearly double the driving current was needed to obtain a signal at similar distances apart for both red and IR LEDs [76]. While augmenting current and LED count boosts radiation power [46], a small PD active area might not capture this, resulting in no amplitude increase [86]. Expanding the PD's active area or count enhances the signal, with amplitude boosts of 42% for wrist-worn red PPG and 73% for IR. Increasing PD count over LED count is advantageous due to reduced power and heat [76, 86]. Wavelengths should be collected starting with the longest, as pulsations first appear in deeper vessels [46]. An optimal sample rate between 21–64 Hz is recommended to

Device	Wavelength of LEDs				Features
	Blue	**Green**	**Red**	**IR**	
Analog Devices ADPD188GG		2			2 Photodiodes I^2C & SPI Communication 2 external sensor inputs 3 LED drivers Ambient Light Rejection
Analog Devices ADPD144RI			1	1	I^2C Communication External LED emitters Ambient Light Rejection
Maxim Integrated MAX30101		1	1	1	I^2C Communication Ambient Light Rejection
Maxim Integrated MAX86150			1	1	I^2C Communication Ambient Light Rejection Electrocardiogram
Maxim Integrated MAX86916	1	1	1	1	I^2C Communication Ambient Light Rejection
OSRAM SFH 7072		2	1	1	Light Barrier to block optical cross-talk Requires AFE
OSRAM SFH 7050		1	1	1	Light Barrier to block optical cross-talk Requires AFE

TABLE 2.1: Summary of Multi-wavelength Photoplethysmography Integrated Sensing Units. Search carried out in 2022. The table lists various PPG sensing units from different manufacturers, highlighting the wavelength of LEDs (Blue, Green, Red, IR), key features, and communication protocols.

compress data and minimise storage efficiently [87].

Contact force is pivotal in PPG sensing [76, 77, 82, 88]. As sensor contact force rises, the pulsatile signal component's amplitude increases until the transmural pressure (difference between external and intra-arterial pressures) becomes zero. Beyond this point, the pulsatile amplitude diminishes with increasing external pressure until arterial walls flatten, halting circulation [77, 82, 88]. For the wrist in reflectance mode, an optimal contact pressure of 80mmHg is suggested for red light [82]. On the upper arm, a 30mmHg pressure yields the highest amplitude for green and IR light in reflectance mode [77]. Minimal contact pressure is required for the forehead in reflectance mode [76].

Device	Drivers	Features
Analog Devices ADPD4000/4001 ADPD4100/4101	8 LED drivers 8 Inputs for PPG, ECG, EDA, impedance and temperature	I^2C & SPI Communication Ambient Light Rejection
Maxim Integrated MAX30110	2 LED 1 Photodiode	SPI Communication Ambient Light Rejection
Maxim Integrated MAXM86146	3 LED 2 Integrated Photodiode	SPI Communication Ambient Light Rejection Integrated Micro Controller
Texas Instruments AFE4950 AFE44S30	8 LED 4 Photodiode	1/2/3 Lead ECG (AFE4950) I^2C & SPI Communication Ambient Light Rejection
Texas Instruments AFE4900	4 LED 3 Photodiode	1 Lead ECG I^2C & SPI Communication Ambient Light Rejection
Texas Instruments AFE4404	3 LED 1 Photodiode	I^2C Communication Ambient Light Rejection

TABLE 2.2: Summary of Multi-wavelength Photoplethysmography Analog Front Ends. Search carried out in 2022. The table provides a comparison of various PPG analog front-end devices, detailing the number of LED drivers and photodiodes, as well as key features such as communication protocols, ambient light rejection, and additional capabilities like ECG and impedance measurements.

2.1.5 Motion Artefact Reduction

Motion artefacts significantly impact the accuracy of PPG sensing. Motion artefacts distortions in the PPG signal arise from body movement and the varying light penetration depths depending on sensor placement. These artefacts can be periodic or non-periodic and often have larger amplitudes than the signal's pulsatile component [72, 89]. Blanos et al. showed that green (525 nm) and orange (590 nm) light were less affected by motion artefacts than red light (650 nm) [69]. Matsumura et al. concurred, noting a higher SNR ratio for green (530 nm) and blue (470 nm) light compared to red (640 nm) during motion [90]. Shorter wavelengths, like green and blue, offer better SNRs due to their penetration depths and in-vivo optical path lengths making them less prone to motion noise [72]. They also experience less attenuation from optical processes and capture less noise from deeper tissues, like bone movement [69]. However, some shorter wavelengths due to shallow penetration depths do not reveal much cardiac activity [72].

The typical frequency range of a PPG signal is 0-4 Hz, whilst motion artefacts fall within 0-10 Hz, making the removal of motion artefacts challenging. Many methods use a motion reference signal, often collected from accelerometers or gyroscopes [89]. Conversely, researchers have utilised multi-wavelength PPG as a means for motion artefact reduction. For example, Wang et al. utilised the isobestic wavelength (800

nm) as a motion reference and applied a noise-cancelling algorithm to refine the PPG signal [91]. Similarly, Zhang et al. used an IR (940 nm) PPG signal for motion reference, leveraging its deep penetration and motion sensitivity. They used a wavelet transform for signal cleaning and reconstruction, reducing heart rate estimation errors to less than 2 BPM for all motion types [89].

Yao et al. developed a method to separate motion artefacts from PPG signals using an algorithm based on the Beer-Lambert law, which utilised red (660 nm) and two IR (850 and 940 nm) wavelengths [42]. Chang et al. applied a maximal-ratio combined algorithm to 15 PPG signals, achieving a 50% reduction in variations relative to a single-wavelength reference sensor [73]. Chen et al. implemented a similar algorithm on an all-wavelength wrist-worn PPG device, revealing a superior SNR compared to single-wavelength [74]. Lee et al. developed a motion artefact reduction algorithm using 12-channel PPG signals with green (530 nm), red (660 nm) and IR (940 nm) wavelengths. Using a two-step analysis method—first independent component analysis, then principal component analysis with a truncated singular value decomposition approach—the method showed impressive performance in high-motion scenarios. It achieved 82.49% sensitivity (correctly identifying true positives), 99.83% positive predictive value (accuracy of positive predictions), and a very low 0.17% false detection rate (incorrect identifications) [72].

2.1.6 PPG Applications and Wavelength Selection

PPG offers diverse physiological measurements and clinical applications [31, 67]. These include vital signs such as heart rate [92–95], Blood Oxygen Saturation [96], Respiration Rate [97], Blood Pressure [98] and Heart Rate Variability (HRV) [99] as well as clinical insights to Hypertension [100], Atrial Fibrillation [29], Vascular Aging and Atherosclerosis [16, 101], Coronary Heart Disease [102] and Cardiovascular risk [67].

Beyond cardiovascular-related monitoring, PPG sensing has seen several developments, including the detection and monitoring of epileptic seizures [103], diagnosis of respiration diseases [104], monitoring of infectious diseases [66], mental stress and affect recognition [105, 106], monitoring of sleep conditions [107, 108], estimation of blood glucose [109], and medicinal drug delivery monitoring [110–112]. This highlights PPG sensing as a cost-effective continuous monitoring tool to advance a more proactive approach to healthcare and lifestyle choices and establish a technological approach to improving healthcare equity.

The choice to utilise green light in many commercial single-wavelength PPG devices is

due to its optimal light-tissue interactions. Green light is highly absorbent to haemoglobin and penetrates deep enough to sense blood pulsations but not too deep to collect additional physiological information and noise. Nevertheless, alternate light wavelengths have exhibited enhanced signal quality in specific circumstances. This underscores the potential of multi-wavelength approaches to improving the accuracy, robustness and generalisability of PPG sensing [7, 113].

The most common application for multi-wavelength PPG sensing is pulse oximetry, which requires two wavelengths to calculate blood oxygen saturation levels. The blood oxygen saturation level can be estimated from the ratio of pulsatile and non-pulsatile components of each wavelength [96]. Typically, the wavelengths used are red (622–780 nm) and IR (780-2400 nm) [73]; however, researchers have identified orange and green light to perform better due to their robustness to motion artefacts [69, 114]. For blood pressure estimation, multi-wavelength approaches consistently outperform single-wavelength methods [73, 115, 116]. This superiority is especially evident when harnessing the distinct interactions of various wavelengths with skin and blood [73]. In blood glucose estimation, multi-wavelength usage has been linked to reduced error rates [117, 118]. Additionally, multi-wavelength PPG introduces novel applications such as medicinal drug delivery monitoring [110–112].

2.2 Photoplethysmography Heart Rate Monitoring

The effectiveness of heart rate monitoring through PPG hinges on both the methodology employed for heart rate estimation and the quality of the signal acquired. This section examines the datasets utilised for validating heart rate estimation algorithms, emphasising the prerequisites for such data and the pivotal design considerations. Additionally, various methodologies for assessing the quality of the signals collected. The section then examines two distinct conventional approaches for heart rate estimations, clarifying their strengths and weaknesses and the diverse ways they have been implemented in existing literature. The section concludes with an exhaustive review of deep learning methods for heart rate estimation from PPG signals.

2.2.1 Wrist-worn Heart Rate Monitoring Datasets

As of early 2024, more than 30 PPG research datasets are available, covering diverse applications such as blood pressure monitoring, cardiovascular disease detection, emotion detection, heart rate monitoring, pulse oximetry, and respiratory monitoring [67]. The largest dataset is the UK Biobank, with over 200,000 participants [119].

Every dataset is characterised by three fundamental elements that determine its use case:

1. **Participants:** This aspect encompasses the demographic and physical details of the cohort from which the data was gathered. It includes attributes like age, biological sex, weight, height, skin type, health condition, and, in specific scenarios, even the species.
2. **Protocol:** This component outlines the environment and conditions under which the data was collected. It specifies whether the setting was a hospital, laboratory, or a more naturalistic environment. Additionally, it describes the tasks or activities that participants were engaged in during the data collection process.
3. **Devices:** This facet provides insights into the kind of bio-signals recorded. It details the measurement site of the device, the variety of distinct signals each device captured, including channels, axes or wavelengths, and the resolution and sample rate of each signal.

Wrist-worn heart rate monitoring datasets with the intended use of validating heart rate estimation algorithms typically have a chest-worn electrocardiogram (ECG) and a wrist-worn PPG. The ECG serves as the reference device, providing "ground truth" heart rate values extracted over designated time intervals; without validation against an ECG, the method would measure pulse rate instead of heart rate. A motion reference is typically included from a triaxial accelerometer or gyroscope. The protocol typically involves a series of activities with varying levels of intensity collected in either a laboratory or naturalistic setting. Summarised in Table 2.3 are the datasets that fulfil these criteria.

Notably, some heart rate estimation algorithms validate their methodology using emotion detection datasets such as WESAD [120] and CLAS [121], respiratory monitoring datasets such as BIDMC [122] and CapnoBase [123], as well as hospital setting datasets such as MIMIC PERform [124]. While these datasets provide the essential signals for validation, their protocols are specifically designed for different applications. A key feature of wrist-worn heart rate monitoring datasets is to encompass known sources of interferences that evaluate the methodology's robustness, such as diverse motion types, varying motion intensities, a broad spectrum of heart rate values and a diverse cohort in terms of age, biological sex, BMI and skin type [125].

Most wrist-worn heart rate monitoring datasets employ laboratory-based protocols, typically on a treadmill with varying speeds [126–128]. This protocol strategy captures varying motion intensities and a range of heart rate values from increasing and decreasing the workload. However, the protocol may only capture periodic motion types due to the cyclical nature of running in a controlled environment. Whilst beneficial

in exploring the 'crossover effect' of having a similar movement cadence to cardiac activity [125], the protocol lacks diversity.

Treadmill-based protocols generally capture two primary scenarios: elevated heart rates associated with high motion intensities (when running) and lower heart rates linked with minimal motion (when walking or at rest). Capturing scenarios of elevated heart rates with low motion intensities can be achieved using an ergometer [129] or getting participants to hold onto the treadmill bar (BAMI-2) [127]. Lower heart rates linked with high motion intensities and aperiodic motion types can be captured via arm and wrist movements (IEEE Test) [126]. Interestingly, research suggests that the motion type rather than the activity intensity has more impact on the signal quality. Changes in activity and erratic wrist movements were found to cause more inaccuracies than prolonged elevations in motion intensity from running and cycling [125].

An alternative approach to protocol design is to select activities that are performed daily aimed at collecting realistic motion types and intensities. PPG DaLiA was the only dataset to employ a naturalistic protocol incorporating activities with low (driving), medium (walking), and high-intensity arm movements (table soccer), as well as a mix of periodic (walking) and aperiodic motion (eating). Additionally, tasks with differing physical demands (driving vs. ascending stairs) were chosen to induce varied heart rates [130].

Regarding devices, wrist-worn triaxial accelerometers are standard across datasets with BAMI and Casson et al., also including wrist-worn triaxial gyroscopes [127, 129]. ECG choices varied across the datasets; IEEE Train/Test and Casson et al. elected a single-lead ECG [126, 129], PPG DaLiA used a three-lead ECG [130], and BAMI 1 and 2 used a 24-hour Holter Monitor [127]. The accuracy of the ECG device is paramount, as any inaccuracies in the "ground truth" values can inadvertently be reflected in the subsequent heart rate estimation algorithms. For PPG sensor configurations, Casson et al. collect one green (510 nm) channel [129], IEEE Train and Test collects two green (515 nm) PPG channels [126], whilst BAMI 1 & 2 collects three green (525 nm) PPG channels [127]. Both PPG DaLiA and DWL collect multi-wavelength PPG signals [128], [130]; however, PPG DaLiA uses an Empatica E4, which uses green and red LEDs to produce a single PPG signal [130]. On the other hand, DWL used a single PPG channel for blue (undefined), green (520 nm) and IR (940 nm) [128].

Examining the cohorts of wrist-worn heart rate monitoring datasets, there is a noticeable inconsistency in reporting. While age and biological sex are generally reported, exceptions exist, such as in DWL and IEEE Train [126, 129]. Only IEEE Test and PPG DaLiA provided both height and weight, which are indicative of BMI [126, 130]. Uniquely, PPG DaLiA reported skin type using the Fitzpatrick scale [130]. It is key to report the

demographic and physical details of the cohort consistently. This ensures a comprehensive evaluation of heart rate estimation algorithms, especially concerning potential interferences like skin melanin content, obesity, and biological sex [7,23]. Equally key is ensuring diversity in these details. For instance, while PPG DaLiA reported skin types, it had no representation for types 1, 5, and 6 of the Fitzpatrick skin type scale [130]. Following the data collection and reporting guidelines outlined in [125] and best practices outlined in [131] is recommended.

Dataset	Participants	Protocol	Devices
IEEE Train 2015 [126, 132,133]	*12 subjects* **Biological Sex:** 12 Male **Age:** 18 - 35 years **Weight:** Unreported **Height:** Unreported **Skin Types:** Unreported	*Laboratory-Based Protocol on Treadmill* **Protocol 1:** Rest (0.5 min), 8 km/h (1 min), 15 km/h (1 min), 8 km/h (1 min), 15 km/h (1 min), Rest (0.5 min) **Protocol 2:** Rest (0.5 min), 6 km/h (1 min), 12 km/h (1 min), 6 km/h (1 min), 12 km/h (1 min), Rest (0.5 min)	**Accelerometer:** Three-axis wrist-worn **Electrocardiogram:** One-channel using wet ECG sensors. **Photoplethysmogram:** Two channels. LED: Green - 515 nm **Data:** All signals (125 Hz).
IEEE Test 2015 [126, 132,133]	*8 subjects* **Biological Sex:** 7 Male, 1 Female **Age:** 25.9 ± 13.4 years **Weight:** 66.9 ± 7.9 kg **Height:** 172.9 ± 10.4 cm **Skin Types:** Unreported	*Laboratory-based protocol with various arm movements.* **Protocol 1:** various forearm and upper arm exercises, running, jumping, and push-ups. **Protocol 2:** intensive forearm and upper arm movements (e.g. boxing).	**Accelerometer:** Three-axis wrist-worn **Electrocardiogram:** One-channel using wet ECG sensors. **Photoplethysmogram:** Two channels. LED: Green - 515 nm **Data:** All signals (125 Hz). Data was transmitted via Bluetooth.
Casson et al. 2016 [129,134]	*8 subjects* **Biological Sex:** 3 Male, 5 Female **Age:** 22-32 years (mean: 26.5) **Weight:** Unreported **Height:** Unreported **Skin Types:** Unreported	*Laboratory-Based Protocol on Treadmill and Ergometer* Participants were asked to complete one or more of four exercises for up to 10 minutes, setting the pace themselves. Exercises Walk, Run, Low Resistance Ergometer and High Resistance Ergometer.	**Accelerometer:** Wrist-worn Shimmer 3 GSR+ unit. **Electrocardiogram:** Actiwave recorder - Electrodes are positioned on either side of the heart. **Photoplethysmography:** Single Channel. LED: Green - 510 nm. **Gyroscope:** Wrist-worn Shimmer 3 GSR+ unit. **Data:** All signals were sampled at 256 Hz.

Table 2.3 continued from previous page

Dataset	Participants	Protocol	Devices
BAMI-1 2020 [127, 135]	*25 subjects* **Biological Sex:** 10 Male, 14 Female **Age:** 26.9 ± 4.8 years **Weight:** Unreported **Height:** Unreported **Skin Types:** Unreported	*Laboratory-Based Protocol on Treadmill* Rest (1 min), 2.5 km/h (2 mins), 6 km/h (3 mins), 3 km/h (2 mins), 7 km/h (3 mins) 2.5 km/h (2 mins), Rest (1 min)	**Accelerometer & Gyroscope:** 3-axis using a 6-axis inertial measurement unit **Photoplethysmogram:** Three channel LED: Green (525 nm), 1 LED on either side of each photodiode. Three PPG sensors were placed 6 mm apart. Photodiodes: 3 photodiodes **Electrocardiogram:** 24-h Holter monitor **Data:** ECG (125 Hz), all other signals (50 Hz).
BAMI-2 2020 [127, 135]	*23 subjects* **Biological Sex:** 17 Male, 4 Female **Age:** 22.0 ± 1.7 years **Weight:** Unreported **Height:** Unreported **Skin Types:** Unreported	*Laboratory-Based Protocol on Treadmill* Rest (1 min), 3.5 km/h (2 mins), 7 km/h (2 mins), 7 km/h Holding treadmill bar (2 mins), 3.5 km/h (2 mins), 3.5 km/h Holding treadmill bar (2 mins), Rest (1 min)	**Accelerometer & Gyroscope:** 3-axis using a 6-axis inertial measurement unit **Photoplethysmogram:** Three channel LED: Green (525 nm), 1 LED on either side of each photodiode. Three PPG sensors were placed 6 mm apart. Photodiodes: 3 photodiodes **Electrocardiogram:** 24-h Holter monitor **Data:** ECG (125 Hz), all other signals (50 Hz).

Table 2.3 continued from previous page

Dataset	Participants	Protocol	Devices
PPG DaLiA 2019 [130]	*15 subjects* **Biological Sex:** 7 Male, 8 Female **Age:** 30.6 ± 9.6 years **Weight:** 69.0 ± 12.4 kg **Height:** 175.3 ± 8.8 cm **Skin Types:** 1 (0), 2 (1), 3 (11), 4 (3), 5 (0), 6 (0)	*Naturalistic Protocol* Sitting Still (10 mins), Ascending/Descending stairs (5 mins), Table Soccer (5 mins), Cycling (8 mins), Driving Car (15 mins), Lunch break (30 mins), Walking (10 mins), Working (20 mins) Double tap accelerometer signal pattern used for signal synchronisation.	**Accelerometer:** 3 axes Empatica E4 Device **Photoplethysmogram:** Empatica E4 LEDs: 2 Green, 2 Red. 2 Photodiodes with a total area of 15.5 mm^2. **Electrocardiogram:** RespiBAN Professional Device. **Respiration:** RespiBAN Professional Device **Electrodermal Activity:** Empatica E4 Range (0.01 µS – 100 µS) **Temperature:** Empatica E4 Resolution of 0.02 °C **Data:** All RespiBAN Professional (700 Hz), PPG (64 Hz), Accelerometer (32 Hz), Electrodermal Activity and wrist Temperature (4 Hz)
DWL (Wrist) 2022 [128, 136]	*14 subjects* **Biological Sex:** Unreported **Age:** Unreported **Weight:** Unreported **Height:** Unreported **Skin Types:** Unreported	*Laboratory-Based Protocol on Treadmill* Rest (1 min), 6 km/h (1 min), 12 km/h (1 min), 6 km/h (1 min), 12 km/h (1 min), Rest (1 min) 12km/h was reduced if needed.	**Accelerometer:** 3 axes **Photoplethysmogram:** 2x IR LED (940 nm) 2x Green LED (520 nm) Blue LED: Unreported 1 x Photo-detector **Gyroscope:** 3 axes **Data:** All signals sampled at 100 Hz

TABLE 2.3: Available Wrist-worn PPG Heart Rate Monitoring Research Datasets. This literature review, conducted in 2023, examines datasets for wrist-worn PPG heart rate monitoring research. The features of each dataset include cohort, devices, and protocol. Some of these datasets were later used for the analysis and validation of the PPG heart rate estimation methodology. Two datasets were excluded: DWL [128] due to its small sample size and Casson et al. [129] because it employed different protocols for each subject, lacking the consistency needed for proper evaluation across subjects.

2.2.2 Signal Quality Analysis

Assessing PPG signal quality is an essential step in any of its applications. Broadly, the quality of a PPG signal is determined by the clarity of the physiological information contained within the signal. As detailed in Section 2.1.3, the quality of PPG signals is affected by several sources of interference, with studies showing up to 86% of the collected signals being of insufficient quality for wrist-worn PPG heart rate monitoring [113]. Signal quality analysis aims to identify segments of the signal that contain interferences, determine the magnitude of the interference to gain insights into the potential source or sources to ultimately establish the recoverability of the segment. Basic sanity checks are generally the first quality checks performed, primarily focusing on identifying interferences stemming from sensor configuration, placement, and communication [137]. These checks can include missing data detection, flat-line detection, and clipping or over-saturation detection.

The approach to quality analysis varies depending on the specific application and context [137]. The analysis generally involves extracting features known as Signal Quality Indices (SQIs) from the signal. SQIs can focus on individual beats and waveform morphology or segments of the signal. As detailed in Section 2.2.3, isolating individual beats from wrist-worn PPG is challenging, especially in periods of motion, making beat and waveform-specific features less applicable to wrist-worn applications.

Elgendi proposed using the agreement of two distinct beat detectors to estimate noise within the PPG signal [138]. Elgendi examined this metric along with seven others and found skewness to be the most optimal SQI for finger-worn PPG in clinical settings [138]. Still, the estimation of higher-order statistics requires a relatively long time window [139]. A common SQI in the literature is the SNR, albeit with diverse definitions. Other research used additional reference signals in their PPG signal quality assessment, such as accelerometer [140,141] and ECG [142].

While some methods further this analysis by categorising the quality of segments [138,140–144], this necessitates the use of human-annotated labels and introduces the risk of error propagation as well as additional computational overhead [113]. It has been recommended that to guarantee the optimal performance of the application, a more nuanced consideration of PPG signal quality within a PPG signal processing pipeline is essential [113].

2.2.3 Conventional Beat Detector Algorithms

heart rate estimation from PPG signals can be achieved by detecting individual heartbeats, relying on PPG waveform features like systolic and diastolic peaks, the dicrotic

notch and the diastolic trough [124] (Figure 2.1). This method enables the extraction of detailed features from waveform morphology and inter-beat intervals, aiding in-depth physiological and cardiovascular analysis [124]. However, this approach is susceptible to inaccuracies. Misidentification or omission of heartbeats can lead to significant errors. Motion artefacts and demographic variations further challenge its robustness.

Charlton et al. evaluated 15 PPG beat detectors against ECG-derived heartbeats using data from eight datasets, including hospital and daily living settings, as summarised in Table 2.3 [124]. Data was sourced from eight datasets encompassing hospital and everyday living settings, including the PPG DaLiA dataset. Hospital data utilised transmissive PPG sensing, while daily activity data used wrist-worn reflectance PPG sensors. Evaluation metrics were the F1 score and the Mean Absolute Percentage Error (MAPE), which assesses the accuracy of computed heart rate values. Heart rate is computed from the detected beats using:

$$HR(BPM) = 60 \cdot \frac{Numberof DetectedBeats}{ElapsedDurationof DetectedBeats(Seconds)} \tag{2.1}$$

where the fraction represents the mean inter-beat interval (IBI), notably The Association for the Advancement of Medical Instrumentation (AAMI) standard prescribes acceptable limits for heart rate monitoring within ±10%, as measured by MAPE [27, 145, 146]. Additionally, the F1 score is a metric, ranging from 0 to 1, that evaluates a model's accuracy by combining its ability to make correct predictions with its consistency in identifying relevant instances.

In minimal movement conditions, beat detector performance varied. Median F1 scores spanned 50.7% to 99.9%, and MAPE ranged from 0.2% to 59.7%. The top eight detectors had F1 scores between 90% and 99% [124]. Intense physical activities reduced performance, with F1 scores dropping to 17.9%-90.6% and MAPE values rising to 7.0%-69.0%. The eight top detectors had F1 scores between 55% and 91%, with the lowest accuracy during table soccer and stair climbing [124].

For the beat detectors skin melanin differences had minimal impact in hospital settings with minimal motion. Subjects with higher melanin had median F1 scores of 91.2%-98.5% and MAPE values of 1.4%-9.9%, compared to 86.6%-97.5% and 2.1%-14.6% for those with lower melanin in hospital settings. Additionally, high inter-subject variability was observed in the wrist-worn datasets [124].

Overall, the most effective detectors were MSPTD [147] and QPPG [148], with high accuracy in minimal movement conditions and reduced performance during intense activities with median MAPE values ranging from 4.3% to 20.1% [124]. While the study

was comprehensive in its scope, it had limitations due to the constraints of the available datasets. Specifically, it did not investigate the combination of factors that could affect PPG signal quality, such as skin melanin content, the type of motion involved and biological sex [7]. Additionally, the study did not investigate the influence of individual wavelengths on beat detection efficacy.

2.2.4 Conventional Heart Rate Estimation Algorithms

An alternative approach for PPG-based heart rate estimations focuses on calculating the average heart rate over a designated interval rather than isolating individual heartbeats. While PPG heart rate estimation algorithms offer a coarser-grained analysis, they compensate by providing a more stable representation of heart rate due to enhanced robustness against errors such as missed or misidentified heartbeats.

Regarding methodology, PPG heart rate estimation algorithms generally include four main steps: prepossessing, motion artefact reduction, heart rate estimation and heart rate tracking or post-processing [92–95]. The prepossessing step typically includes filtering, re-sampling, windowing, transformation and normalisation. A notable feature across all heart rate estimation algorithms is an 8-second sliding window with a 2-second shift. Each window undergoes analysis and has an assigned "ground truth" value derived from a chest-worn ECG. Perhaps the most essential step is motion artefact reduction, which may incorporate motion reference signals gathered from accelerometers, gyroscopes, and the PPG sensor itself [7, 92–95]. Evaluation of the methodology generally occurs in terms of mean absolute error (MAE) of the predicted values against the 'ground truth' values.

Research into methods for wrist-worn PPG heart rate estimation began with the seminal work of TROIKA. This three-step technique focuses on de-noising, high-resolution spectral analysis, and spectral peak tracking. It employs independent component analysis and adaptive filtering to mitigate motion artefacts in PPG signals without requiring additional sensors or reference signals [126]. The validating dataset, referred to as IEEE 2015 SPC, was later used in the 2015 IEEE Signal Processing Cup, which popularised the topic in academia [133].

Many conventional computational techniques have been employed to enhance the accuracy and robustness of PPG heart rate estimations. These methods range from basic thresholding and filtering to more advanced techniques like spectral analysis and signal decomposition, as extensively summarised in [92–95]. For instance, SpaMA employs power spectral density analysis of PPG and accelerometer signals to identify and eliminate motion artefacts and HR, using a thresholding heart rate tracker for further

Name	**Pre Processing**	**Strategy**	**Post Processing**
ABD [120]	Windowing	Three-stage filtering identifies pulse peaks using adjusted Kaiser windows.	Inter-beat interval correction
AMPD [149]	Detrending and Windowing.	The local Maxima Scalogram method identifies beats by locating scale-dependent maxima, where these maxima are only considered within scales smaller than the one containing the most maxima	None
ATM [150]	Filtering and Normalisation	Peak detection uses an adaptive threshold proportional to PPG amplitude, dynamically adjusting to signal variations.	Inter-beat interval correction
CoPPG [151]	Windowing	Percentile-based thresholds set for adaptive filtering. Peaks exceeding the 90th percentile were identified.	Inter-beat interval correction
ERMA [152]	Filtering, Rectifying, and Squaring	Short and long-term averages identify interest blocks. Within valid blocks, beats are detected when the short-term average exceeds the long-term plus threshold.	None
HeartPy [153]	Normalising and Squaring	Rolling mean threshold approach, testing different moving average percentage values to find the most stable heart rate estimate within a valid BPM range.	Inter-beat interval correction.
IMS [154]	None	Positive gradient segmentation approach using dynamic thresholds based on amplitude and duration to detect beats	None
MSPTD [147]	Detrending and Windowing	Improves AMPD by calculating Local Maxima Scalograms for both local maxima and minima.	None
PDA [155]	None	Upslope sequences tracking approach with dynamic thresholds based on sequence length and amplitude for peak identification.	None
PWD [156]	Filtering	Zero-crossing analysis of the PPG derivative with dynamic thresholds, artefact compensation, and peak verification.	None
PPG Pulses [157]	None	Peak detection using a differentiated PPG and an adaptive filter. Filter threshold adjustment based on previous peak amplitude and inter-beat interval.	None
QPPG [148]	Scaling	Detects peaks using a slope sum function and adaptive thresholding.	None
SPAR [158]	Windowing and Filtering	Time delay coordinates generate 7-dimensional phase space for PPG windows. Symmetric Projection Attractor Reconstruction creates a 2-dimensional projection with beat detection at x-axis crossings.	Inter-beat interval correction.
SWT [159]	None	Selected Stationary Wavelet Transform detail subbands used to emphasise upslopes. Beats are detected using an extracted envelope and Gaussian derivative filter.	None
WFD [160]	Filtering and Resampling	PPG is decomposed with wavelet transform. Beats are identified using signal thresholds and derivative analysis.	None

TABLE 2.4: Overview of Open Source PPG Beat Detection Algorithms. This table summaries the key characteristics of various open-source PPG beat detection algorithms evaluated by Charlton et al [124].

refinement [130, 161]. In contrast, Schack et al. developed a multi-channel technique that utilises cross-correlation and auto-correlation between PPG signals to enhance

signal periodicity. The spectra of the PPG signals are combined to amplify common components and minimise noise. Motion artefacts are reduced through harmonic noise damping using accelerometer spectra, and heart rate is recursively tracked using a Gaussian window and linear least squares fitting. [130, 162].

Researchers have also developed multi-wavelength approaches, for example, Warren et al. developed a multi-channel, forehead-worn PPG sensor using red and IR wavelengths. The advanced multi-channel template matching algorithm selects the least artefact-affected channel for real-time heart rate estimation. The results show that the accuracy of heart rate estimates increased by up to 2.7 BPM when using the multichannel-switching algorithm compared to individual channels [71]. Similarly, Alkhoury et al. produced a dual-wavelength method using green and IR wavelengths for heart rate estimation during physical activity. Noise components were extracted from the IR signal and removed from the green PPG signal using a cascading adaptive filter. The outcomes indicate a notable enhancement in performance. Specifically, the Mean Absolute Error (MAE) was recorded at 1.2 ± 0.6 BPM for the wrist and 1.3 ± 0.8 BPM for the palm. In contrast, the single-wavelength TROIKA method yielded an MAE of 3.2 ± 2.8 BPM on the wrist and 1.8 ± 0.9 BPM on the palm [128].

Conventional PPG heart rate estimation algorithms typically have adjustable parameters that can be tuned to improve performance. It is common practice in the literature to adjust these parameters per subject to achieve the highest accuracy for each subject. However, this approach is limited when transitioning from a controlled experimental setting to real-world applications. The absence of "ground truth" values or comprehensive signal data in real-world scenarios to retrospectively tune the parameters limits the practical relevance of the reported results.

In real-world applications, an effective PPG heart rate estimation algorithm is anticipated to function accurately on data from individuals it has not previously encountered. To rigorously assess this generalisability, a 'Leave-One-Subject-Out' (LOSO) cross-validation (CV) scheme is recommended [26, 130]. Within this validation scheme, the data from one session or subject is left out of the parameter tuning and is used to evaluate the performance of unseen data. This is repeated for all subjects in the dataset. Riess et al. employed the LOSO CV scheme to assess the above-mentioned methods. Their findings revealed a significant increase in MAE results, escalating from 1.33 ± 1.4 BPM to 13.1 ± 20.7 BPM on the IEEE Train dataset and from 2.53 ± 2 BPM to 9.2 ± 11.4 BPM on the IEEE Test dataset for the SpaMA method. Similarly, the method proposed by Schack et al. exhibited comparable increases in MAE, rising from 1.3 ± 1.3 BPM to 2.91 ± 4.6 BPM on the IEEE Train dataset and from 6.5 ± 8.3 BPM to 24.7 ± 24.0 BPM on the IEEE Test dataset [130].

Furthermore, the validation datasets present limitations that preclude insights into the robustness of the methods concerning varying skin melanin content. As elaborated in Section 2.1.3, a burgeoning body of evidence indicates that skin melanin negatively impacts PPG sensing performance. While PPG-DaLiA provides Fitzpatrick skin type information, it does not include subjects at the extreme ends of the scale, precisely skin types 1, 5, and 6 [130]. This highlights a gap in the diversity of skin types in current PPG heart rate estimation research data.

2.2.5 Deep Learning PPG Heart Rate Estimations

Supervised deep learning involves learning patterns and rules from data within a defined hypothesis space guided by feedback [163]. This space is formed by network layers that transform inputs using weights and biases, with non-linear activation functions expanding the range of transformations [163]. Unlike traditional machine learning methods that require manual feature engineering, deep learning models learn relevant features directly from raw data, removing the need for explicit feature extraction [163].

Training involves passing batches of data through a 'network' to produce predictions, comparing these against target values using a loss function, and adjusting network parameters via backpropagation and gradient descent [163]. A network's 'architecture' encompasses the selected layers, configurations, and connections. These choices delineate the network's hypothesis space, the potential functions that gradient descent explores, determined by the model's parameters [163]. A good hypothesis space incorporates prior knowledge about the data.

Supervised deep learning aims for generalisation, the ability to perform well on unseen data. The balance influences this capability struck between over-fitting and under-fitting [163]. Over-fitting occurs when a model captures noise and anomalies in the training data, making it perform poorly on new data. Conversely, under-fitting is when the model cannot capture the underlying patterns in the data [163]. Regularisation techniques, such as dropout, where random network connections are 'dropped' during training to prevent co-adaptation, and batch normalisation, which standardises inputs, are employed to combat over-fitting [163].

Deep neural networks (DNNs) are distinguished from shallow neural networks by their multiple hidden layers, and different types of DNNs are further distinguished by their specific architectures. These architectures integrate prior knowledge of the data to create a comprehensive hypothesis space [163]. For instance, CorNet was the first to use the Long-term Recurrent Convolutional Network (LRCN) for time-domain signals, leveraging convolutional layers to capture local patterns and Long Short-term Memory

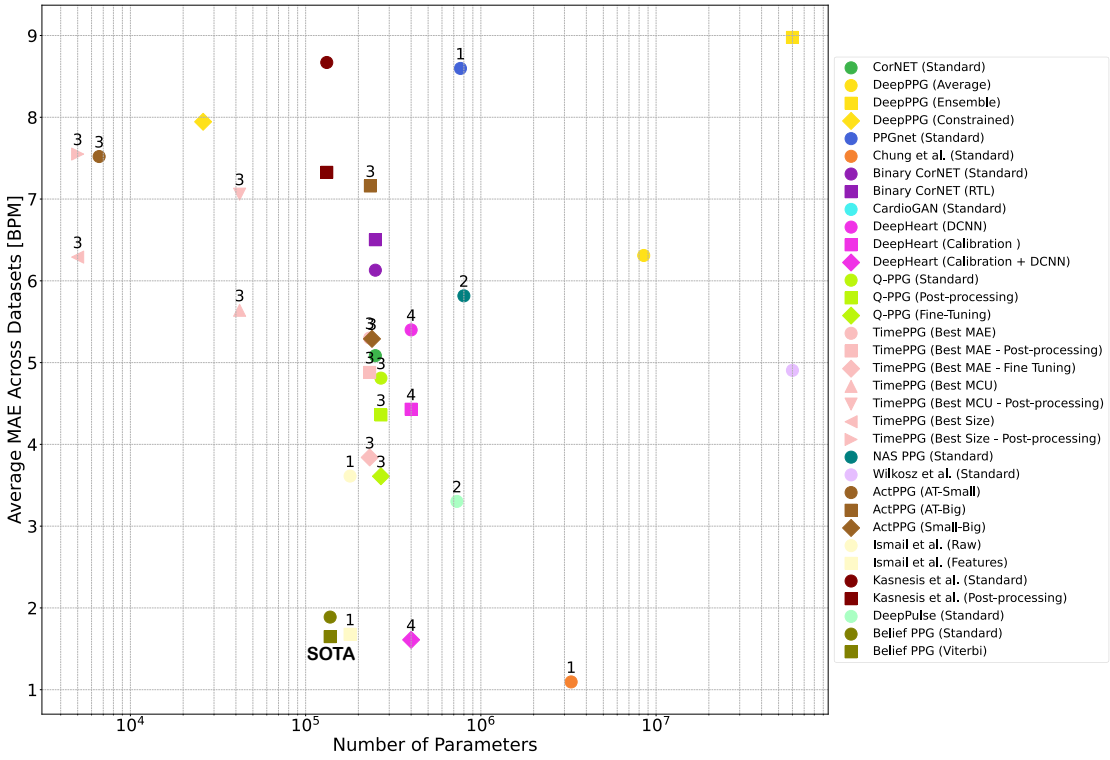
(LSTM) layers for long-term patterns [164]. This combination forms the basis of many heart rate (HR) estimation methods [26, 27, 165–169]. Alternatively, Temporal Convolutional Networks (TCNs) use causal dilated convolutional layers to hierarchically capture both local and extended patterns [170–172], with the receptive field expanding across the network to consider a larger portion of the signal for predictions [170–172]. Some strategies adopt an AlexNet-like structure, employing convolutional and pooling layers to reduce data dimensionality while preserving significant features [130, 173], while others use inception blocks to extract multi-scale features [165]. Attention layers dynamically allocate importance to specific time steps, enhancing performance, interpretability, and flexibility [27, 174, 175]. Additionally, the U-Net architecture, with its symmetric encoder-decoder structure and skip connections, provides precise localisation and efficient feature extraction and reconstruction [27, 174].

Researchers have utilised network architecture search (NAS) as a data-driven approach to architecture generation for heart rate estimations [168, 170–172, 176]. Ray et al., in earlier research, investigated three distinct NAS techniques but found them too resource-intensive due to a large search space [176]. To mitigate this, researchers employed seed architectures to reduce the search space. Burrello et al. and subsequent works used a temporal convolutional network as a seed network leveraging MorphNet and Pruning-in-time NAS techniques to optimise the network [170–172]. Song et al. employed an LRCN seed architecture using the Efficient Neural Architecture Search algorithm and Tree-structured Parzen Estimator hyperparameter optimisation to find an optimal solution [168].

Generally, pre-processing includes filtering the signals to be within the typical heart rate frequency range of 0-4 Hz, resampling the signal, and standardising the signals, subject-wise, to be zero mean unit variance and an 8-second sliding window with a 2-second shift. Most methods use the time domain representation of the signal as input to the network [26, 164–167, 170–172, 175, 177] whilst other methods utilise the Fourier Transform to attain better frequency resolution [127, 130, 168, 174]. Bieri et al. used both time and frequency domain representations, preserving the resolution of both domains [27]. Ismail et al. extracted statistical, time and frequency domain features alongside processing the time domain signals, seeing MAE improvements from 5.4 ± 6.3 BPM to 2.4 ± 2.9 BPM on the IEEE Datasets [169]. Bieri et al. also investigated augmenting the data to expand the number of samples in the training set by employing techniques such as time stretching and jitter. They found that including data augmentation caused the error rates to almost halve on some datasets [27].

Predominantly, methods formulate heart rate estimation as a regression task, whereby the model predicts a continuous heart rate value [26, 164–167, 169–172, 175, 177, 178].

Conversely, an alternate approach is to formulate heart rate estimation as a classification task [27, 127, 168]. One such approach classified the heart rate value within a predetermined set of bins, making it prone to quantisation errors. Researchers have also formulated PPG beat detection as a classification task, using temporal convolutional networks to classify systolic beats [179]. Another approach formulates heart rate estimation as a generative one, producing models that generate de-noised PPG signals [173] or ECG signals from PPG signals [174].



1: Didnt use LOSO, 2: Report results not Held out set, 3: Subject-specific Tuning, 4: Post-processing without Cross Validation, SOTA: State of the Art

FIGURE 2.6: Relationship Between Accuracy and Complexity in Deep Learning PPG Heart Rate Estimation Algorithms. This figure shows how model accuracy relates to complexity, with parameter counts ranging from 65 million [130] to 5,000 [172]. Some methods achieve high accuracy with fewer parameters, while others perform poorly despite greater complexity, highlighting that complexity does not guarantee better results. The algorithms must also support real-time predictions and edge device compatibility, with generalisability evaluated via LOSO CV.

In real-world applications, heart rate algorithms must deliver real-time predictions on individuals new to the system and maintain a model complexity suitable for edge devices, often measured in the number of parameters. These complexities vary from 65 million [130] to 5,000 parameters [172]. Some strategies aim for edge compatibility by using binarised [167] or quantised [170] networks, while others prioritise efficient parameter counts [27]. As highlighted in Section 2.2.4, generalisability on new users is gauged using LOSO CV. This balance of complexity and generalisability is pivotal in

evaluating the efficacy of such methods, as illustrated in Figure 2.6. Notably, Bieri et al. set a benchmark by deploying a 138K parameter network, showcasing the lowest error rates across various datasets, including MAE of 1.5 ± 0.6 BPM on IEEE Train and 1.5 ± 0.3 BPM on the BAMI-2 datasets [27].

Ray et al., in earlier research, introduced uncertainty quantification to PPG heart rate estimation deep learning methods. The approach employed the Monte Carlo dropout method to quantify epistemic uncertainty and a distributional prediction strategy with a negative log-likelihood (NLL) loss function to quantify aleatoric uncertainty. While they applied practical methods to validate these uncertainty quantifications, they did not use standardised calibration techniques [26]. Bieri et al. employed a belief propagation method to quantify predictive uncertainty, both uncertainty types combined, using a quantised probability distribution. This method demonstrated robust overall calibration, though it was slightly overconfident at higher confidence levels [27].

In multi-wavelength deep learning PPG heart rate estimation methods, Ngoc-Thang et al. developed an LRCN model for a finger-based transmissive mode PPG sensor using red (660 nm) and IR (880 nm) wavelengths. This methodology yielded a correlation coefficient 0.996 with heart rate values generated by a pulse oximeter. However, it lacked CV and a data acquisition protocol encompassing motion [180]. Mehrgardt et al. similarly developed a finger-based transmissive mode PPG sensor using IR (880 nm), red (660 nm) and green (537 nm) wavelengths as well as accelerometer and gyroscope data. Using a network of four fully connected layers, they analysed combinations of the signal data. During stationary, the combination of green, red, and IR PPG and accelerometer and gyroscope data produced the lowest MAE results of 2.63 ± 30.05 BPM. Conversely, whilst walking green, PPG combined with accelerometer and gyroscope data exhibited the lowest MAE of 6.52 ± 43.68 BPM. During running, the green, red, and IR PPG ensemble, complemented by accelerometer and gyroscope data, showcased the lowest MAE of 5.8 ± 34.7 BPM [181].

One common limitation across all approaches is the lack of exploration of fairness. It is essential to guarantee that these systems do not reflect discriminatory or unfair behaviour toward specific individuals or populations. Mehrabi et al. describe fairness in deep learning as "the absence of prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics" [182].

Regrettably, there have been many examples of such discriminatory behaviour within machine learning systems. One of the most notable examples is COMPAS, a machine learning system that measures a person's risk of committing another crime. A study found it to have a higher false positive rate for African Americans than Caucasians, inaccurately predicting their risk of re-offending [183]. Similarly, in the healthcare sector,

a machine learning algorithm used by more than 200 million patients in the USA was less likely to refer equally sick black patients than white patients to programs aimed at improving care for patients with complex needs [184].

Due to limitations in wrist-worn PPG heart rate monitoring validation datasets, no method gives insights into the robustness to demographic variations and consequently the fairness of the method. This raises concerns considering the substantial evidence indicating that PPG sensing is influenced by certain demographic factors, including higher skin melanin levels, being a biological female, and increased BMI.

Paper	Signal Used	Preprocessing	Architec-ture	Training	# Parameters	Learning Task	Additional Comments
CorNet *2019* [164]	1 PPG Channel	**Filter:** 4^{th} order Butterworth bandpass filter (0.1-18 Hz) **Resample:** 125 Hz **Transform:** None **Normalise:** Z Score	LRCN	**Batch Size:** 25 **Optimiser:** RMSProp **CV:** 5 Fold **Loss:** Undefined	256K	Regression	Predicts biometric ID as well as HR.
Deep PPG *2019* [130]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** Spectrogram Clipping (0-4 Hz) **Resample:** None **Transform:** Short Term Fourier Transform **Normalise:** Z Score	CNN	**Batch Size:** 128 **Optimiser:** Adam **CV:** LOSO **Loss:** MAE	S: 8.5M, E: 60M, R-c:26K	Regression	Introduced three methods: Standard, En-semble, and Resource-constrained.
PPGNet *2019* [165]	1 PPG Channel	**Filter:** 2^{nd} order Butterworth band-pass filter (0.5-5 Hz) **Resample:** None **Transform:** None **Normalise:** Z Score	Inception Block and LRCN	**Batch Size:** 128 **Optimiser:** SGD **CV:** 5 Fold **Loss:** MAE	765K	Regression	Transfer learning was examined using the weights trained from one dataset on an-other.
Chung et al. *2020* [127]	3 PPG Channel and 3 Axis Accelero-meter	**Filter:** 4^{th} order Butterworth bandpass filter (0.4-4 Hz) **Resample:** 25 Hz **Transform:** Fast Fourier Transform **Normalise:** Z Score PPG Channels and Accelerometer spectra averaged	LRCN	**Batch Size:** 32 **Optimiser:** Adam **CV:** 4 Fold **Loss:** Cross Entropy	3.27M	Classification	
PP Net *2020* [166]	1 PPG Channel	**Filter:** None **Resample:** 31.25 Hz **Transform:** None **Normalise:** Z Score Truth Values also normalised	LRCN	**Batch Size:** 100 **Optimiser:** Adam **CV:** 10 Fold **Loss:** MSE	124k	Regression	
Binary CorNet *2020* [167]	1 PPG Channel	**Filter:** 4^{th} order Butterworth bandpass filter (0.2-4 Hz) **Resample:** 125 Hz **Transform:** None **Normalise:** Z Score Input Quantisation (5-bit)	LRCN	**Batch Size:** 32 **Optimiser:** Adam **CV:** LOSO **Loss:** Undefined	Memory: 260 Kbits	Regression	Gradient-based moving average post-processing used to compensate for accuracy loss due to binarisation.
Cardio GAN *2020* [174]	ECG and PPG	**Filter:** ECG: FIR bandpass filter (3-45 Hz), PPG: But-terworth bandpass filter (1-8 Hz) **Resample:** 128 Hz **Transform:** Fast Fourier Transform **Normalise:** Min Max	Attentional U-Net	**Batch Size:** 128 **Optimiser:** Adam **CV:** Train-test split **Loss:** Combination of Cross-Entropy and MSE	Undefined	Generative	Network generates ECG signals from PPG signals.
DEEP HEART *2021* [173]	1 PPG Channel	**Filter:** 3^{rd} Order Butterworth bandpass filter (0.4-5Hz) **Resample:** 32 Hz **Transform:** None **Normalise:** None Synthetic clean PPG signal generated from ECG R peaks	CNN	**Batch Size:** 128 **Optimiser:** Adadelta **CV:** LOSO **Loss:** Cosine Similarity	400k	Generative	
Q-PPG/Time PPG *2021* [170, 171]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** Bandpass filter (0.5-4 Hz) **Resample:** 32 Hz **Transform:** None **Normalise:** None	TCN	**Batch Size:** 128 **Optimiser:** Adam **CV:** LOSO **Loss:** LogCosh	BestMAE: 232K, BestMCU: 42k, BestSize: 5k	Regression	The post-processing step compares the latest TCN prediction with the average of the previous N predictions and clips the es-timate if the difference exceeds a threshold.

Table 2.5 continued from previous page

Paper	Signal Used	Preprocessing	Architec-ture	Training	# Parameters	Learning Task	Additional Comments
NAS PPG *2021* [168]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** 4^{th} order Butterworth bandpass filter (0.4-4 Hz) **Resample:** 32 Hz **Transform:** Fast Fourier Transform **Normalise:** Min Max	LRCN	Batch Size: 128 **Optimiser:** Undefined **CV:** LOSO **Loss:** Cross Entropy	800K	Classification	Efficient neural architecture search and Tree-structured Parzen estimator hyperparameter optimisation techniques used to find optimal architecture and hyperparameters.
Wilkosz et al. *2021* [177]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** None **Resample:** 32 Hz **Transform:** None **Normalise:** Z Score	Multi-branch LRCN	**Batch Size:** 128 **Optimiser:** Adam **CV:** LOSO **Loss:** MSE	60M	Regression	
Act PPG *2022* [172]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** Bandpass filter (0.5-4 Hz) **Resample:** 32 Hz **Transform:** None **Normalise:** None	TCN	**Batch Size:** Undefined **Optimiser:** Undefined **CV:** LOSO **Loss:** LogCosh	6.63k/234k	Regression	The post-processing step compares the latest TCN prediction with the average of the previous N predictions and clips the estimate if the difference exceeds a threshold.
Ismail et al. *2022* [169]	2 PPG Channel and 3 Axis Accelero-meter	**Filter:** 30^{th} Order Elliptic bandpass filter (0.4-5 Hz) **Resample:** 12 5Hz **Transform:** Fractional Brownian motion, Statistical, Time domain and Spectral feature extraction. **Normalise:** Undefined	LRCN	Batch Size: Undefined **Optimiser:** Undefined **CV:** Subject-specific and subject-independent. **Loss:** MAE	179K	Regression	
Kasnesis et al. *2022* [175]	All PPG Channels and 3 Axis Accelero-meter	**Filter:** None **Resample:** 32 Hz **Transform:** None **Normalise:** Z Score	Attentional CNN	Batch Size: 256 **Optimiser:** Adam **CV:** LOSO **Loss:** Undefined	132K	Regression	The architecture allows for the visualisation of attentional maps, which enhances model explainability. In the post-processing stage, output values undergo clipping if they deviate by more or less than 10% from the average of the last 10 estimated values.
Deep Pulse *2022* [26]	1 PPG Channel and 3 Axis Accelero-meter	**Filter:** 2^{nd} order Butterworth bandpass filter (0.5-4 Hz) **Resample:** 64 Hz **Transform:** None **Normalise:** Z Score	Multi-branch LRCN	**Batch Size:** 32 **Optimiser:** Nadam **CV:** LOSO **Loss:** NLL	730K	Regression	Quantification of Epistemic and Aleatoric Uncertainty using Monte Carlo dropout and outputting distribution parameters were used with a Negative Log Likelihood loss function.
Belief PPG *2023* [27]	All PPG Channels and 3 Axis Accelero-meter	**Filter:** Time Domain: 4^{th} order Butterworth bandpass filter (0.1-18 Hz) Frequency domain: Spectra clipping (0.5-3.5 Hz) **Resample:** 64 Hz **Transform:** Fast Fourier Transform **Normalise:** Z Score	LRCN and attentional U-Net	**Batch Size:** 128 **Optimiser:** Adam **CV:** LOSO, 5 Fold and Leave one Dataset out **Loss:** Categorical Cross Entropy	138K	Classification	Data Augmentation used (Jitter and Time Stretch) Predictive uncertainty is quantified using a quantised probability distribution obtained using belief propagation.

TABLE 2.5: Summary of PPG Deep Learning Heart Rate Estimation Algorithms. This table presents a literature review conducted in 2023, summarising 16 papers on PPG deep learning heart rate estimation algorithms. The columns include: Signal Used (including accelerometer use and the number of PPG channels), Preprocessing (such as filtering, transforming, resampling, and normalising), Architecture (type and components), Training (parameters such as optimiser, batch size, and epochs), Number of Parameters (in the model), Learning Task (regression, classification, or generation), and Additional Comments (including uncertainty quantification and other relevant notes).

2.3 Summary

This chapter offers a thorough review of the advancements and challenges in the field of multi-wavelength PPG, with a particular focus on wrist-worn devices used for heart rate monitoring. It highlights several critical research gaps and limitations affecting the effectiveness and fairness of current approaches.

A significant gap is the limited diversity in existing PPG datasets, which inadequately represent various skin types, especially darker skin tones (Fitzpatrick types 5 and 6). This lack of diversity hinders the evaluation of wrist-worn PPG heart rate estimation methodology's fairness and robustness.

Moreover, current datasets often focus on controlled, periodic motions, such as treadmill running, and fail to encompass the wide range of motion types and intensities encountered in daily life. This narrow scope limits the applicability of wrist-worn PPG heart rate estimation methodology in real-world scenarios.

Validation methods also pose a challenge. Many studies do not employ Leave-One-Subject-Out (LOSO) cross-validation, a key technique for assessing the generalisability of wrist-worn PPG heart rate estimation methodology. Without robust validation, the reliability of these methods for new users remains unknown.

The chapter also addresses the scarcity of research on multi-wavelength approaches in wrist-worn PPG heart rate estimation methodology. While using multiple wavelengths has the potential to enhance accuracy and robustness, there is a notable lack of studies exploring deep learning techniques that integrate this approach for wrist-worn devices.

Uncertainty quantification is another area needing attention. Few studies incorporate uncertainty estimation into wrist-worn PPG heart rate estimation methods, which is essential for building trust in healthcare applications of PPG technology.

Fairness considerations are similarly under-explored. There is a notable absence of research examining how heart rate estimation methods perform across different demographic groups, particularly concerning skin tone, biological sex, and body mass index (BMI).

Additionally, many existing algorithms are not optimised for real-time performance or deployment on wearable devices, which limits their practical use. Enhancing real-time capabilities is key for the effectiveness of these systems in everyday applications.

Lastly, the chapter highlights the need for a comprehensive analysis of how various interference sources—such as skin melanin, motion artifacts, and biological differences—affect PPG signal quality. Addressing these issues could improve the robustness

and reliability of heart rate monitoring systems.

Overall, the chapter underscores the need for more diverse and comprehensive datasets, improved validation methodologies, and advanced estimation methodologies that address fairness, uncertainty, and real-world applicability in wrist-worn PPG heart rate monitoring.

Chapter 3

Research Design and Methodology

3.1 Gaps in Existing Research

In the previous chapter, the literature review identified three primary gaps in current research: Firstly, heart rate estimation datasets show a lack of diversity in terms of skin melanin content and the variety of motion types and intensities. Secondly, there is an absence of wrist-worn multi-wavelength deep-learning methods for heart rate estimation. Thirdly, most existing heart rate estimation algorithms have not been robustly analysed for demographic variations, such as biological sex and skin melanin content. Additionally, the influence of skin temperature on heart rate estimation algorithms is another gap, though not covered in this thesis.

3.2 Research Objectives and Questions

In this section, the primary research questions are systematically bridged with the corresponding objectives, which are established to address the intricacies of wrist-worn Photoplethysmography (PPG) heart rate estimation using multi-wavelength approaches. This mapping is pivotal in providing a coherent and strategic alignment between the core investigative queries and the structured objectives that underpin the thesis:

1. To what extent does the robustness and generalisability of wrist-worn PPG heart rate estimations vary across different wavelengths or combinations of wavelengths, compared to the green light conventionally used in consumer wrist-worn devices?
    - Objective 5: Influence of Wavelength Selection on PPG Heart Rate Estimation. Related Chapter: 6.
2. What is the impact on performance based on variations in skin melanin content and biological sex in wrist-worn PPG heart rate estimation?

- Objective 6: Impact of Skin Melanin and Biological Sex on PPG Heart Rate Estimation. Related Chapter: 6.

3. In PPG heart rate estimation, does deep learning demonstrate superior performance compared to conventional signal processing and statistical methods?
    - Objective 9: Comparative Evaluation of PPG Heart Rate Estimation Methods. Related Chapters: 6 and 8.
4. How can uncertainty be most reliably estimated in the context of PPG heart rate estimations?
    - Objective 7: Evaluation of Uncertainty Methods in Deep Learning. Related Chapter: 7.
5. To what extent does the inclusion of uncertainty metrics in post-processing enhance the reliability of wrist-worn PPG heart rate estimations?
    - Objective 8: Development of Post-Processing Methods for PPG Heart Rate Estimations. Related Chapter: 7.

3.3 Definitions

Throughout the thesis, several key terms are consistently used and defined here for clarity. The first term, "Accuracy", describes the precision of predictions. It is quantified using metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the standard set by AAMI, which is a MAPE of 10%.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3.2}$$

where, y_i is the truth value of the i^{th} sample, $\hat{y}_i$ is the predicted value of the i^{th} sample and n is the number of samples. "Generalisability" is the second term and refers to the ability of a model or method to maintain its accuracy when applied to new data not previously used in its training or validation. This aspect is key in determining the model's applicability in real-world scenarios beyond the controlled settings of training and testing. The third term, "Robustness", addresses the model's accuracy under challenging or adverse conditions, such as during intense physical motion. This trait is essential for models used in environments where conditions can significantly vary. Lastly, "Fair" pertains to the uniformity of the model's accuracy across different

demographic groups, such as biological males and females. Ensuring fairness is critical in developing models that perform equitably across diverse populations.

In addition to these metrics, the Bland-Altman plot will be used in this research, serving as an analytical tool for assessing the agreement between two different measurement methods, as shown in Figure 3.1. It is particularly valuable for comparing new techniques with established standards. The plot effectively visualises the difference between two measurements against their mean.

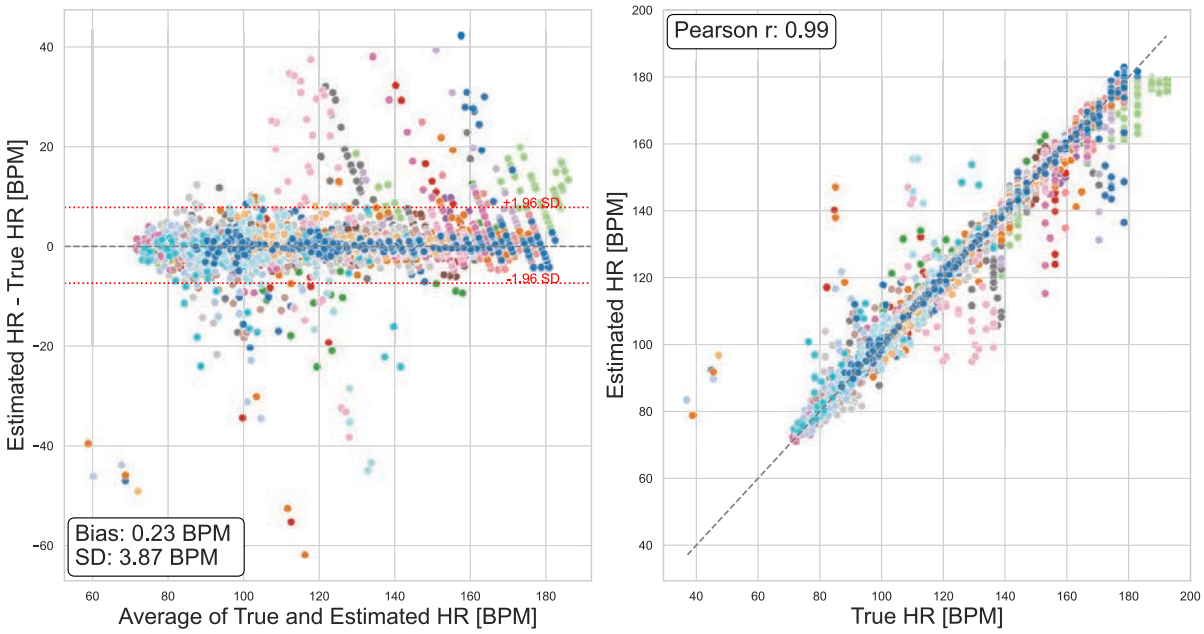


FIGURE 3.1: Bland-Altman and Correlation Plots for Heart Rate Predictions. The Bland-Altman Plot (left) shows the agreement between predicted and true heart rate (HR) values, with the difference between them plotted on the y-axis and their mean on the x-axis. This plot highlights any systematic bias or trends in prediction accuracy. The Correlation Plot (right) illustrates the linear relationship between predicted and true HR values, indicating the strength and direction of their correlation.

For uncertainty estimates in a regression setting, average calibration is used to evaluate the reliability of a model's predictions. This assessment is known as "average calibration," and it evaluates how well the uncertainties predicted by a model align with the actual errors observed in the predictions [178].

To conduct this evaluation, the model's predictions are grouped into several "bins" based on their predicted uncertainty levels. For each bin, the average actual error of the predictions is then calculated. The purpose of this process is to determine whether the predictions in each bin are as uncertain as the model predicts [178]. In other words, the model's confidence in its predictions should match the actual outcomes: if the model predicts a high uncertainty, the errors should indeed be larger; if it predicts low

uncertainty, the errors should be smaller. This concept is expressed as:

$$p_{avg}^{obs}(p) := \mathbb{E}_{x\sim\mathbb{F}_X}[\mathbb{F}_{Y|x}(\mathbb{Q}_p(x))], \forall p \in (0,1) \tag{3.3}$$

X and Y are random variables, with x and y being specific values of X and Y. $\mathbb{F}$ represents the true cumulative distribution function (CDF) of a random variable. $\mathbb{E}_{x\sim\mathbb{F}_X}$ denotes the expected value (average) over the distribution of X, meaning we are averaging over all possible values of X. $\mathbb{Q}_p(x)$ is an estimate of the quantile function at percentile p, which is essentially the inverse of the CDF, evaluated at x. For perfect calibration, the relationship $p_{\text{avg}}^{\text{obs}}(p) = p$ must hold true for all p values between 0 and 1. This means that the model's predicted uncertainty should ideally equal the actual proportion of times the model's prediction is correct [178]. Points above this line indicate under-confidence (where predicted probabilities are too low), while points below the line indicate overconfidence (where predicted probabilities are too high). Average calibration is often depicted graphically, as shown in Figure 3.2, to compare

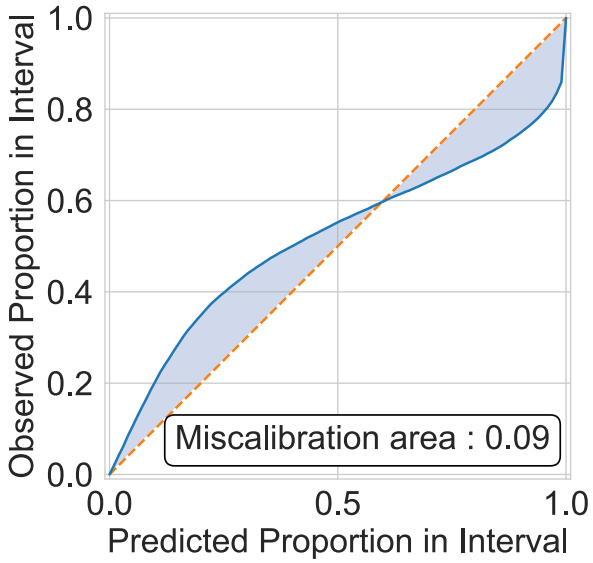


FIGURE 3.2: Calibration Plot for Probability Estimates. This figure shows the calibration of predicted probabilities by plotting the observed proportion of outcomes against the predicted proportion across various probability intervals. The orange line represents perfect calibration, where predictions match observed frequencies. Points above this line indicate under-confidence (where predicted probabilities are too low), while points below the line indicate overconfidence (where predicted probabilities are too high). The miscalibration area is also displayed to highlight deviations from perfect calibration.

perfect calibration (a straight line) with the model's actual calibration (which might be a curve) [178]. The difference between these two lines can be measured by calculating the "miscalibration area," which is the area between the perfect and actual calibration lines.

This measurement is key for models where the reliability of predictions is important. The smaller the miscalibration area, the better the model's calibration, meaning its uncertainty predictions are more reliable.

In conclusion, the methodologies outlined in this thesis, including the use of Mean Absolute Error, Mean Absolute Percentage Error, Bland-Altman and correlation plots, and calibration plots and metrics, provide a comprehensive framework for evaluating the accuracy, generalisability, robustness, and fairness of the models developed. These methodologies not only ensure the scientific rigour of the research but also enhance the applicability and relevance of the models in real-world scenarios. By meticulously examining the accuracy and reliability of predictions across different conditions and populations, this thesis contributes valuable insights into the field, fostering advancements in predictive modelling and its practical applications.

3.4 Research Timeline

The chronological progression of this research did not align with the thesis's order. Initially, the research explored various avenues in deep learning for heart rate estimation, including network architecture search [176], but some paths were abandoned after preliminary investigations with existing datasets. A literature review was conducted [7], leading to the identification of multi-wavelength approaches in wrist-worn PPG heart rate estimation [186] and the potential of incorporating uncertainty quantification, both showing promise in preliminary data [26]. Subsequently, a dataset was designed and collected to address identified gaps, which then informed the refinement of the uncertainty-aware deep learning method for this new dataset, both of which are detailed in this thesis.

It should be noted that the global COVID-19 pandemic had an impact on the research timeline, necessitating adjustments to the original research plan and data collection procedures. Despite these challenges, the core objectives of the research were maintained and successfully pursued.

3.5 Software Ecosystem

The software ecosystem of this thesis includes a range of tools for data analysis, signal processing, and deep learning. Python 3.11 is the primary language used throughout, with Pandas 2.1.4 for data manipulation. SciPy 1.11.2 and NumPy 1.24.4 support scientific computing and numerical operations, respectively.

Tkinter 8.6 is used for GUI development, and the study-watch-sdk 4.4.0 integrates with study-watch devices. ECG signals are processed with py-ecg-detectors 1.3.3. For PPG beat detection, PPG-beats 1.01 (August 2022) and Matlab R2022b are utilised.

Machine learning tasks are carried out with scikit-learn 1.3.2, and deep learning is handled by TensorFlow 2.15, with TensorFlow Probability 0.20.0 applied for uncertainty quantification in Section 7. Statistical modelling is performed using statsmodel 0.14.1, while Uncertainty_Toolbox 0.1.1 supports further uncertainty analysis. Data visualisation is achieved with matplotlib 3.8.2 and Seaborn 0.13.1.

Computational tasks not involving deep learning were executed on an Intel i7-1355U CPU with 1.70 GHz and 16.0 GB of RAM. In contrast, deep learning tasks were performed using NVIDIA T4 GPUs with 16 GB of GDDR6 memory and a memory bandwidth of 320 GB/s, hosted on the Google Cloud Platform.

Chapter 4

Specification and Processing of A Wrist-worn Multi-wavelength Photoplethysmography Heart Rate Monitoring Dataset

The preceding chapter detailed the research design and methodology, framing the approach undertaken in this study and thesis. This chapter addresses Objective 2, beginning with an explanation of the design and implementation of key elements in a photoplethysmography (PPG) heart rate monitoring dataset: protocol, cohort, and devices. The chapter then expands on the data processing steps, including signal alignment for PPG and electrocardiogram (ECG) signals and heart rate extraction from ECG signals. The chapter concludes with details on skin tone classification along with the rationale of other computed metrics.

4.1 Protocol

In designing the study protocol, careful attention was given to the insights and limitations discovered during the review of existing PPG heart rate estimation datasets (Section 2.2.1). The aim was to develop a protocol encompassing a variety of activities to mirror diverse real-world scenarios and physical conditions, addressing the limitations related to the diversity of motion and activity intensities found in previous laboratory-based studies [126–129].

The protocol was structured into four main phases: Active Rest, Running, Rest, and Cycling, summarised in Table 4.1. The Active Rest phase involved activities focused on wrist movements designed to capture erratic aperiodic movement and periodic contractions of the posterior forearm muscles. The Running phase included different

Phase	Activity	Duration (Minutes)	Fitness Level (PAR)	Expected Motion Type	Additional Comments
Active Rest	Stress Ball	2	1-7	Periodic	Contraction of Posterior Forearm Muscles
Active Rest	Hand Gripper	2	1-7	Periodic	
Active Rest	Finger Stretcher	2	1-7	Periodic	
Active Rest	Writing	2	1-7	Aperiodic	Erratic Wrist Movements
Active Rest	Typing	2	1-7	Aperiodic	
Running	3 km/h	3	1-7	Periodic	Potential capture of the 'crossover effect'
Running	5 km/h	3	1-7	Periodic	
Running	7 km/h	3	3-7	Periodic	
Running	11 km/h	3	4-7	Periodic	
Running	15 km/h	3	5-7	Periodic	
Rest	Hands on Table	2	1-7	No Movement	
Rest	Free Movement	2	1-7	Aperiodic	Participants could move freely around the laboratory
Rest	Hands on Table	2	1-7	No Movement	
Rest	Free Movement	2	1-7	Aperiodic	Participants could move freely around the laboratory
Rest	Hands on Table	2	1-7	No Movement	
Cycling	0.5 kg	3	1-7	Periodic	Participants were asked to keep their hands on the handlebar to potentially capture elevated heart rates associated with low upper-body movements
Cycling	1 kg	3	1-7	Periodic	
Cycling	2 kg	3	3-7	Periodic	
Cycling	3 kg	3	4-7	Periodic	
Cycling	5 kg	3	5-7	Periodic	

TABLE 4.1: Data Collection Protocol Overview. This table details the phases, activities, and conditions used in the data collection protocol, including the duration, fitness level (Physical Activity Rating, PAR), expected motion type, and additional comments. Activity duration is variable and adjusted based on the participant's PAR, ranging from active rest tasks, such as using a stress ball or typing, to more intense exercises like running at different speeds and cycling with varying resistance levels. The protocol accommodates both periodic and aperiodic motions, with notes on potential outcomes, such as the 'crossover effect' during running or elevated heart rates during low upper-body movement in cycling.

intensities ranging from 3 km/h to 15 km/h. This was followed by a Rest phase, during which participants placed their hands on a table, allowing for the observation of reductions in heart rate and absence of movement. The concluding Cycling phase employed an ergometer at different resistances, ranging from 0.5 kg to 5 kg, aimed at observing increases in heart rates with minimal movement, as participants were instructed to keep their hands on the handlebars. The total duration of the protocol ranged between 32 and 50 minutes, depending on the participants' fitness levels. The implemented equipment

comprised an h/p/cosmos Pulsar 3p treadmill and a Monark 874E weight ergometer. The laboratory's temperature was unmonitored and unregulated, and a single operator conducted data collection for all participants.

Running and cycling intensities were determined based on each participant's self-reported physical activity level, preventing overexertion or risk of injury. Participants were free to stop any activity or the entire protocol at any time, ensuring their comfort and safety throughout the study. Three minutes were allocated for running and cycling activities to allow the stabilisation of heart rate before a change in intensity, and two minutes were assigned for the other activities to capture a representative portion of the activity adequately. This approach to protocol design was anticipated to yield a rich and reliable dataset, facilitating the improvement and validation of heart rate estimation methods and their adaptability to various real-life applications and conditions, albeit with the inherent limitations of laboratory-based protocols.

4.2 Cohort

The recruitment of participants strictly followed Manchester Metropolitan University's ethical guidelines (EthOS ID: 40624). Interested individuals who could register their interest online were reached through flyers and posters. After registration, they received an information sheet and a scheduling form. To be included, participants had to be healthy adults over 18 years old with no known cardiovascular diseases. Those taking medications that alter heart function, such as asthma medicines, decongestants, illegal drugs, and certain prescription medications, were excluded. Individuals who did not meet these criteria were informed of their ineligibility and thanked for their interest.

Upon arrival, participants received a briefing on health and safety along with detailed information about the study. After providing consent and receiving a copy of the signed consent form, basic measurements were recorded, and the designated devices were attached. Participants were instructed to wear the wrist-worn device on the wrist of their dominant hand, adjusted to a fit similar to how they typically wear watches, as shown in Figure 4.4 B. The chest-worn strap was applied following the guidelines provided in the device documentation [187].

To capture skin tone accurately and ensure precise colour calibration, a colour checker card featuring the Fitzpatrick scale was placed under the subjects' arms during image capture, illustrated in Figure 4.1. Photographs were taken using a Samsung Galaxy S8+ smartphone, employing a 12 MP, f/1.7, 26 mm (wide) camera with Dual Pixel Phase Detection Auto Focus and Optical Image Stabilisation. This method ensured the

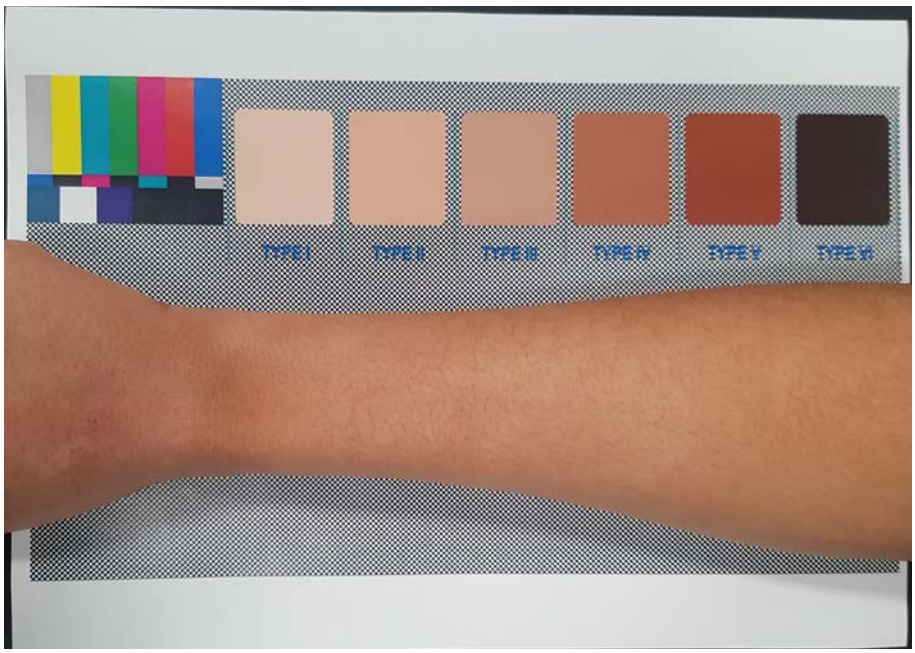


FIGURE 4.1: Example of Participants' Arm on the Colour Checker and Fitzpatrick Scale Card.

accurate capturing of skin tones and white balance for post-processing. Participants were thanked and gifted for their valuable involvement and contribution to the study.

Measurement	**Scale**
Age	The International System of Units scale for time, specified in years
Biological Sex	The World Health Organisation and the European Institute for Gender Equality define sex as 'Biological and physiological characteristics that define humans as female or male.' [188]
Dominant Hand	Left or Right
Fitness Level	Self-administered Physical Activity Rating [190]
Height	The International System of Units scale for length, specified in centimetres
Skin Type	Self-administered Fitzpatrick Scale [189]
Weight	The International System of Units scale for mass, specified in kilograms
Wrist Circumference	The International System of Units scale for length, specified in centimetres

TABLE 4.2: Overview of Basic Physiological and Demographic Measurements. This table outlines the key physiological and demographic measurements collected from participants, including their corresponding scales.

While determining the appropriate study size, power analysis is often utilised to ascertain the number of subjects needed to confirm or reject a hypothesis. Based on the works of Bent et al. and Fallow et al., a study size of ≥ 48 participants is considered necessary to achieve 80% power to reject the null hypothesis concerning differences in PPG accuracy between Fitzpatrick skin types, with an ANOVA power calculation suggesting 8 participants for each of the 6 skin types [52, 59]. However, Colvonen et al. contend that conclusions drawn from such analyses might be misleading due to factors affecting PPG sensing accuracy, within-group variance of skin tone types, and potential administrative errors in classification. Consequently, Colvonen et al. advocate for larger sample sizes, especially including more individuals with darker skin tones, to limit false negatives and account for possible interactions with skin tone [191]. Given the budget constraints of this research, a cohort of the recommended size was not feasible. Therefore, the aim was to maintain a proportional representation across different skin types within an attainable sample size of $n = 20$, acknowledging the inherent limitations and potential biases in the findings due to the restricted cohort size.

4.3 Devices

This section outlines the devices employed for gathering signals, encompassing both chest-worn and wrist-worn devices. Insight into the electrical components and operation of each device, as well as the various types of data they capture, is provided. Particular attention is paid to the PPG sensor geometry, which is key for the signal quality. The section concludes with a discussion on data transmission and storage, emphasising the assurance of accuracy and precision through the strategic use of graphical user interfaces.

4.3.1 Electrocardiogram

In this study, the QardioCore chest strap was utilised to acquire ECG signals from a single channel, with an input dynamic range of 50 mV peak-to-peak and a DC dynamic span of ±300 mV. The device maintained a gain accuracy of 5% and a differential range of ±5 mV. The amplitude resolution of the ECG was 0.8 μV, and the signal bandwidth ranged from 0.05 to 40 Hz. The device employed an A/D sampling rate of 600 Hz, with the internal sampling rate ensuring precise signal acquisition. The sampling resolution was 16-bit, and the common mode rejection was 60 dB, with an input impedance of over 100 MΩ. Additionally, the device featured automatic calibration to maintain the accuracy of the measurements [187]. The QardioCore chest strap was previously validated against medical-grade Holter monitors in two separate instances, with sample

sizes of 50 and 31, yielding correlation scores of 0.92 and 0.95, respectively [192, 193]. The placement of the chest strap was meticulously performed as per the specifications delineated in the accompanying documentation, illustrated in Figure 4.2.

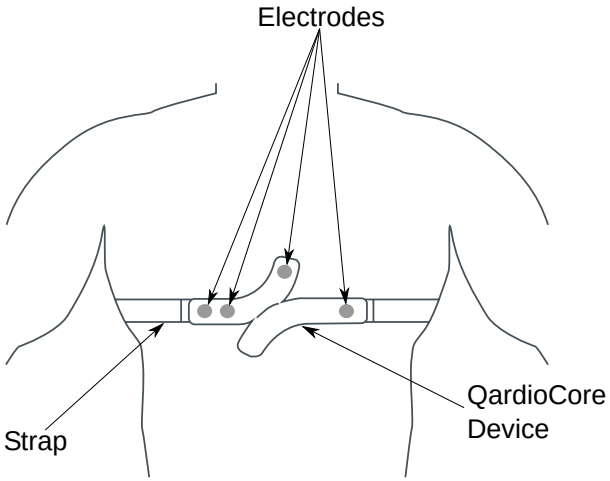


FIGURE 4.2: The QardioCore Chest Strap Placement, adapted from [187].

ECG signal acquisition was facilitated through the Android QardioDirect application (version: 2.8.8), specifically utilising a Samsung S8+ smartphone via Bluetooth connectivity. Subsequently, the gathered ECG signals were securely stored within the QardioMD web application, where they were anonymised and retrieved in HL7 aECG format. The detailed versions of the software used are outlined in Section 3.5

4.3.2 Wrist-worn Device

This study utilised the EVAL-HCRWATCH4Z (firmware version: 5.14), a wrist-worn device developed by Analog Devices. The research-grade bio-sensing device is equipped to acquire synchronised multi-wavelength PPG, triaxial accelerometer data, skin temperature, electrodermal activity, and ECG [194]. Notably, only multi-wavelength PPG, triaxial accelerometer and skin temperature were employed in the research.

A comprehensive overview of the electrical components embedded within the device is provided in Figure 4.3. The ADXL362 accelerometer facilitated motion detection with a ±8 g digital output range and SPI digital interface, sampled at 100 Hz [194, 195]. Skin temperature was collected using the NTCG104EF104FTDSX sensor from TDK Corporation, thermally coupled to the device's underside, with performance intricacies linked to its mechanical connection to the body, operating within a temperature range of -30°C to +50°C [194, 196].

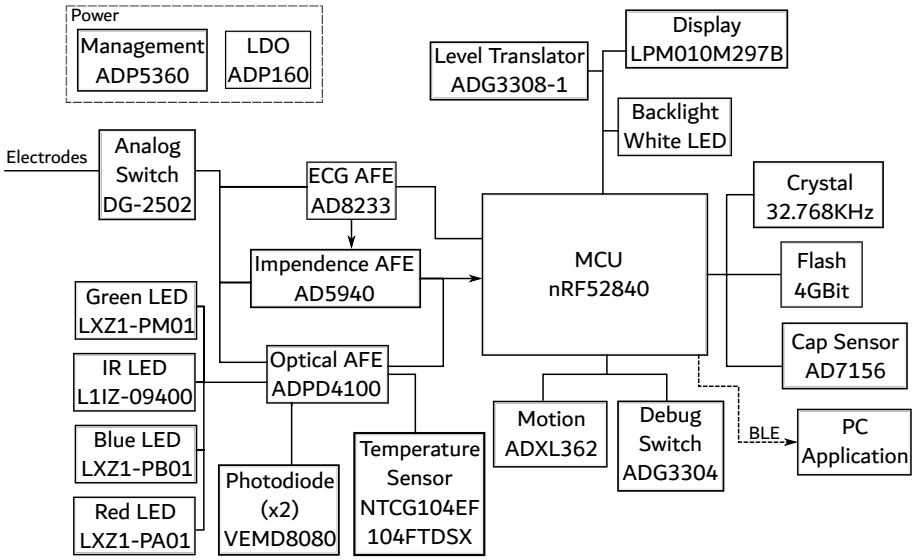


FIGURE 4.3: Overview of the electrical components of the wrist-worn device, adapted from [194]. The figure illustrates the PPG sensor (comprising the AFE, PDs, and LEDs), temperature sensor, bio-impedance AFE, ECG AFE, accelerometer, micro-controller, display and power management components.

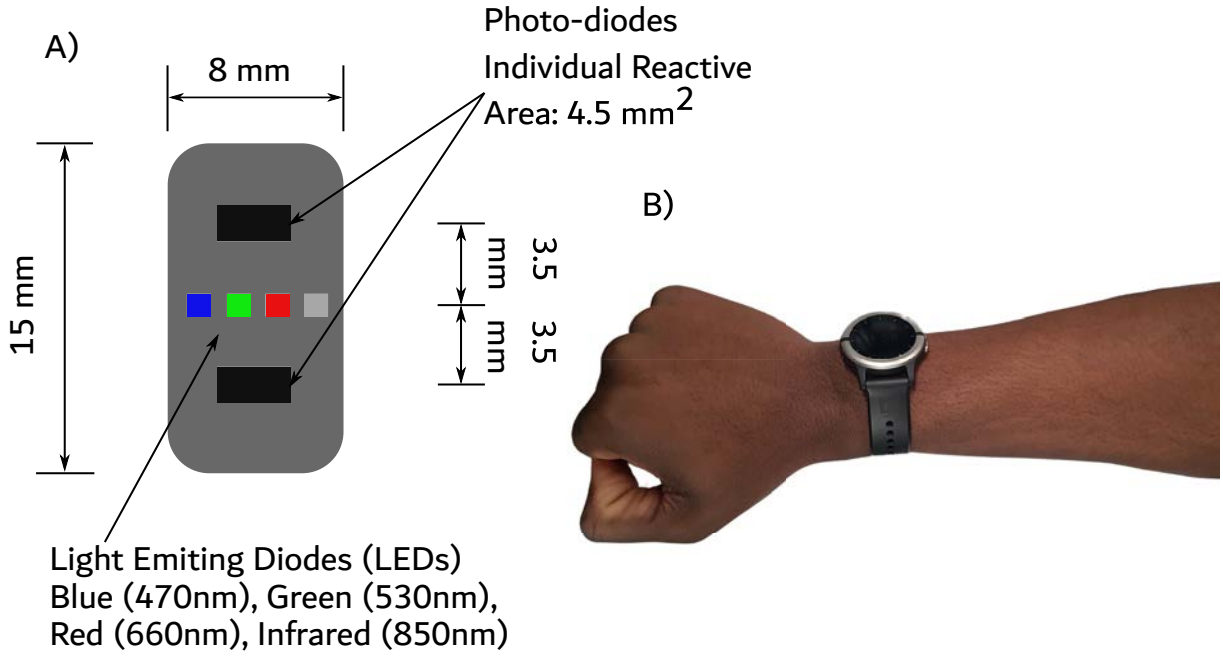


FIGURE 4.4: Illustration of the experimental setup for PPG measurement: A) Geometry and configuration of the PPG sensor. B) Wrist-worn device, including the PPG sensor, attached to the wrist of the participant's dominant hand, adjusted to the typical tightness of a watch.

The ADPD4100 [197], functioning as a multimodal AFE, incorporated inputs from the VEMD8080 Photodiodes [198] and the aforementioned temperature sensor, among others not utilised in this study. The ADPD4100 also controls the PPG sensors' blue (470 nm), green (530 nm), red (660 nm) and IR (850 nm) LEDs [199, 200]. Two channels were collected for each wavelength from a photodiode with a reactive area of 4.5 mm^2, positioned 3.5 mm from the LEDs on each side, illustrated in Figure 4.4 A. The LEDs operate at 7.5 mA and are gathered using a flexible input multiplexer in the sequence of green, IR, red, and blue. The device features a programmable timing controller capable at managing LED pulses, with a specified AFE width of 3 μs, pulse width of 2 μs, and a pulse offset of 16 μs. This precise control of pulse characteristics is key for capturing accurate PPG signals, ensuring the reliability of the data collected during the study. The configuration facilitates effective navigation through synchronised multi-wavelength PPG data acquisition complexities. Each channel maintained Transimpedance Amplifier gains of 200 kΩ and was sampled at 100 Hz, utilising an I^2C serial communication interface with a 100 kHz clock frequency. Through the employment of synchronous demodulation and other techniques, the ADPD4100 mitigates ambient light interference [194].

4.3.3 Collection Graphical User Interface

To facilitate accurate data collection from the wrist-worn device and synchronise it with activity timings, a custom graphical user interface (GUI) was developed. The connection between the device and the GUI was established using Bluetooth Low Energy (BLE) via a Nordic BLE nRF52840 USB dongle [201], which also served to minimise power line interference (50Hz/60Hz) and enhance the quality of the output signal [194].

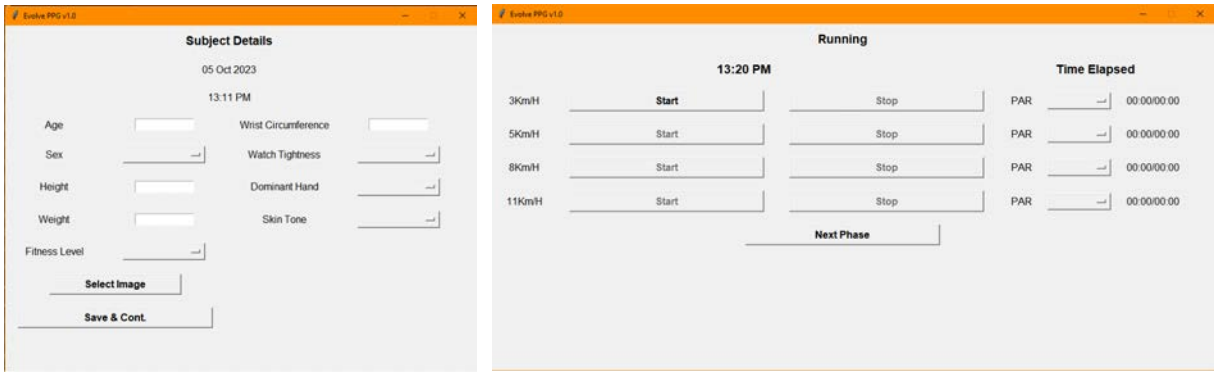


FIGURE 4.5: Data Collection Graphical User Interface for Subject Basic Measurements (Left) and Running Phase (Right)

Upon connection, the GUI provided an input form to record participant measurements, as detailed in Section 4.2, and additional forms to log the start and stop timings of each activity phase within the protocol, ensuring precise alignment with the collected

data. The resulting watch data, participant measurements, and activity timings were subsequently stored in CSV files for further processing. The detailed versions of the software used are outlined in Section 3.5

4.4 Signal Processing and Data Extraction

This section details the processes of signal processing and data extraction of the collected signals, which are pivotal for subsequent analysis and PPG heart rate modelling. Details encompass the alignment of ECG and PPG signals and the methodology for heart rate extraction from ECG, with comparisons and validations against alternative techniques. Additionally, the methods employed for skin tone classification and the computation of BMI and exercise effort are defined, laying a foundational framework for ensuing analyses and modelling endeavours.

4.4.1 Electrocardiogram and Photoplethysmogram Alignment

As elaborated in Section 4.3, the chest-worn ECG device and the wrist-worn bio-signal device recorded data via either a smartphone or a laptop. Each device has an internal clock that may not be perfectly synchronised, causing the timestamps recorded by each device to be misaligned. To correct these inter-device time delays, several computational strategies have been applied, including cross-correlation [202, 203], dynamic time warping [204], and region-of-interest matching methods like peak alignment [124] or the double tap method [130]. The double tap method creates a unique marker at the start and end of the protocol by having the participant tap all devices twice at the same time.

Dynamic time warping can adjust for non-linear distortions and shifts in time but is computationally expensive. Peak alignment depends heavily on the accuracy of the peak detection method employed. In this study, cross-correlation was chosen as a simple and effective method to align the signals. It's also important to note that since PPG devices are worn on the wrist, there's a natural delay between the ECG and PPG heartbeat due to the pulse transition time (PTT), causing a delay that can be between 100 - 250 ms depending on blood pressure [205].

To achieve an accurate lag estimate from cross-correlation, there needs to be sufficient variation in frequency, meaning there should be fluctuations in heart rate exceeding 10 BPM. The transition from a state of running to a state of rest provides this necessary variation in BPM, making it an ideal scenario to align the signals effectively. To calculate

the lag, three segments from each subject were utilised: a short, a medium, and a long segment, each centred around the transition from running to rest.

To align with the PPGs sample rate and optimise lag calculations, the ECG was downsampled to 100 Hz. All eight PPG signals were averaged to produce a single PPG signal to simplify and streamline the cross-correlation process. Segments were selected during the transition between running and resting phases, characterised by a sufficient variation in heart rate. This allowed analysis of the lag between ECG and PPG signals during this physiologically significant change. Three segment lengths were chosen to provide a comprehensive understanding of the temporal ECG-PPG relationship: a short 1-minute segment, a medium 2-minute segment, and a long 3-minute segment. For each segment—short, medium, and long—three distinct ranges of lags were systematically examined, as shown in Figure 4.6. Initially, a comprehensive full cross-correlation was conducted to scrutinise every possible lag. Subsequently, a method was applied that focused on the discrepancies between the total lengths of the PPG and ECG signals, addressing both additive and subtractive variances. Finally, a heuristic approach involving a visual examination of the spectrograms was applied to select a suitable range for lag comparison. This produced nine lag estimates, each visually examined to discard outliers, with the median value of the remaining selected as the optimal lag.

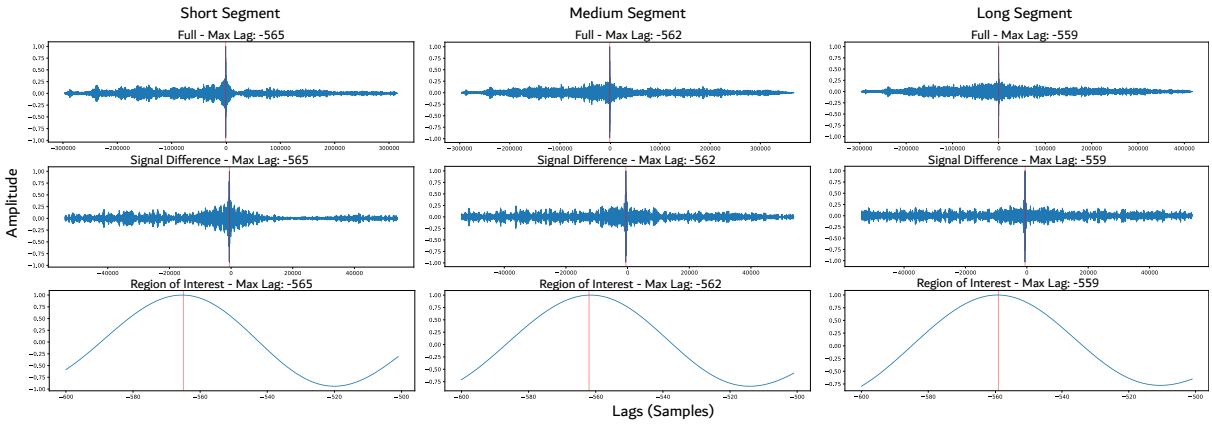


FIGURE 4.6: Cross-correlation analysis of ECG and Processed PPG signals for Subject 1, with systematic examination of three segment durations (short 1-minute, medium 2-minute, long 3-minute) and three lag estimation methods (full cross-correlation, signal differences, region of interest). This multimodal approach provides a comprehensive assessment of the temporal relationship between these cardiovascular signals during the physiologically significant transition from running to resting.

The sign of the optimal lag determined whether the ECG or the PPG was trimmed at the recording's start to synchronise the signals, and any remaining misalignment at the end was also rectified. This nuanced methodology ensured a thorough understanding of the alignment intricacies between the ECG and PPG signals, accommodating the inherent

variability in the recorded physiological data. The detailed versions of the software used are outlined in Section 3.5

4.4.2 Electrocardiogram Heart Rate Extraction

Heart rate 'ground truth' values from a chest-worn ECG play an integral role in Photoplethysmography heart rate datasets. These values serve to validate PPG heart rate estimation methods, necessitating their reliability and accuracy. Hence, the extraction of heart rate from the ECG is a key step in the construction of such a dataset.

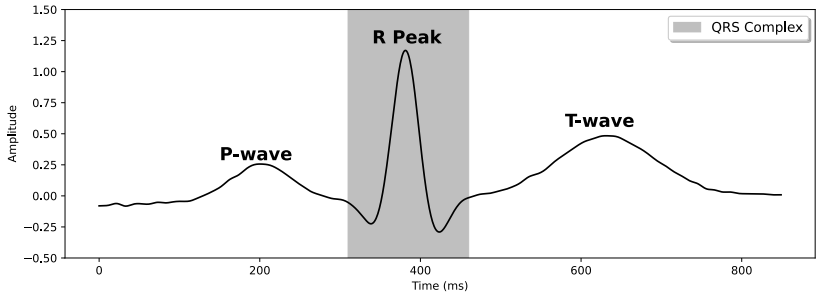


FIGURE 4.7: Typical Electrocardiogram Waveform Highlighting the QRS Complex and R Peak. This figure shows a standard ECG waveform, emphasising the QRS complex and the R peak. The QRS complex represents ventricular depolarisation and appears as a series of sharp, high-amplitude deflections. The R peak, the highest point within the QRS complex, marks the peak of ventricular depolarisation and is key for assessing heart rate and rhythm. The R-R interval, the time between successive R peaks, is used to calculate heart rate and monitor cardiac health.

Similar to PPG heart rate estimation, there exists a myriad of methods to extract heart rate from ECG signals. A prominent approach is to detect key features of the ECG waveform, such as the QRS complex (Figure 4.7). Hamilton and Tompkins designed a method that efficiently detects QRS complexes by analysing slope, amplitude, and width [206]. Hamilton refined this approach, enhancing its efficiency [207]. Conversely, Christov proposed an adaptive thresholding method [208], Elgendi et al. developed a moving average method [209], and Kalidas et al. utilised the Stationary Wavelet Transform [210]. However, similar to PPG beat detectors, this approach is susceptible to inaccuracies such as misidentifying or omitting QRS complexes, leading to significant errors.

Choosing frequency domain methods provides an alternative to time-domain QRS complex detection. The R peaks, serving as pivotal indicators of a heartbeat within an ECG, predominantly reside within the frequency range of 15-35 Hz [209]. Employing a Butterworth bandpass filter specifically tuned to this frequency range and squaring

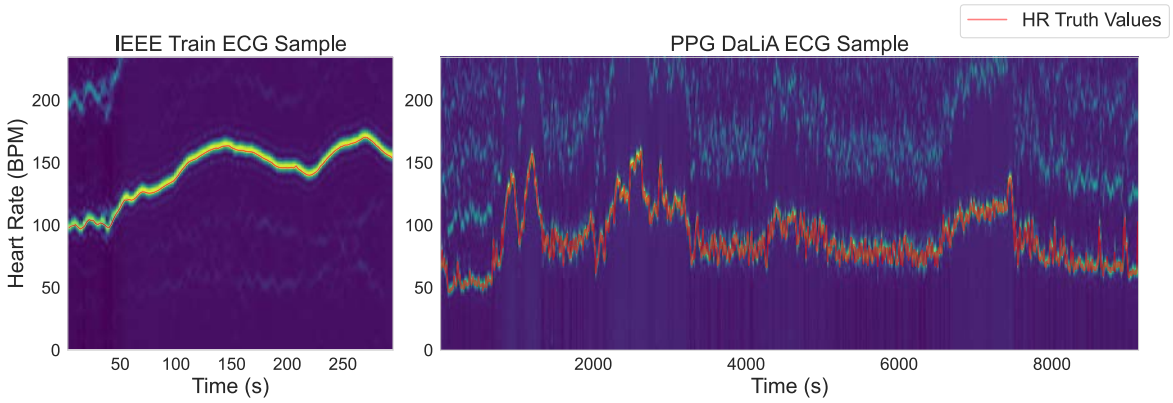


FIGURE 4.8: Samples of Pre-Processed ECG Signals from IEEE Train and PPG DaLiA Datasets with Extracted Heart Rate Truth Values. This figure presents pre-processed ECG signals from the IEEE Train [126] and PPG DaLiA [130] datasets, emphasising the extraction of heart rate truth values. The application of a Butterworth bandpass filter, tuned to the 15-35 Hz range where R peaks are most prominent, followed by signal squaring, effectively isolates these R peaks. The resulting spectrogram clearly reveals a heart rate trace that aligns with the ground truth heart rate values (shown in red).

the signal reveals a discernible heart rate trace within the spectrogram, as illustrated in Figure 4.8. This heart rate trace is corroborated by alignment with the ground truth heart rate values of the datasets, underlining the precise isolation of R peaks within the signal and attesting to the validity and efficacy of the frequency domain approach in isolating accurate heart rate information from ECG signals.

In light of the aforementioned analysis, a frequency domain methodology was proposed, as depicted in Figure 4.9. The ECG signal is first subject to a bandpass Butterworth filter to accentuate the essential R peaks and minimise noise interference. Subsequently, the filtered signals were squared, and a spectrogram was calculated and normalised, focusing on enhancing the visibility of R peak frequencies within the signal. Each point in the spectrogram undergoes rigorous analysis to identify the frequency that corresponds to the HR, factoring in amplitude range thresholds and frequency range. The methodology entails dynamic recalibration of the frequency range at each spectrogram point based on the previously identified frequencies while filtering out sub-threshold values or those that fall outside the accepted range.

To validate the efficacy of the described methodologies, multiple PPG heart rate estimation datasets were employed, each with ground truth heart rate values and ECG signals, acquired through varied devices and adhering to diverse protocols. This multifaceted approach facilitated a thorough examination of the methodologies' efficacy under diverse conditions, framing the evaluation as a supervised task. ECG signals from each

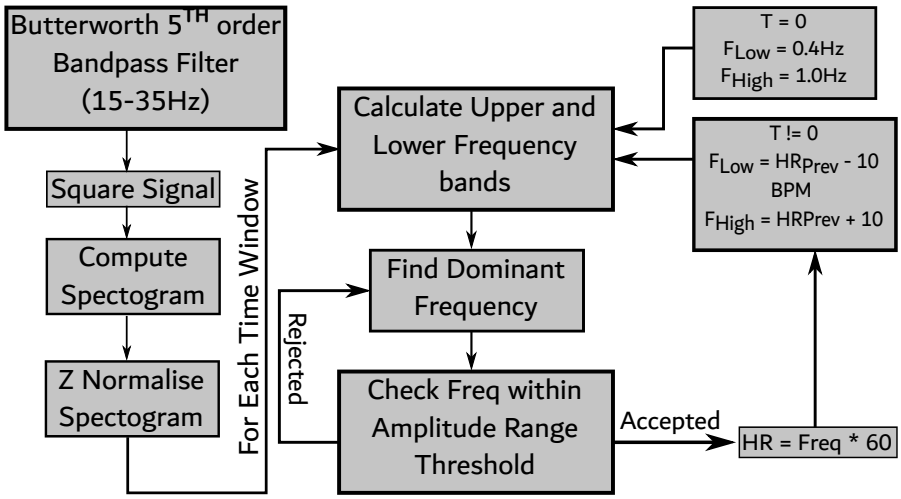


FIGURE 4.9: Block Diagram of the Proposed ECG Heart Rate Extraction Method. This diagram outlines the proposed frequency domain approach for heart rate extraction. It starts with a Butterworth bandpass filter (15-35 Hz) to isolate R peaks, followed by squaring the signal. A normalised spectrogram is then used to highlight R peak frequencies. Each spectrogram point is analysed to determine the heart rate, with dynamic recalibration to filter out irrelevant frequencies and ensure accuracy.

subject were segmented into 8-second windows with a 2-second slide, accompanied by a corresponding heart rate value from the dataset. The detailed versions of the software used are outlined in Section 3.5.

Method	IEEE Train [126]	PPG DaLiA [130]
Christov [208]	3.1 ± 4.1	2.7 ± 1.3
Elgendi et al. [209]	1.1 ± 1.6	1.4 ± 0.5
Kalidas et al. [210]	2.5 ± 1.2	2.9 ± 0.6
Hamilton [207]	1.3 ± 0.8	4.7 ± 1.5
Hamilton and Tompkins [206]	2.2 ± 0.6	3.2 ± 0.8
Proposed Method	**0.4 ± 0.2**	**1.2 ± 0.6**

All Values are MAE in BPM.

TABLE 4.3: Results of ECG Heart Rate extraction validation experiment on IEEE Train [126] and PPG DaLiA [130] datasets. **Bold** values indicate the lowest MAE distribution.

Table 4.3 shows that the proposed method demonstrated superior performance on the IEEE Train dataset with the lowest MAE of 0.4 ± 0.2 BPM, whilst maintaining competitive results on the PPG DaLiA dataset. This consistent performance across both datasets, coupled with low standard deviations, indicates a robust and reliable approach for heart rate estimation tasks. A visual inspection of the performance of the proposed

method on the collected ECG data substantiated its effectiveness, as depicted in Figure 4.10. The empirical observations and analytical outcomes collectively underscore the reliability and accuracy of the proposed method in extracting accurate heart rate data from diverse ECG signals.

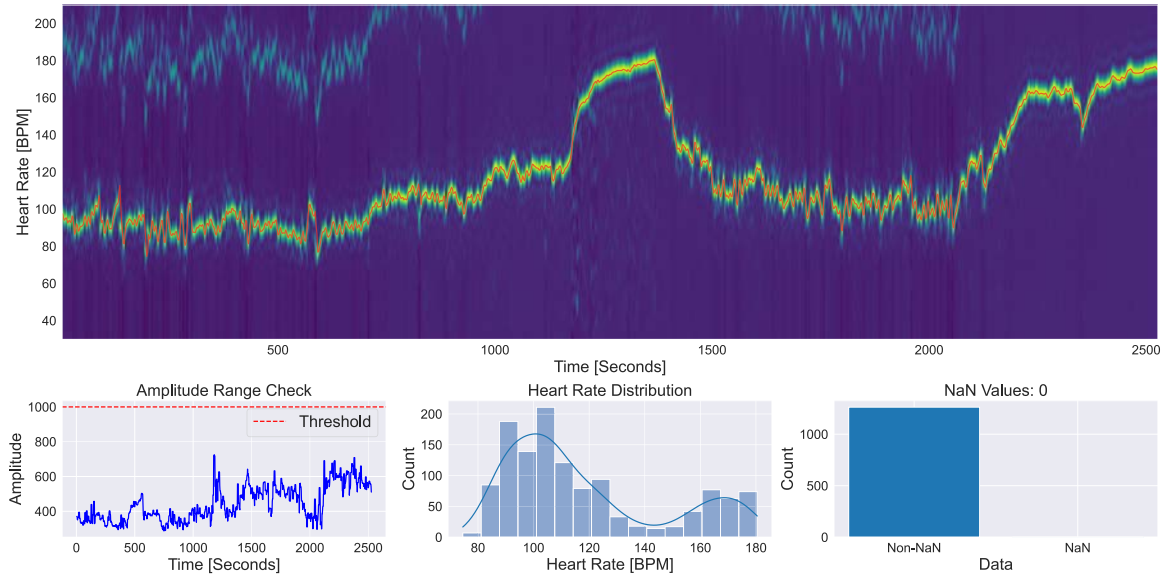


FIGURE 4.10: Analysis of ECG Heart Rate Extraction Method for Subject 1. This figure demonstrates the proposed ECG heart rate extraction method (red line) applied to the spectrogram of processed ECG signals. For this subject, the amplitude range threshold (bottom left) was not exceeded, resulting in the inclusion of all HR values (bottom right). The spectrogram shows an increase in heart rate during the running and cycling phases of the protocol, indicating a well-distributed range of heart rate values (bottom middle).

During the experiment, instances were identified where sub-optimal adherence of the ECG device to the subject compromised the integrity of the signal. Figure 4.11 depicts intervals where the heart rate trace in the spectrogram is obscured, rendering it non-distinct (see mid-section of spectogram). Examining the amplitude range during these obscured intervals uncovered considerable discrepancies attributed to noise. To rectify and mitigate the perturbations induced by such discrepancies, a threshold was applied to each subject's ECG signal, leading to the exclusion of windows that lacked a coherent and distinguishable ECG signal. This refinement led to the omission of 1643 windows, representing 5.1% of the entire dataset, reinforcing the reliability and accuracy of the data processed in subsequent analytical stages of the research.

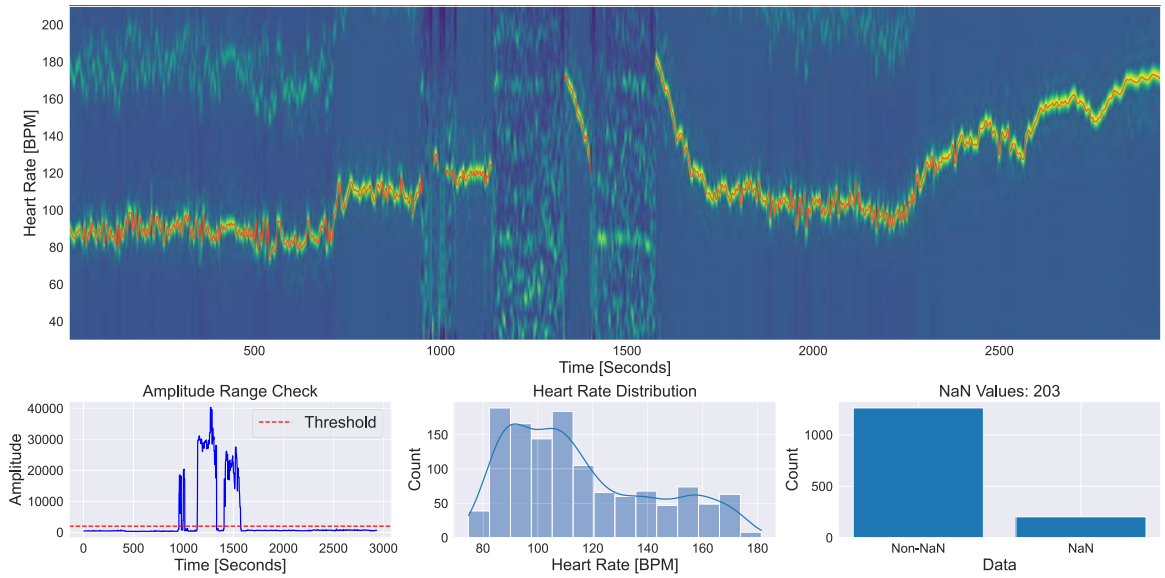


FIGURE 4.11: Analysis of ECG Heart Rate Extraction Method for Subject 21. This figure demonstrates the proposed ECG heart rate extraction method (red line) applied to the spectrogram of processed ECG signals. For this subject, the amplitude range threshold (bottom left) was exceeded in three instances, resulting in the exclusion of 203 heart rate values (bottom right).

4.4.3 Skin Type Classification

In PPG settings, the Fitzpatrick Skin Type Scale and Von Luschan's Chromatic Scale are commonly used to classify skin tones. These methodologies, however, are fundamentally subjective and can vary significantly depending on the assessor. They have also been criticised in dermatology research for focusing too much on lighter skin tones, which can lead to mistakes in evaluating risks and reactions in different skin types [212]. To overcome these limitations, a spectrophotometer has been recommended as the 'gold standard' for objective skin tone assessment [213]. Nonetheless, studies indicate that the results obtained from spectrophotometer assessments are in alignment with visual evaluations of skin colour, highlighting that inaccuracies can permeate both objective and subjective measurement methods due to improper application of techniques [214].

In this research, a robust approach was applied to impartially classify skin tones. As referenced in Section 4.2, a photograph was taken of each participant's arm alongside a colour checker and a Fitzpatrick scale card. Participants were also asked to self-administer the Fitzpatrick scale. To correct any colour variations across the photographs, a systematic white balancing method was utilised. This method focused on each of the three colour channels—Red, Green, and Blue—in the images. For each channel, the following steps were performed:

1. Histogram Calculation: A histogram was created to represent the distribution of pixel intensities in the channel. This histogram counts the number of pixels at each intensity level from 0 to 255.
2. Identifying Primary Range: The histogram was analysed to determine the primary range of pixel intensities. This was done by discarding the pixel colours at each end of the histogram that are used by only 0.05% of the pixels in the image, which helps in ignoring outliers that might be caused by artefacts such as bits of dust.
3. Clipping and Normalisation: The minimum (b_{min}) and maximum (b_{max}) intensity values within the primary range were identified. The pixel values in the original image were then clipped to this range and subsequently normalised. Specifically, any pixel values below b_{min} were set to b_{min} and any pixel values above b_{max} were set to b_{max}. This clipped range was then stretched to the full 0-255 range of possible intensity values using the formula:

$$\text{balanced}_{\text{img}}[..., i] = \left(\frac{\text{clipped}_{\text{img}}[..., i] - b_{\text{min}}}{b_{\text{max}} - b_{\text{min}}} \right) \times 255 \tag{4.1}$$

   where i represents the colour channel index (0 for Blue, 1 for Green, and 2 for Red).

This process ensures that the pixel values are adjusted and balanced, resulting in uniform colour distributions and improved image contrast within each refined image. The outcome is a white balanced image that more accurately represents the true colours of the scene, which is key for the accurate classification of skin tones.

After the image processing, a panel of three individuals, self-identifying as Fitzpatrick skin types 1, 3, and 5, independently assessed the participant's images using the Fitzpatrick scale. Figure 4.12 illustrates the observed variability in skin type categorisation across the study cohort, reinforcing the prevailing concerns regarding the reliability of this methodology. A weighted average classification was then calculated, giving self-administered classifications 1.5 times more weight than those from panel members. This approach aimed to achieve a more balanced and objective assessment of skin tones using the Fitzpatrick scale.

4.4.4 Additional Computed Metrics

As elucidated in Section 2.1.3, BMI has been reported to affect PPG signal quality. The BMI is typically calculated using the Quetelet Index, which is mathematically defined as an individual's weight in kilograms divided by the square of their height in

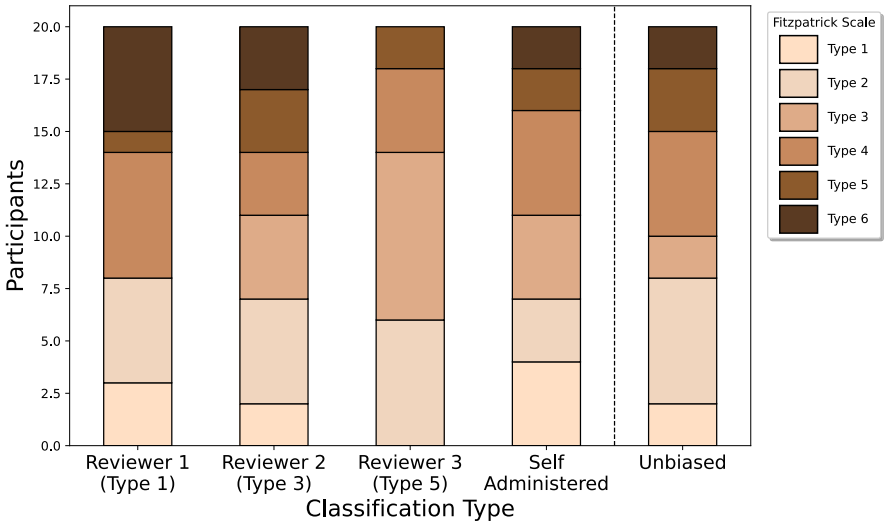


FIGURE 4.12: Classification of Cohort Skin Type Using the Fitzpatrick Skin Type Scale. This stacked bar chart shows the skin type classification of the study cohort using the Fitzpatrick Skin Type Scale. The first three bars represent classifications by independent reviewers, self-identifying as skin types 1, 3, and 5. The fourth bar reflects self-administered classifications by the participants. The fifth bar displays a weighted average, giving more weight to self-administered classifications. The variability in skin type classification by different reviewers is evident, highlighting the importance of using pragmatic methods when using the Fitzpatrick scale.

meters (kg/m^2) [215]. The resultant metric is generally categorised for interpretative convenience: a BMI below 18.5 is classified as underweight, between 18.5 and 24.9 as healthy, between 25 and 29.9 as overweight, and 30 or above as obese [216]. Notably, the National Health Service (NHS) has adjusted these categorisation boundaries for Black, Asian, and some other minority ethnic groups to account for the differential health risks associated with BMI in these populations [216]. Additionally, while the Quetelet Index serves as a globally recognised metric, it is subject to criticism for its failure to account for muscle mass, potentially providing a misrepresentative portrayal of an individual's percentage of body fat and, consequently, their overall health and fitness status [215].

In the context of physical exercise, heart rate serves as a pivotal indicator of the level of effort being exerted by an individual, as it exhibits a proportionate increase with intensifying physical activity. Consequently, using raw heart rate values as a consistent and reliable measure of exercise intensity becomes challenging across different age groups due to the inherent variations in the heart rate range caused by the ageing process. The concept of effort, when represented in terms of HR, is typically described as a percentage of the individual's maximal heart rate (MHR) [217]. Various methodologies have been proposed to estimate MHR; while the gold standard involves conducting a Maximal Aerobic Test, which pushes an individual to their absolute physical limits, this

is not always feasible in all research or practical contexts. Consequently, several equations have been derived to estimate MHR, including the widely adopted Fox equation $(220 - Age)$, despite its recognised variability. Alternative formulations include the Tanaka equation $(208 - 0.7 \times Age)$ and the Fairbarn equation, which is gender-specific ($208 - 0.8 \times Age$ for males and $201 - 0.63 \times Age$ for females), amongst others [217]. In this study, the Fox equation was employed to estimate MHR and, subsequently, effort, primarily due to its widespread adoption and ease of application across diverse populations.

4.5 Summary

The chapter meticulously details the protocol design, aimed at overcoming limitations in existing dataset protocols. It highlights the inclusion of active rest and cycling phases to capture low- and high-heart-rate scenarios with varying movement patterns. The cohort section elaborates on participant measurements and methodologies. Device selection is also discussed, focusing on the use of ECG and wrist-worn devices, along with software for activity timings.

The chapter proceeds to examine signal alignment techniques, selecting cross-correlation for its simplicity and efficacy and detailing a comprehensive approach to ensure accurate signal alignment. Heart rate extraction from ECG is given special attention, essential for evaluating PPG heart rate estimation algorithms. A novel frequency domain method was developed for more accurate heart rate extraction, outperforming existing methods in comparative analysis. The chapter also addresses skin tone classification, which is key in this research. Despite debates on the Fitzpatrick scale's accuracy, a pragmatic approach was adopted, using a diverse panel for skin tone classification, revealing the subjectivity in the process. Other computed metrics like BMI and MHR-derived physical effort are also discussed.

The collected dataset comprises data from 20 participants selected from an initial pool of 30 due to data collection and signal integrity issues. The cohort has an age distribution of 25.9 ± 8.2 years and includes 13 female and 7 male participants. Skin types are evenly split with 10 participants having Fitzpatrick skin types I-III and 10 having Fitzpatrick skin types IV-VI. The dataset includes close to 15 hours of data, resulting in a total of 26,442 samples of 8-second windows with 2-second slides. Notably, the dataset features the largest representation of heart rates in the 160-180 BPM range across available datasets, with close to 6,000 samples indicating physical effort rates of 60% or higher. Furthermore, it includes the most comprehensive collection of PPG wavelengths, with two channels each for blue, green, red, and IR. The data collection

protocol incorporates erratic wrist movements, cross-over effects, motion-free periods, and periods of increased heart rates with minimal motion. These attributes provide a robust foundation for evaluating wrist-worn PPG heart rate estimation methods.

Chapter 5

Analysis of A Wrist-worn Multi-wavelength Photoplethysmography Heart Rate Monitoring Dataset

The preceding chapter gave details of the design, collection and processing of the wrist-worn multi-wavelength Photoplethysmography (PPG) heart rate monitoring dataset. This chapter evaluates the collected dataset, primarily focusing on its efficacy in reflecting the diversity and robustness intended in its design. The dataset, designed for validating heart rate estimation methods, captures wrist-worn PPG signals across a diverse cohort—accounting for variations in biological sex, skin melanin content, age, and BMI—under different motion types and physical effort levels. It utilises a chest-worn electrocardiogram (ECG) and a wrist-worn multi-wavelength PPG device, which collects two channels of PPG signals from the two photodiodes, using blue, green, red, and IR LEDs. Additionally, the wrist-worn device is equipped with sensors such as an accelerometer.

The assessment begins with a comparative analysis of existing single-wavelength wrist-worn PPG heart rate estimation datasets (see Table 2.3), focusing on how well the dataset represents heart rate and physical effort levels within the cohort. This is followed by a thorough evaluation of the accelerometer's effectiveness as a reference for PPG motion artifacts across different types of movement. The analysis further includes an examination of various SQIs to evaluate the dataset's reliability under diverse conditions. The final phase involves using a range of beat detectors to determine the optimal wavelength for wrist-worn heart rate estimation and to test the compatibility of conventional algorithms with this multifaceted dataset. The outcomes of this investigation

aim to affirm whether the dataset successfully meets its intended objectives, thereby contributing to wrist-worn PPG heart rate monitoring research.

5.1 Comparative Dataset Analysis

This section evaluates the dataset based on cohort demographics, heart rate measurements, motion, and skin temperature. It involves a comparison with several single-wavelength PPG heart rate estimation datasets, including IEEE Train [126], IEEE Test [126], BAMI 1 [127], BAMI 2 [127], and PPG DaLiA [130]. As outlined in Section 2.2.1, two additional datasets of this type were considered but excluded from this analysis due to their limited sample size and inconsistent protocols across subjects, which did not meet the standards required for a robust evaluation.

5.1.1 Cohort

A primary objective of this dataset is to facilitate a comprehensive assessment of the impact of demographic variables on the accuracy of PPG heart rate estimation algorithms and evaluate the signal quality of the PPG signals obtained. A critical aspect of this endeavour is ensuring a diverse cohort representation. As delineated in Table 5.1, it is observed that a substantial number of datasets do not provide detailed cohort demographics, thereby constraining the scope for analytical exploration. Only datasets such as IEEE Test [126] and PPG DaLiA [130] offer such demographic information.

In examining age diversity, the PPG DaLiA dataset stands out with its broad age range, with an mean age of 31 years and a standard deviation of 10 years. This contrasts with the dataset collected explicitly for this study, which tends to comprise younger participants, evidenced by a mean age of 26 years and a standard deviation of 8 years. Regarding biological sex distribution, the dataset collected for this study demonstrates a balanced representation of both sexes, albeit with a marginal inclination towards female participants. This starkly contrasts the IEEE Test dataset, which exhibits a pronounced male bias.

The diversity of the Fitzpatrick skin types in the collected dataset is particularly noteworthy. It encompasses all categories on the Fitzpatrick scale, albeit with a caveat: three of the six skin types are represented by merely two participants each. To address within-group variance and to better understand factors affecting PPG sensing accuracy, skin melanin content was categorised into two groups: low (Fitzpatrick types 1, 2, and 3) and high (types 4, 5, and 6), with each group comprising ten participants. From now on in this thesis, skin melanin content will be utilised as a primary variable for

	IEEE Train (incl. 13) [126]	**IEEE Test [126]**	**BAMI 1 [127]**	**BAMI 2 [127]**	**PPG DaLiA [130]**	**MW PPG HR (This Work)**
Age	—	25 ± 12	—	—	31 ± 10	26 ± 8
Biological Sex	—	Female: 1 Male: 9	—	—	Female: 8 Male: 7	Female: 13 Male: 7
Fitzpatrick Skin Type	—	—	—	—	II: 1 III: 11 IV: 3	I: 2 II: 6 III: 2 IV: 5 V: 3 VI: 2
BMI	—	22.4 ± 2.9	—	—	22.3 ± 1.8	22.8 ± 3.0

TABLE 5.1: Comparison of Cohorts Across All Utilised Datasets. The table compares cohorts in terms of age, biological sex, Fitzpatrick skin type, and BMI. The MW PPG HR dataset (this work) has the most diversity in Fitzpatrick skin type, with representation of all six types. Additionally, the table highlights the amount of non-reported demographics for each dataset.

analysis due to these considerations. Finally, the BMI parameter exhibits ample diversity within the collected dataset, with a slightly higher average BMI but a more significant standard deviation than the other datasets, highlighting the range of body compositions encompassed in the study.

5.1.2 Heart Rate

The dataset was designed to encompass a broad spectrum of heart rate values, essential for analysing PPG heart rate estimation algorithms across different heart rate intervals. Table 5.2 presents a detailed breakdown of the sample distribution across various heart rate intervals for each dataset. The PPG DaLiA dataset peaks in the 60-80 BPM range, typical for adult resting heart rates, while the collected dataset shows highest representation in 80-100 BPM and extends to 180-200 BPM. This broader range facilitates rigorous evaluation of the methods' robustness across diverse physiological states.

In contrast, treadmill-based protocol datasets, such as IEEE Train, BAMI 1, and BAMI 2, show their most significant occurrence of samples in the 120-160 BPM range, corresponding to moderate to intense physical activity. The collected dataset particularly stands out for its substantial inclusion of the 160-200 BPM range (2,358 samples), highlighting heart rates associated with vigorous activity and demonstrating the dataset's coverage of various physiological conditions. The PPG DaLiA dataset is the largest, with about 65,000 samples, while the collected dataset, with around 26,000 samples,

Heart Rates (BPM)	IEEE Train (incl. 13) [126]	IEEE Test [126]	BAMI 1 [127]	BAMI 2 [127]	PPG DaLiA [130]	MW PPG HR (This Work)
0 - 40	0	0	0	2	0	0
40 - 60	0	6	0	4	3,746	5
60 - 80	76	235	252	95	21,585	3,298
80 - 100	193	240	1,188	836	22,374	9,087
100 - 120	263	153	2,475	2,325	9,884	5,859
120 - 140	391	348	2,607	2,480	4,878	3,302
140 - 160	696	247	2,097	1,813	1,679	2,533
160 - 180	256	99	574	670	512	1,936
180 - 200	0	0	21	78	39	422
Total	1,875	1,328	9,214	8,303	64,697	26,442

TABLE 5.2: Comparison of Heart Rate Samples Across Utilised Datasets. This table shows the number of heart rate samples across various BPM ranges. The MW PPG HR dataset is particularly notable for its extensive coverage, especially in the 60-180 BPM range, which is proportionality underrepresented in other datasets. This broad sample range aims to enhance the validation and verification of heart rate estimation methods across both extreme and normal heart rate ranges

is the second-largest. This extensive data across various heart rate ranges is key for a thorough and precise PPG heart rate estimation research evaluation.

Analysing physical effort as a percentage of MHR across datasets like IEEE Test, PPG DaLiA, and MW PPG HR (This Work) reveals critical trends in exertion levels. It's important to consider age's impact on MHR, as heart rate values alone can be misleading. Detailed in Table 5.3, the IEEE Test dataset predominantly features samples in the higher exertion ranges, with the majority (45.9%) in the 60-80% MHR range, followed by 30.4% in the 40-60% MHR range. This indicates a focus on moderate to high exertion levels. The PPG DaLiA dataset, in contrast, is concentrated in the low to moderate exertion range, with 57.3% of samples in the 40-60% MHR range and a significant 29.0% in the 20-40% MHR range, showing a preference for moderate physical effort.

The MW PPG HR dataset (this work) displays a broader distribution but leans towards moderate exertion, with the 40-60% MHR range accounting for 57.4% of its samples. It also includes a notable representation in the high exertion range (10.7% in the 80-100% MHR range). These trends highlight the varied focus of each dataset, with IEEE Test and MW PPG HR (This Work) covering a more comprehensive range of exertion levels and PPG DaLiA focusing more on moderate exertion. This diversity in physical effort levels is key for comprehensively evaluating PPG heart rate estimation algorithms.

Physical Effort (% of MHR)	**IEEE Test [126]**	**PPG DaLiA [130]**	**MW PPG HR (This Work)**
0 - 20	0	0	0
20 - 40	226	18,776	2,473
40 - 60	404	37,074	15,132
60 - 80	609	7,956	5,949
80 - 100	89	891	2,831

TABLE 5.3: Comparison of Sample Counts per Physical Effort Level Across Utilised Datasets Reporting Age. Physical effort is defined as the percentage of heart rate over MHR, derived from the Fox Equation. Notably, the MW PPG HR dataset demonstrates a proportionally higher number of samples at elevated effort levels ($\geq$ 60% MHR) compared to the other datasets analysed.

5.1.3 Motion

In PPG signal analysis, accelerometers serve as a critical tool for motion reference and artefact reduction, yet their efficacy in capturing diverse motion types and intensities merits investigation. Under the premise that heart rate escalates with increased physical workload, particularly in treadmill-based protocols, it is hypothesised that higher heart rates correspond to intensified motion. However, this correlation may not hold for cycling or wrist/arm movements, where high motion can coexist with lower heart rates.

To evaluate how well the accelerometer captures different motion the Euclidean norm of the three accelerometer axes is calculated for each segmented window (8-second duration with a 2-second shift) to quantify motion intensity. The mean value of these windowed segments is then taken as the indicator of accelerometer intensity. Subsequently, min-max normalisation is applied to these values, facilitating comparison across different datasets.

As illustrated in Figure 5.1, the accelerometer effectively captured the escalation in motion intensity concurrent with increased activity intensity for treadmill-based protocols. This finding is further supported by a positive correlation (0.52) between accelerometer intensity and true HR, notably with an expanded spread of accelerometer intensities observed beyond 100 BPM.

The collected dataset, particularly during its treadmill-based running phase, also demonstrated a range of accelerometer intensities, as shown in Figure 5.2. However, a weaker correlation (0.36) was observed between true heart rate and accelerometer intensity, with a noticeably broader spread past 100 BPM. Contrarily, the accelerometer's effectiveness in capturing motion types diminished in wrist-based movements and cycling scenarios,

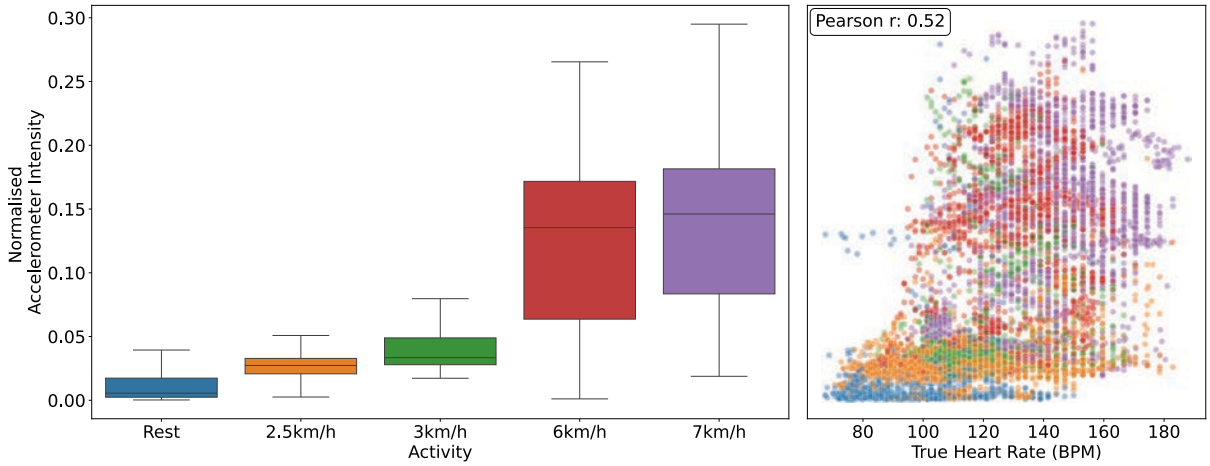


FIGURE 5.1: Relationship between Normalised Accelerometer Intensity, Activity, and True Heart Rate for BAMI 1 Dataset [127]. The figure presents box plots of normalised accelerometer intensity—derived by calculating the Euclidean norm of the 3D acceleration signal and applying min-max normalisation across datasets. The box plots show the median, IQR, and 1.5 IQR whiskers. Notably, a broader distribution is observed at running speeds of 6 km/h and above. Additionally, there is a greater spread of normalised accelerometer intensity at True Heart Rates (ECG-derived) exceeding 100 BPM.

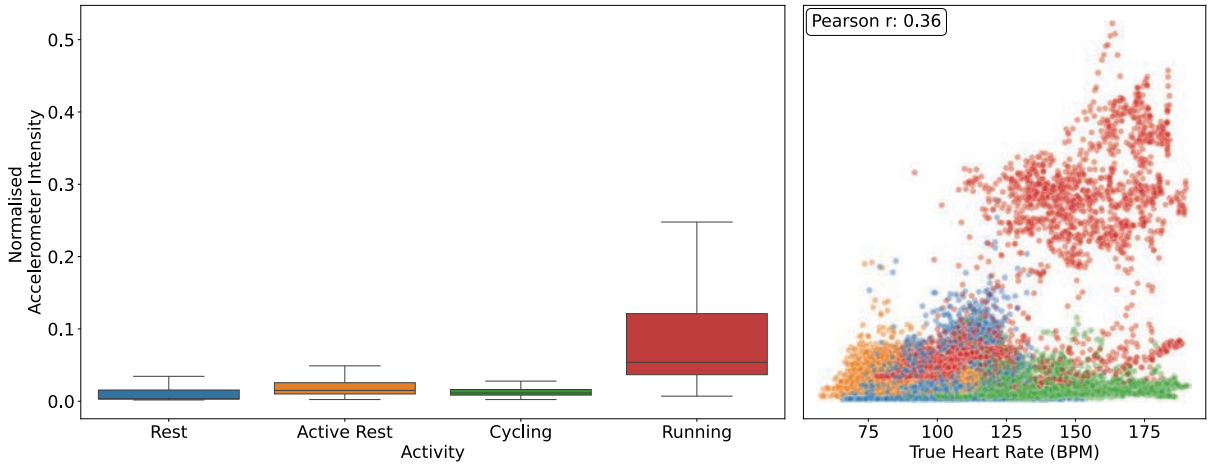


FIGURE 5.2: Relationship between Normalised Accelerometer Intensity, Activity, and True Heart Rate for MW PPG HR Dataset (This Work). The figure presents box plots of normalised accelerometer intensity, calculated using the Euclidean norm of the 3D acceleration signal and applying min-max normalisation across datasets. The box plots show the median, IQR, and 1.5 IQR whiskers. Running shows the widest spread of accelerometer intensity, while active rest has a distribution similar to rest and cycling, indicating potential limitations in capturing all movement types. A similar trend of greater spread is observed at True Heart Rates (ECG-derived) above 100 BPM, primarily from running, whereas minimal motion is recorded during cycling, as expected.

with notable differences in distribution observed between rest, cycling, and active rest conditions.

These observations underscore that while accelerometers are useful for tracking certain motion types, they may not comprehensively represent all motion-induced noise within PPG signals. This highlights the need for cautious interpretation of accelerometer data, especially in activities where motion patterns, such as erratic wrist-movements, differ significantly from treadmill-based protocols.

5.1.4 Local Skin Temperature

As highlighted in Section 2.1.3, skin temperature is a factor that can significantly impact the quality of PPG signals. However, upon examining the range of skin temperature values recorded during the study, it was observed that there was minimal variation in these temperatures across different activities. Notably, the active rest phase exhibited the broadest range of skin temperature values. This variation can be attributed to the fact that active rest was the initial phase of the protocol, during which subjects were exposed to the colder conditions prevalent during the winter season. This exposure potentially led to a greater spread of skin temperature than in other phases.

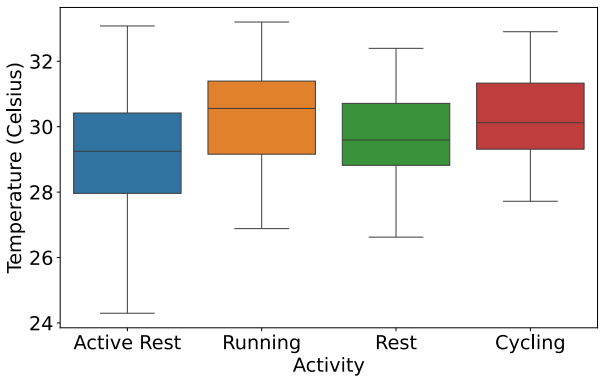


FIGURE 5.3: Distribution of Local Skin Temperature across Different Activities. The box plots show the median, IQR, and 1.5 IQR whiskers. Activities are ordered as: active rest, running, rest, and cycling. Winter data collection led to lower temperatures, especially during the first phase.

Given the overall lack of significant variation in skin temperature throughout the various phases of the study, it has been determined that an in-depth exploration of skin temperature as a variable affecting PPG signal quality will not be included in this thesis. The minimal variability observed suggests that skin temperature, under the conditions of this study, does not markedly influence the PPG signal quality to a degree that necessitates further investigation within the scope of this research.

5.2 Multi-wavelength Photoplethysmography Signal Quality Analysis

This section will evaluate various Signal Quality Indices (SQIs) to determine their efficacy in assessing the quality of collected PPG signals, thus addressing objective 3. This analysis is based on the premise that PPG signal quality will likely vary with physical activity intensity. For instance, signal quality is anticipated to degrade during high-intensity activities like running on a treadmill compared to lower-intensity activities such as walking or resting. The aim is to establish whether these SQIs accurately reflect changes in signal quality under different physical conditions.

The effectiveness of SQIs will be analysed in two key areas: firstly, assessing how PPG signal quality varies across different light wavelengths, as wavelength significantly impacts signal accuracy due to differences in light penetration and absorption by skin and blood. Secondly, exploring how these SQIs perform under different motion types, which is key for understanding their reliability amidst motion-induced noise. Additionally, the impact of participants' demographic characteristics on PPG signal quality, as measured by the SQIs, will be examined. The detailed versions of the software used are outlined in Section 3.5

5.2.1 PPG and Accelerometer Correlation

The relationship between the accelerometer and PPG signals is a critical aspect of this study, focusing on the premise that the motion detected by the accelerometer should correspond to the motion artefacts present in the PPG signal. To evaluate this, a specific SQI was developed involving the calculation of the Euclidean norm of the accelerometer axes for each windowed segment. This accelerometer intensity signal was then correlated with the corresponding windows of the PPG signal to establish the SQI.

However, the findings, as illustrated in Figure 5.4, reveal an unexpected pattern in the correlation between accelerometer and PPG signals across different activities. Contrary to initial assumptions, cycling exhibited the highest spread of correlation values, suggesting that it contains the most motion-induced artifacts within its PPG signal. Interestingly, the motion correlation for running was similar to that of rest, which deviates from the expected trend of increased motion in more intense activities. This unexpected similarity could potentially be attributed to a 'cross-over effect' in running.

Comparing true heart rates with this SQI showed a weak correlation, lacking a discernible trend. This outcome indicates that this particular SQI may not effectively quantify the quality of the collected PPG signals. Specifically, it is anticipated that running,

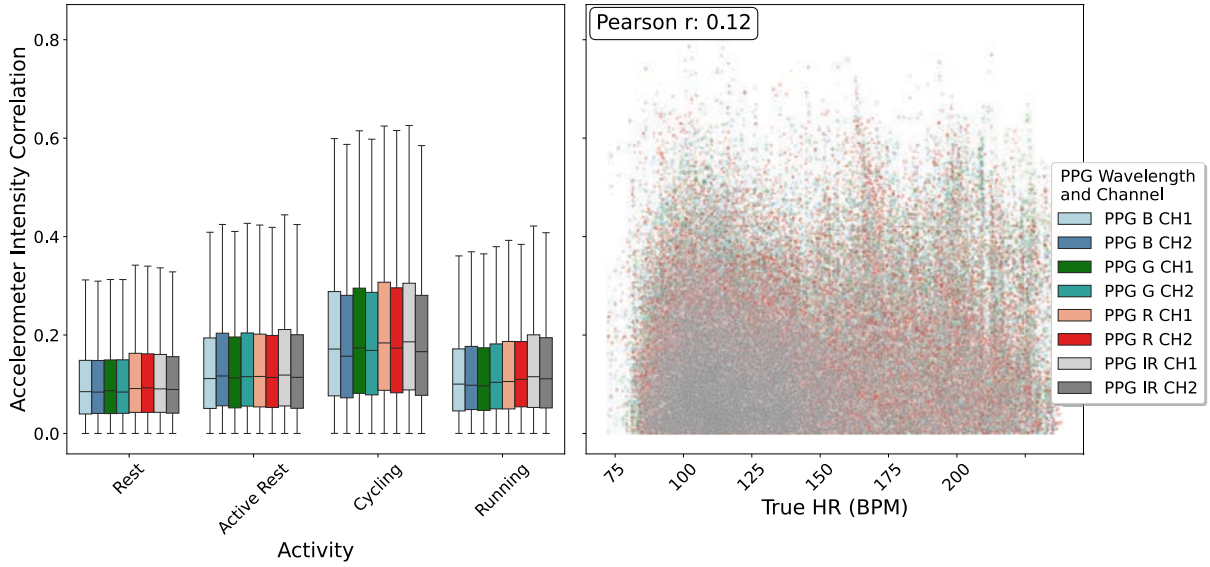


FIGURE 5.4: Relationship Between Accelerometer Intensity and PPG Correlation, Activity, and True Heart Rate for MW PPG HR Dataset (This Work). This figure shows box-plots of the correlation between accelerometer intensity (calculated as the Euclidean norm of the three axes) and the PPG signal. The box plots show the median, IQR, and 1.5 IQR whiskers. PPG signals are analysed by LED wavelength and photodiode channel. The correlation's relationship with ECG-derived heart rate is shown. Lab-based protocols often find motion increases with heart rate.

typically associated with significant motion, would exhibit the highest motion artefacts in the PPG signal. However, the results show the opposite, suggesting this SQI may not accurately reflect the expected motion impact. This discrepancy also implies that the accelerometer data used as a motion reference in this SQI might not appropriately represent the motion artifacts in the PPG signal. Furthermore, the observation that the quality of longer wavelengths during motion is not significantly worse than that of shorter wavelengths challenges common assumptions about wavelength-dependent signal deterioration due to motion.

5.2.2 Signal-to-Noise Ratio

Signal-to-noise ratio (SNR) is a widely employed signal quality index (SQI) across various domains, including PPG signal analysis, for assessing the proportion of the desired "signal" to background "noise" [138]. Elgendi's research examined several SQIs, with SNR being one, although in this study skewness was determined the optimal SQI [138]. In Elgendi's approach, the "signal" component is defined as the standard deviation of the absolute values of a filtered PPG signal, whereas the "noise" component is the standard deviation of the filtered PPG signal itself. [138].

However, this definition reveals counter-intuitive results, as shown in Figure 5.5. For example, activities with significant aperiodic motion, like table soccer, exhibit higher SNR, whereas more static conditions, like sitting, show lower SNR. Additionally, a minimal correlation (0.1) is observed between SNR and the true HR, challenging the effectiveness of this SNR definition in accurately reflecting noise levels.

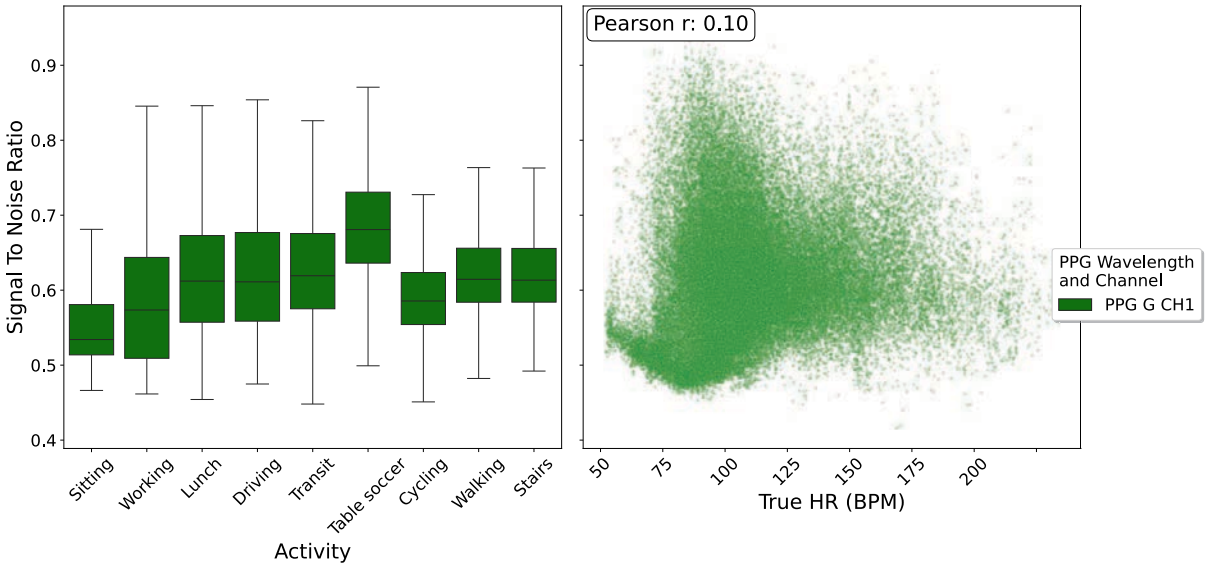


FIGURE 5.5: Relationship between Elgendi Signal-to-Noise Ratio, Activity and True Heart Rate for PPG DaLiA Dataset [130]. This figure shows box-plots of the correlation between Elgendi Signal-to-Noise Ratio and the PPG signal. The box plots show the median, IQR, and 1.5 IQR whiskers. PPG signals are analysed by LED wavelength and photodiode channel. The correlation's relationship with ECG-derived heart rate is shown. Lab-based protocols often find motion increases with heart rate.

Another approach to defining SNR utilises ECG-derived heart rate values as a means to separate signal from noise [128]. Our implementation of this method begins by segmenting the PPG signal into windows using the Tukey window function to mitigate transient effects. It then applies a bandpass Butterworth filter within a specific frequency range (0.5 - 4 Hz). Subsequently, harmonics based on the ECG-derived heart rate are identified and eliminated using a band-stop Butterworth filter, carefully removing harmonics beyond a certain threshold, as dictated by the Nyquist criterion. This process aims to isolate the noise component within the signal, which is then subtracted from the original signal to obtain a 'clean' signal. The SNR is subsequently computed by comparing the power of the 'clean' signal and the noise components, expressed in decibels on a logarithmic scale.

This refined SNR calculation method exhibits more consistent and expected trends, as illustrated in Fig 5.6. Lower-intensity activities like sitting display higher SNR values, while more motion based activities such as table soccer and stair climbing show lower

SNR values. Significantly, a stronger negative correlation between true heart rate and SNR is observed, indicating a more accurate and meaningful representation of the relationship between SNR and activity intensity in PPG signal analysis.

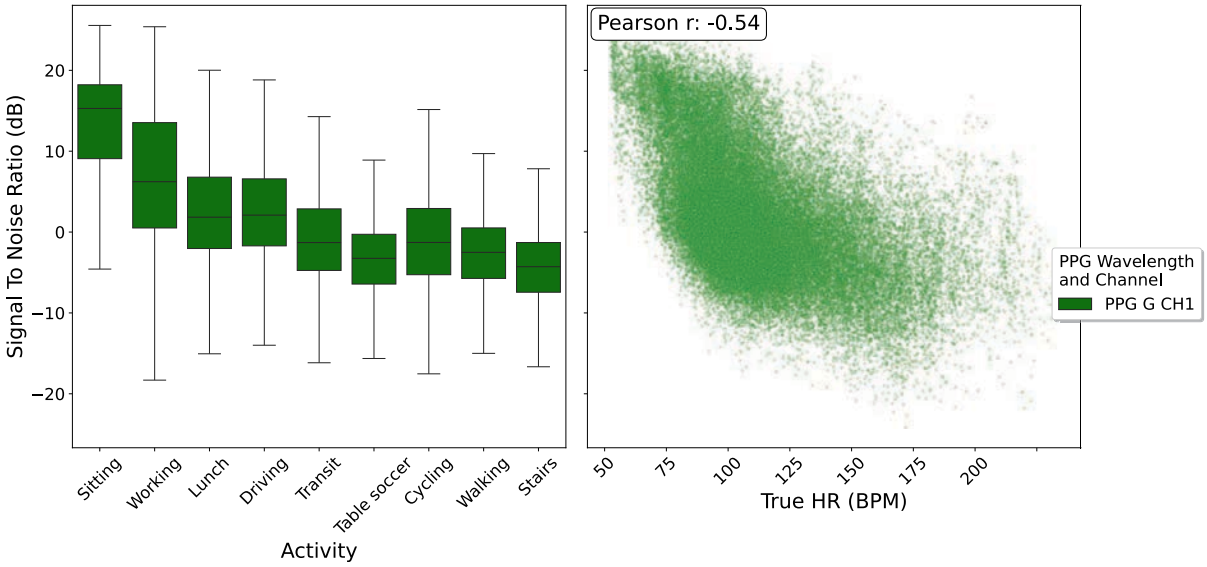


FIGURE 5.6: Relationship Between The Proposed ECG-derived Signal-to-Noise Ratio, Activity and True Heart Rate for PPG DaLiA [130]. This figure shows box-plots of the correlation between the proposed ECG-derived Signal-to-Noise Ratio and the PPG signal. The box plots show the median, IQR, and 1.5 IQR whiskers. PPG signals are analysed by LED wavelength and photodiode channel. The correlation's relationship with ECG-derived heart rate is shown. Lab-based protocols often find motion increases with heart rate.

The analysis of the ECG-derived SNR on the collected dataset across various activity phases suggests that longer wavelengths exhibit poorer SNR than shorter wavelengths, as illustrated in Figure 5.7. This pattern is particularly pronounced during the cycling and resting phases, where the discrepancy in SNR between the different wavelengths is more marked. Conversely, during running and active rest, the differences in SNR among various wavelengths are less substantial, indicating a potential interaction effect between activity type and the impact of wavelength on SNR.

The trend within shorter wavelengths indicates a decline in SNR as the activity intensity increases, with the cycling phase displaying the broadest distribution of SNR values. This could be attributed to the nature of cycling, which may induce more aperiodic motion or physiological changes that affect the PPG signal differently than other activities. The moderate negative correlation between SNR and true heart rate underscores the inverse relationship between physiological exertion—as evidenced by increased HR—and signal quality. The correlation coefficient of -0.41 indicates that as the heart rate rises, possibly due to increased physical activity, the SNR tends to

diminish, reflecting a degradation in signal quality amidst heightened physiological activity.

Developing an effective SQI for wrist-worn PPG sensing presents challenges due to the complex interplay of signal and noise. The chapter examines three SQIs, ultimately determining the ECG-derived SNR as most effective in reflecting the expected data trends and assumptions. Consequently, this SQI will be adopted in subsequent chapters to assess proposed methodologies, providing a robust tool for evaluating the quality and reliability of PPG-based physiological estimation methods in wrist-worn devices.

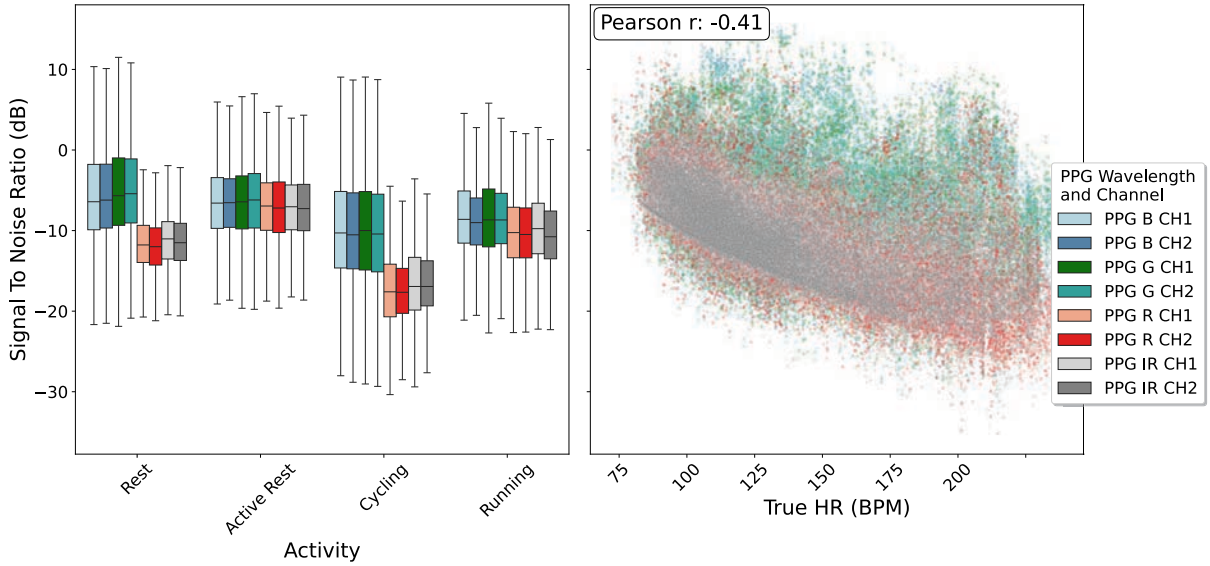


FIGURE 5.7: Relationship Between The Proposed ECG-derived Signal-to-Noise Ratio, Activity and True Heart Rate for MW PPG HR (This Work). This figure shows box-plots of the correlation between the proposed ECG-derived Signal-to-Noise Ratio and the PPG signal. The box plots show the median, IQR, and 1.5 IQR whiskers. PPG signals are analysed by LED wavelength and photodiode channel. The correlation's relationship with ECG-derived heart rate is shown. Lab-based protocols often find motion increases with heart rate..

5.3 Multi-wavelength Photoplethysmography Beat Detectors Analysis

In the domain of PPG-based heart rate monitoring, the prominence of the systolic peak in the PPG waveform is of utmost importance. The sequential analysis of these systolic peaks (beats) facilitates the derivation of vital physiological indicators such as heart rate and heart rate variability, which are key for cardiovascular monitoring [124]. Consequently, the distinctness and clarity of the systolic peaks serve as a significant measure of the quality of the PPG signal. This section aims to explore whether there is a

trend between wavelength and PPG beat detection performance, assess the robustness of beat detectors to different types and intensities of motion, and analyse how demographic variations impact PPG beat detection accuracy. The framework developed by Charlton et al. for evaluating fifteen open-source systolic peak detectors provides the foundation for this investigation [124].

To align with the results of Charlton et al., the same evaluation parameters were adopted [124]. The PPG signals were processed at a sampling rate of 100 Hz, filtered through a bandpass filter from 0.67 to 8.0 Hz, and segmented into 20-second intervals with a 5-second overlap. Only PPG windows that met predefined quality standards were included, and the mid-amplitude points of the detected peaks were the focus of the subsequent analysis. Reference ECG beats were delineated using the 'jqrs' and 'rpeakdetect' QRS detection algorithms based on well-established techniques [206, 207]. A consistency check for ECG beats was enforced, accepting only those detected by both QRS methods within a 150 ms interval, thus excluding discordant ECG beats. Following the detection of PPG beats, an alignment with the ECG beats was performed, ensuring that both were within a 150 ms concordance range. Furthermore, as discussed in Section 4.4.2, compromised ECG signal segments were omitted from the analysis. The detailed versions of the software used are outlined in Section 3.5. To assess the effectiveness of the beat detection algorithms in estimating heart rate, the heart rate values calculated from the detected beats in the PPG signal will be compared with the heart rate values derived from the ECG measurements.

5.3.1 Activity and Wavelength

Figure 5.8 presents the MAPE analysis for each activity in the analysis. As expected, the rest phase exhibited the most accurate results across all detectors. When comparing short and long wavelengths, algorithms ATM, COppg, ERMA, and IMS demonstrated similar MAPE values. Interestingly, PWD displayed marginally improved performance with longer wavelengths compared to shorter ones. Conversely, the remaining detectors performed better with shorter wavelengths.

During active rest and cycling activities, MAPE values remained consistent across various wavelengths. QPPG, ERMA, MSPTD, and WFD consistently displayed the lowest MAPE distributions in these phases. Running, characterised by the most significant motion, produced the least accurate results for all beat detectors.

Table 5.4 provides a comprehensive overview of the median absolute errors (AE) for various PPG wavelengths and activities. The data reveals that shorter wavelengths, particularly blue and green, consistently resulted in lower median errors across different activities.

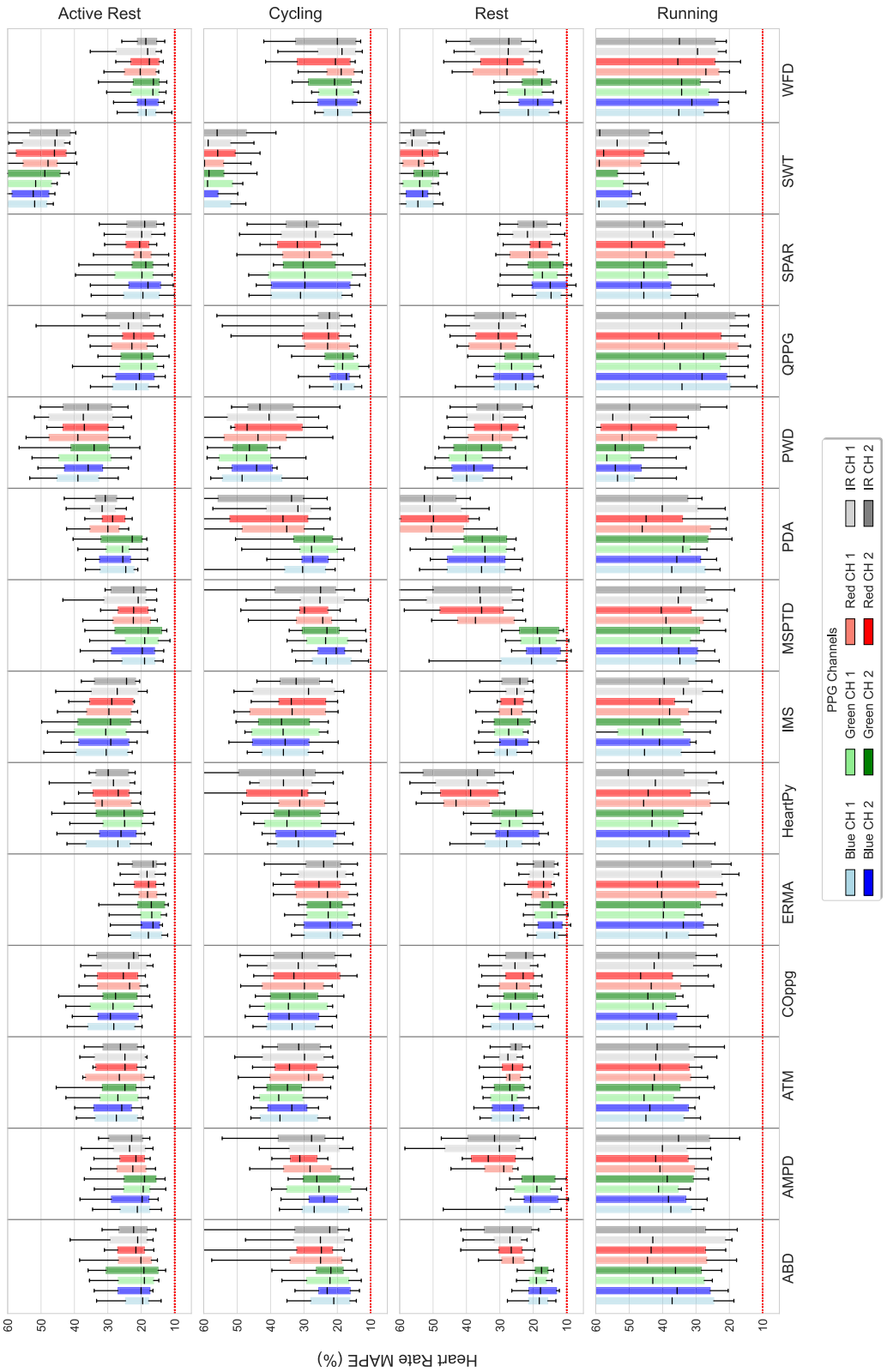


FIGURE 5.8: Performance of PPG Beat Detectors Heart Rate Estimation Across Various Activities for MW PPG HR dataset (This Work). This figure presents box-plots of the absolute percentage error (MAPE) distribution for 15 PPG beat detectors across different activities. The box plots show the median, IQR, and 1.5 IQR whiskers. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. The red line denotes the AAMI standard of 10% MAPE. The code base for the PPG beat detectors was from Charlton et al. [124].

PPG Channel	Active Rest		Cycling		Rest		Running	
	Detector	Median AE (BPM)	Detector	Median AE (BPM)	Detector	Median AE (BPM)	Detector	Median AE (BPM)
Blue CH 1	ERMA	15.1	QPPG	17.4	**ERMA**	**11.5**	QPPG	33.2
Blue CH 2	ERMA	13.6	QPPG	17.3	ERMA	11.8	**QPPG**	**29.7**
Green CH 1	ERMA	13.7	**QPPG**	**17.2**	ERMA	11.9	QPPG	34.5
Green CH 2	**WFD**	**12.9**	QPPG	19.2	SPAR	12.2	QPPG	29.8
Red CH 1	ERMA	14.4	WFD	17.4	ERMA	14.5	QPPG	38.4
Red CH 2	ERMA	14.5	WFD	20.8	ERMA	14.1	WFD	39.1
IR CH 1	ERMA	14.7	ERMA	19.1	ERMA	13.9	WFD	31.0
IR CH 2	ERMA	13.6	WFD	19.4	ERMA	15.0	QPPG	39.3

TABLE 5.4: Activity-Based Performance Analysis of PPG Beat Detectors Heart Rate Estimation Across Various Wavelength for MW PPG HR dataset (This Work). This table shows the median absolute error (AE) in BPM for different PPG heart rate detectors across various activities and PPG wavelengths. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. Each row represents a specific PPG channel, and each column under the activity headings lists the detector with the corresponding median AE. The code base for the PPG beat detectors was from Charlton et al. [124]. **Bold** indicates lowest median absolute error for that activity.

For active rest, the green wavelength (Channel 2) paired with the WFD detector achieved the lowest AE at 12.9 BPM, indicating superior accuracy. In comparison, the blue wavelength (Channel 2) with the ERMA detector was also effective, with a slightly higher AE of 13.6 BPM.

During cycling, the green wavelength (Channel 1) combined with the QPPG detector provided the best performance, showing the lowest AE at 17.2 BPM. Blue light (Channel 2) with the QPPG detector also performed well, with a marginally higher AE of 17.3 BPM.

In a resting state, the blue wavelength (Channel 1) and the ERMA detector were the most accurate, achieving the lowest AE of 11.5 BPM. Green light (Channel 2) paired with the SPAR detector showed a slightly higher AE of 12.2 BPM.

For more intense activities like running, the blue wavelength (Channel 2) coupled with the QPPG detector resulted in the lowest AE of 29.7 BPM, indicating better accuracy under these conditions. Green light (Channel 2) with the QPPG detector followed closely, with an AE of 29.8 BPM.

Overall, the differences in accuracy between blue and green wavelengths across activities were generally within 1 BPM. The ERMA detector showed consistent performance

across various wavelengths for active rest and rest. However, for more dynamic activities like cycling and running, the QPPG detector emerged as the most accurate, with blue light performing slightly better for running and green light excelling in cycling. This analysis underscores the importance of selecting appropriate PPG wavelengths and detectors tailored to specific activities to achieve optimal heart rate estimation accuracy.

5.3.2 Biological Sex and Wavelength

Figure 5.9 presents the comparative analysis of PPG beat detector performance across biological sexes. While most detectors exhibited generally uniform performance, some notable exceptions emerged. Detectors like AMPD, ABD, and ATM displayed a significantly tighter IQR for males compared to females. This indicates more consistent heart rate estimation accuracy in males for these specific detectors.

Wavelength also appears to influence detector performance. Eight out of the fifteen detectors analysed demonstrated a broader MAPE distribution when using longer wavelengths (IR). This suggests that their accuracy is lower with IR light compared to shorter blue and green wavelengths. Conversely, detectors like PWD and ATM maintained consistent performance across the entire wavelength spectrum, highlighting their robustness to wavelength variations.

Table 5.5 provides a detailed analysis of the median absolute error (AE) in beats per minute (BPM) for different PPG wavelengths and biological sexes. The table highlights that blue light (Channel 1) yielded the most accurate results for both females and males, with QPPG as the best-performing beat detector. For females, QPPG achieved a median AE of 23.3 BPM, while for males, it achieved a lower AE of 19.7 BPM, indicating a 3.6 BPM higher accuracy for males.

When examining other wavelengths, green light (Channel 1) showed a slightly higher AE for females at 25.4 BPM, with QPPG as the detector, compared to 20.5 BPM for males. Interestingly, for green light (Channel 2), MSPTD was the most accurate detector for males, with an AE of 20.6 BPM, while QPPG was still the best for females with a similar AE of 23.4 BPM.

For red and infrared wavelengths, the errors increased for both sexes. Red light (Channel 1) had an AE of 30.7 BPM for females and 25.6 BPM for males, with WFD as the most accurate detector for both. Similarly, infrared light (Channel 1) resulted in an AE of 29.8 BPM for females using WFD and 26.6 BPM for males using ERMA.

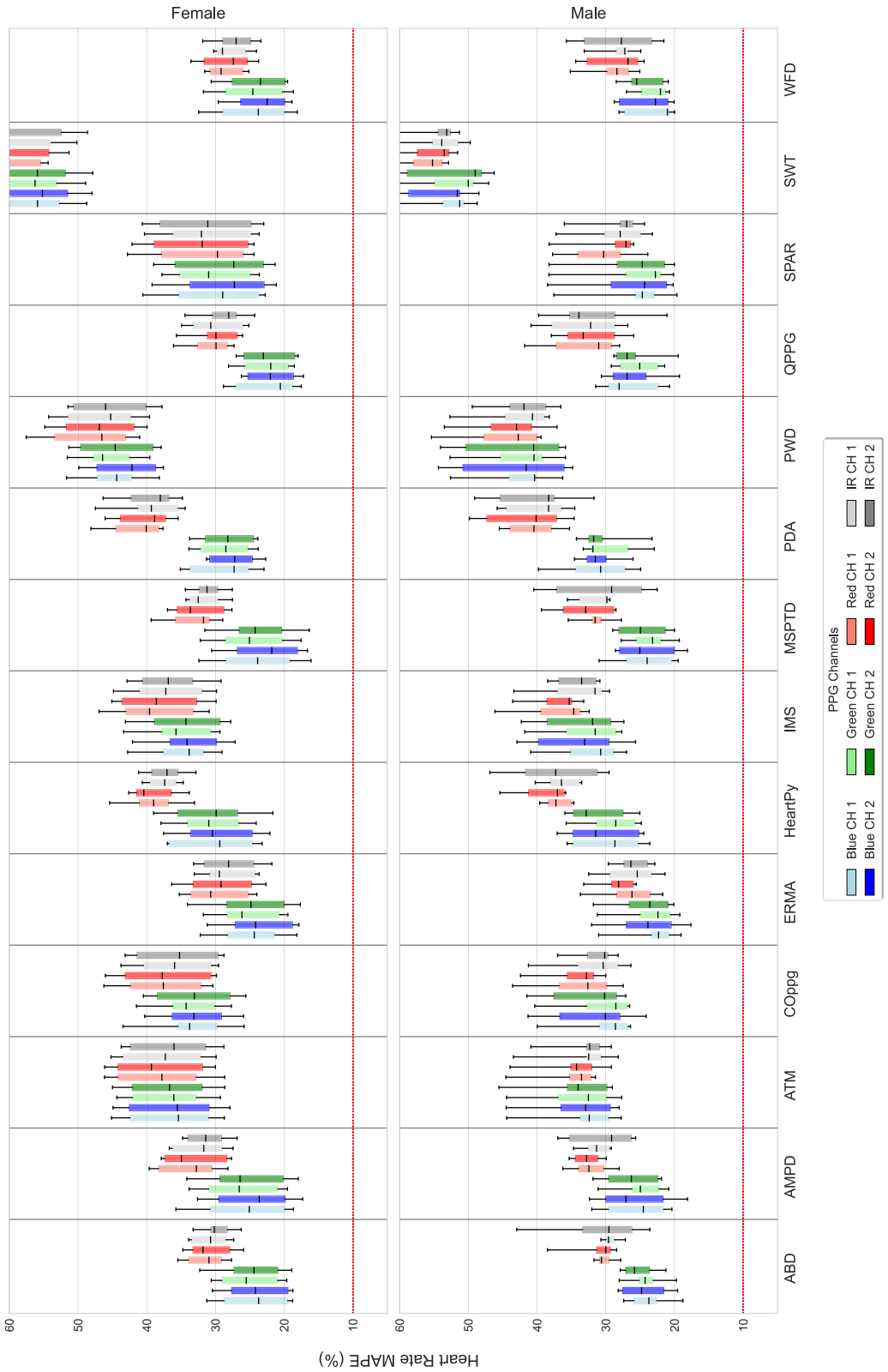


FIGURE 5.9: Performance of PPG Beat Detectors Heart Rate Estimation by Biological Sex for MW PPG HR dataset (This Work). This figure presents box-plots of the absolute percentage error distribution for 15 PPG beat detectors by biological sex. The box plots show the median, IQR, and 1.5 IQR whiskers. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. The red line denotes the AAMI standard of 10% MAPE. The code base for the PPG beat detectors was from Charlton et al. [124].

PPG Channel	Female		Male	
	Detector	Median AE (BPM)	Detector	Median AE (BPM)
Blue CH 1	**QPPG**	**23.3**	**QPPG**	**19.7**
Blue CH 2	**QPPG**	**23.3**	WFD	20.1
Green CH 1	QPPG	25.4	QPPG	20.5
Green CH 1	QPPG	23.4	MSPTD	20.6
Red CH 1	WFD	30.7	WFD	25.6
Red CH 2	ERMA	30.3	ERMA	26.6
IR CH 1	WFD	29.8	ERMA	26.6
IR CH 2	ERMA	28.8	ERMA	25.2

TABLE 5.5: Biological Sex-Based Performance Analysis of PPG Beat Detectors Heart Rate Estimation Across Various Wavelengths for MW PPG HR dataset (This Work). This table shows the median absolute error (AE) in BPM for different PPG heart rate detectors across biological sexes and PPG wavelengths. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. Each row represents a specific PPG channel, and each column under the activity headings lists the detector with the corresponding median AE. The code base for the PPG beat detectors was from Charlton et al. [124]. **Bold** indicates lowest median absolute error for that biological sex.

In summary, across all wavelengths, males exhibited more accurate PPG measurements than females. QPPG was consistently the most accurate detector for both sexes, especially when using blue light. However, for males, WFD and MSPTD detectors also demonstrated competitive accuracy with blue and green lights, respectively.

5.3.3 Skin Melanin Content and Wavelength

Figure 5.10 depicts the relationship between skin melanin content and PPG beat detector performance. The analysis reveals a trend where higher melanin content coincides with larger IQRs in heart rate measurement accuracy. This suggests greater variability in accuracy for individuals with darker skin tones compared to those with lighter skin tones.

However, some detectors, such as ATM, PWD, and SPAR, exhibit minimal variation across different wavelengths, indicating consistent performance regardless of melanin content. Notably, for most detectors, shorter wavelengths (blue and green) appear to yield more accurate measurements. ERMA, MSPTD, QPPG, WFD, ABD, and AMPD all demonstrate comparable accuracy distributions at shorter wavelengths. It's important to note that none of these detectors meet the AAMI standard for medical device accuracy.

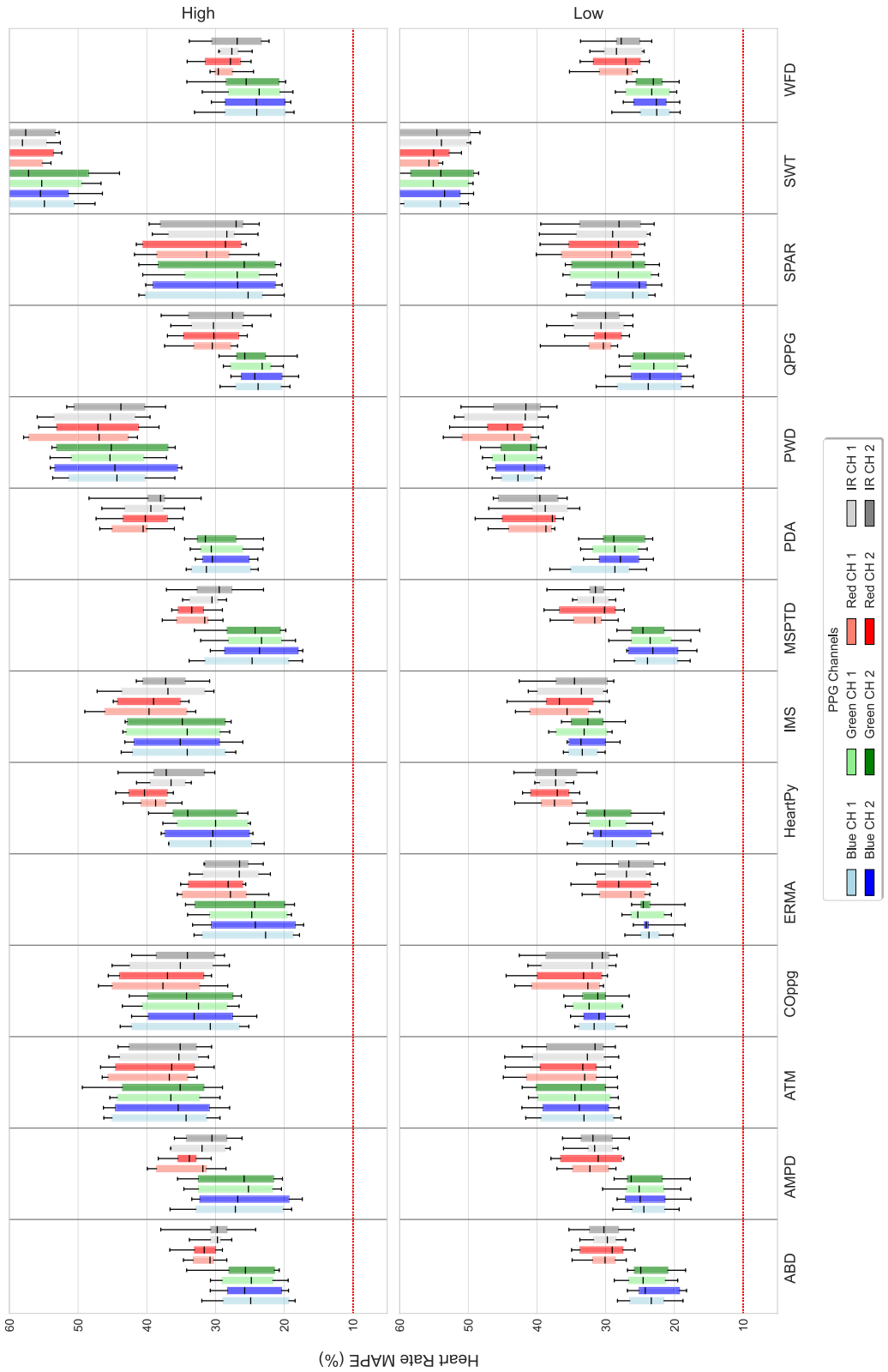


FIGURE 5.10: Performance of PPG Beat Detectors by Skin Melanin Content. This figure presents box-plots of the absolute percentage error distribution for 15 PPG beat detectors by skin melanin content. The box plots show the median, IQR, and 1.5 IQR whiskers. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. The red line denotes the AAMI standard of 10% MAPE. The code base for the PPG beat detectors was from Charlton et al. [124].

Despite this, the consistency observed at shorter wavelengths across various detectors highlights their potential for providing more reliable PPG heart rate estimations.

Table 5.6 presents the median absolute errors (MAE) for PPG heart rate detectors across different skin melanin content levels and PPG wavelengths. The analysis reveals that skin melanin content has a minimal impact on the accuracy of PPG beat detection, with certain trends observed across the various wavelengths.

For individuals with lower melanin content, blue wavelengths showed slightly better performance, with the QPPG detector achieving the lowest MAE of 22.0 BPM using Blue Channel 1. As melanin content increases, the accuracy remains relatively stable, with QPPG still performing well at 23.1 BPM for Blue Channel 2.

PPG Channel	**Low**		**High**	
	Detector	**Median AE (BPM)**	**Detector**	**Median AE (BPM)**
Blue CH 1	**QPPG**	**22.0**	QPPG	23.2
Blue CH 2	QPPG	22.9	**QPPG**	**23.1**
Green CH 1	QPPG	24.0	**QPPG**	**23.1**
Green CH 1	QPPG	23.0	MSPTD	24.9
Red CH 1	ERMA	29.7	ERMA	29.4
Red CH 2	WFD	28.6	ERMA	30.0
IR CH 1	ERMA	27.8	ERMA	28.6
IR CH 2	ERMA	27.4	ERMA	27.7

TABLE 5.6: Skin Melanin Content-based Performance Analysis of PPG Heart Rate Detectors Across Various Wavelengths for MW PPG HR dataset (This Work). This table shows the median absolute error (AE) in BPM for different PPG heart rate detectors across skin melanin content and PPG wavelengths. The analysis includes each PPG wavelength and channel, where wavelength represents the light wavelength of the LED used and channel indicates the photodiode collecting the signal. The data is from the MW PPG HR dataset (This Work). Each row represents a specific PPG channel, and each column under the activity headings lists the detector with the corresponding median AE. The code base for the PPG beat detectors was from Charlton et al. [124]. **Bold** indicates lowest median absolute error for that skin melanin content.

In contrast, green wavelengths exhibited consistent MAE values across different melanin levels. QPPG was the most accurate detector for both low and high melanin content, with MAE values of 24.0 BPM and 23.1 BPM, respectively, using Green Channel 1.

Longer wavelengths, such as red and infrared, also showed minimal variation in accuracy between different melanin content levels. The ERMA detector consistently

performed well across these wavelengths, with MAE values ranging from 27.4 BPM to 30.0 BPM.

Overall, QPPG demonstrated superior accuracy with shorter wavelengths, particularly blue light, while ERMA excelled with longer wavelengths, such as red and infrared. This suggests that while melanin content does not significantly impact PPG beat detection accuracy, the choice of wavelength and detector is key for achieving the best results across different skin types.

5.4 Summary

This chapter addresses Objective 3 of the thesis: *Quality Assessment of PPG Signals*, by developing and comparing various methods for quantifying and assessing the quality of the collected PPG signals across different activities and wavelengths. It provides an in-depth analysis and comparison of the multi-wavelength PPG heart rate monitoring dataset against existing datasets.

The cohort analysis underscores the dataset's unique diversity, featuring balanced representation in biological sex (13 females, 7 males) and encompassing all six Fitzpatrick skin types. The dataset effectively captures a wide range of physiological states, with a substantial number of samples (2,358) exceeding 160 BPM and 9,087 samples within the 80-100% maximal heart rate range.

In the motion analysis, detailed insights from the accelerometer data reveal that while treadmill-based protocols show a correlation between increased speed and accelerometer intensity, wrist-based movements unexpectedly exhibit similar intensities to rest. This finding highlights the limitations of accelerometers in fully capturing the motion types that affect PPG sensing.

The chapter's focus on quality assessment is evident in its evaluation of three SQIs: Elgendi SNR, a proposed ECG-derived SNR, and the correlation between accelerometer intensity and PPG signal. The ECG-derived SNR emerges as the most reliable SQI for this use case, displaying consistent and expected trends, such as a stronger negative correlation (-0.41) with true heart rate and distinct distributions for each activity. This makes it the most suitable SQI for assessing the robustness of proposed heart rate estimation methodology.

In the final section, the chapter evaluates PPG beat detectors across activities, wavelengths, and demographics. QPPG and ERMA perform best, especially during rest with blue light, achieving errors as low as 11.5 BPM. Detector accuracy varies by activity, with WFD excelling in low-intensity activities and QPPG in more intense ones.

Differences between blue and green wavelengths are minimal, often within 1 BPM. Demographic factors influence PPG detector performance to varying degrees. Males generally exhibit higher accuracy across all wavelengths, with QPPG showing a 3.6 BPM advantage for males over females when using blue light. Skin melanin content has a minimal impact overall, but detectors like ERMA perform consistently well with longer wavelengths, such as red and infrared, across different melanin levels.

Overall, this chapter provides insights that inform the development and evaluation of wrist-worn PPG heart rate monitoring methodologies, particularly in addressing motion artefacts, signal quality, and demographic variability.

Chapter 6

A Convolutional Neural Network for Heart Rate Estimation

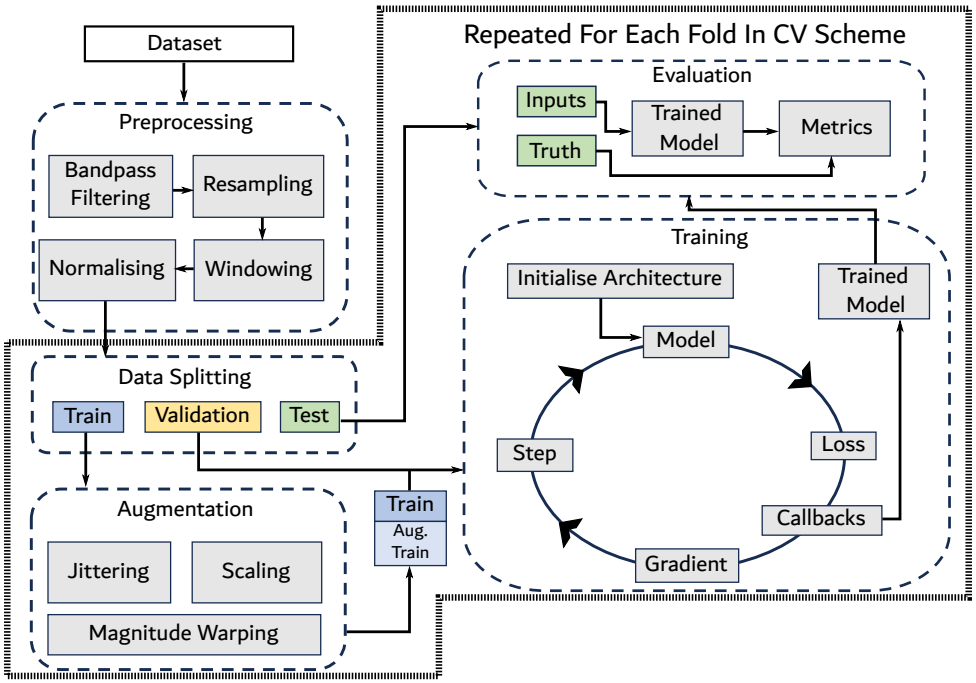


FIGURE 6.1: Overview of the Development Process of the Proposed CNN Heart Rate Estimation Method. This figure illustrates the general methodology for supervised deep learning, detailing the specific steps chosen for this study. The process includes pre-processing, data augmentation, data splitting, model initialisation, model training, model evaluation and cross-validation [163].

Following the detailed analysis of the multi-wavelength wrist-worn photoplethysmography (PPG) heart rate estimation dataset in the previous chapter, this chapter aims to validate a convolutional neural network (CNN) method for heart rate estimation using the collected data. This process, illustrated in Figure 6.1, encompasses several key steps: pre-processing and augmentation, data splitting and CV, architectural design and

training, and thorough evaluation. The chapter provides detailed justifications for each decision in the method's creation.

Furthermore, an in-depth performance analysis is conducted, including assessing the impact of wavelength selection, demographic variations, and application of the method on existing single-wavelength datasets. The chapter concludes with a comparative analysis, evaluating the proposed deep learning approach against conventional statistical methods. The detailed versions of the software used are outlined in Section 3.5

6.1 Signal Pre-processing and Augmentation

6.1.1 Signal Pre-processing

As detailed in Section 2.2, both conventional and deep learning PPG heart rate estimation methods employ a pre-processing stage to prepare and enhance the signal for the estimation methodology. This stage typically includes techniques such as filtering, re-sampling, windowing, transformation and normalisation. Figure 6.2 depicts an unprocessed PPG signal, showcasing the raw data with its inherent challenges for accurate heart rate estimation. The signal exhibits baseline wander and motion artefact spikes, both common issues in PPG analysis. However, an 8-second segment reveals discernible PPG waveforms amidst the noise and interference. This emphasises the necessity of the detailed pre-processing steps outlined in this section for extracting accurate heart rate estimation.

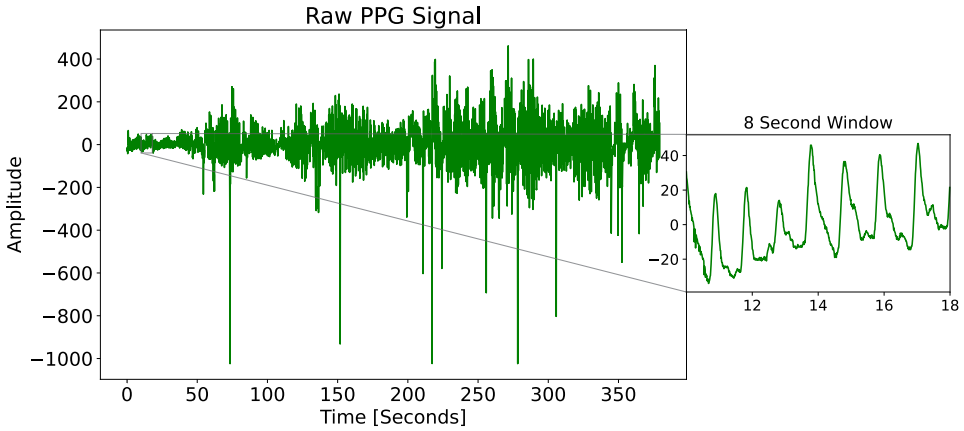


FIGURE 6.2: Raw PPG signal displaying the full signal alongside an 8-second zoomed-in window. The zoomed-in view highlights the characteristic PPG waveform, with evident noise present in both the full signal and the window.

The initial step in the pre-processing stage involves the application of a 4^{th}-order Butterworth bandpass filter with a pass-band of 0.5-4 Hz, corresponding to the typical

heart rate band of 30-240 BPM. The Butterworth filter is selected for its maximally flat frequency response in the pass-band, ensuring minimal amplitude distortion within the range of interest. As evident in Figure 6.3, this filtering attenuates baseline wander, significantly enhancing signal clarity. However, the figure also reveals residual motion artefact spikes, likely caused by overlapping frequency components that remain unaddressed by the bandpass filter alone.

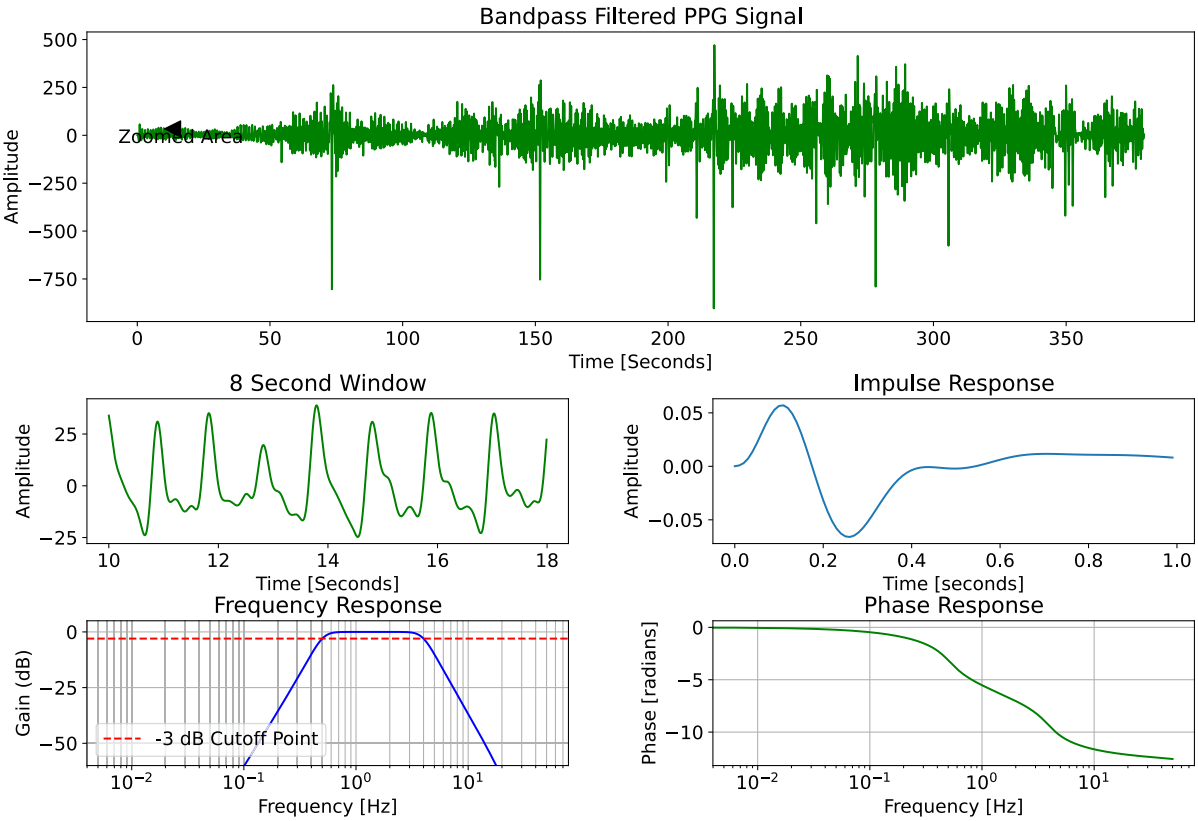


FIGURE 6.3: Band-Pass Filtered PPG Signal and Filter Responses. The figure displays the frequency, phase, and impulse responses of a 4th-order Butterworth band-pass filter with a pass-band of 0.5-4 Hz, corresponding to the typical heart rate range of 30-240 BPM. It also shows an 8-second window of the PPG signal where the effects of filtering are evident. The Butterworth filter is chosen for its maximally flat frequency response in the pass-band, ensuring minimal amplitude distortion within the range of interest.

Before applying the bandpass filter, a 0.05% Tukey window is applied to the signal. The Tukey window serves as a tapering function that helps reduce spectral leakage by smoothly transitioning the signal to an amplitude of zero at the edges, as shown in Figure 6.4. This step is key in minimising the introduction of artefacts during the subsequent filtering process.

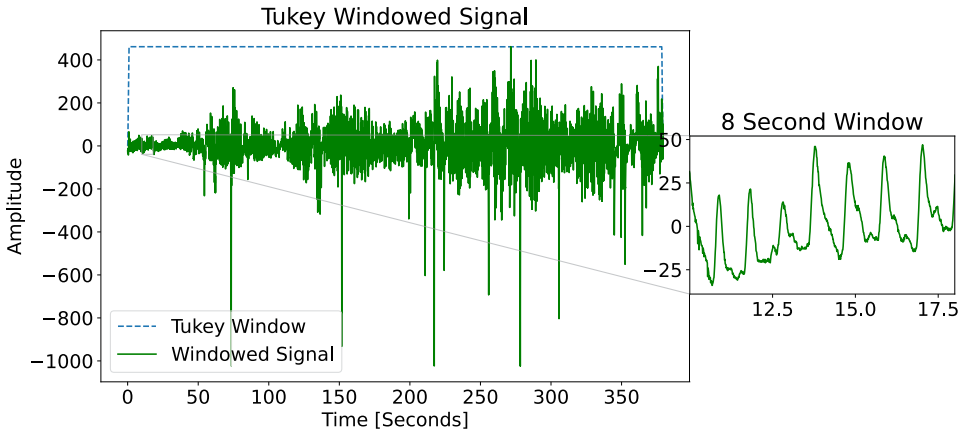


FIGURE 6.4: Tukey Windowed PPG Signal. The figure demonstrates the application of a Tukey window to the PPG signal, which tapers the signal's edges to reduce spectral leakage.

Following filtering, the signal undergoes re-sampling to a standardised rate of 64 Hz for consistency across datasets. A poly-phase filter bank achieves this re-sampling efficiently by combining up-sampling and decimation. This process involves raising the sample rate, applying a low-pass filter to eliminate unwanted high-frequency components, and then reducing the sampling rate through decimation. This method optimises computational efficiency and suppresses aliasing, ensuring the preservation of the core PPG signal information. Next, the resampled signal is segmented into overlapping windows of 8 seconds with a 2-second overlap. This windowing strategy aligns with established practices in PPG heart rate estimation literature, facilitating result comparability across studies.

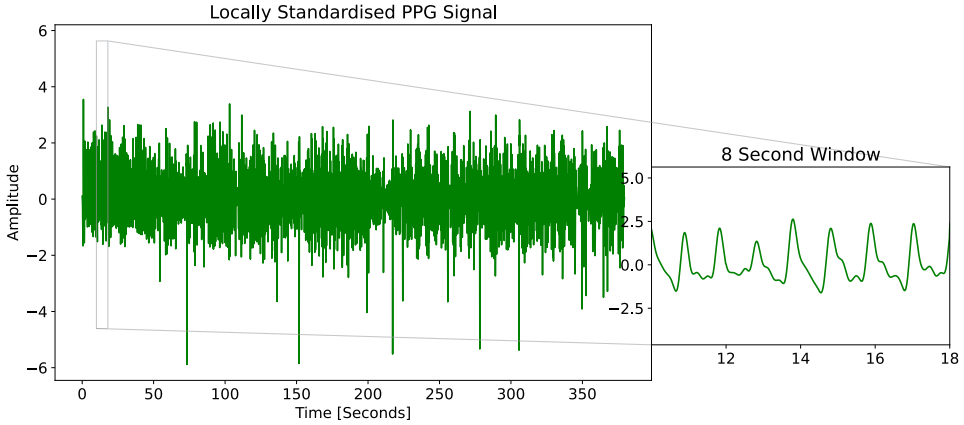


FIGURE 6.5: Z-Normalised PPG Signal. The figure illustrates the z-normalisation process applied to the PPG signal, which standardises the signal by adjusting for mean and variance. This process transforms the data to have a mean of zero and a standard deviation of one, as demonstrated by the normalised PPG waveforms.

Finally, z-normalisation is applied to each individual windowed segment, as shown

in Figure 6.5. This step transforms the data within each window to have a mean of 0 and a standard deviation of 1. By performing z-normalisation on a window-by-window basis rather than on the entire pre-processed signal, the model tailors the normalisation to the specific characteristics of the analysed segment. This approach is particularly advantageous for PPG signals due to their non-stationary nature, where physiological conditions can vary over time. Normalising the entire signal beforehand could potentially mask these variations, hindering the model's ability to detect subtle changes within each window

6.1.2 Signal Augmentation

Data augmentation incorporates prior knowledge of data in-variance under specific transformations. By artificially expanding the training data with these transformations, data augmentation increases the diversity of the input space and discourages overfitting. This, in turn, enhances the model's ability to generalise to unseen data. The effectiveness of this approach has been demonstrated in methods for estimating heart rate from PPG signals [27] and monitoring systems for Parkinson's Disease [218].

As an augmentation technique, jittering adds small random amplitude fluctuations to the PPG signal, as shown in Figure 6.6. This is intended to increase the model's exposure to varied signal patterns, aiding it in learning to process signals with inherent inconsistencies and ultimately enhancing its accuracy in heart rate estimation across diverse conditions.

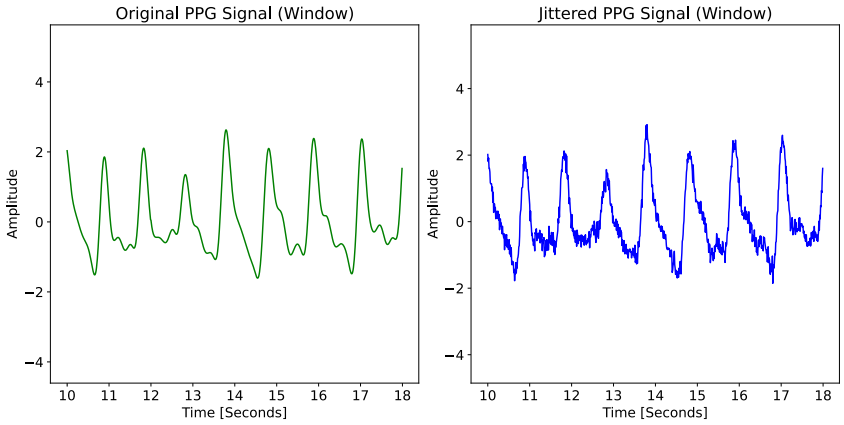


FIGURE 6.6: Signal Jittering of PPG Window. The figure shows the effect of signal jittering on the PPG data, which introduces high-frequency noise into the signal.

The scaling augmentation in the training data adjusts the PPG waveform's amplitude and diversifies the training dataset, as shown in Figure 6.7. By introducing a variety of amplitude profiles, the model is trained to handle a broader spectrum of signal

strengths, making it more versatile and robust in different operational contexts. By training the model on signals with diverse amplitude profiles, the method becomes more robust to the differences in signal magnitude that can occur across different users or sensor positions.

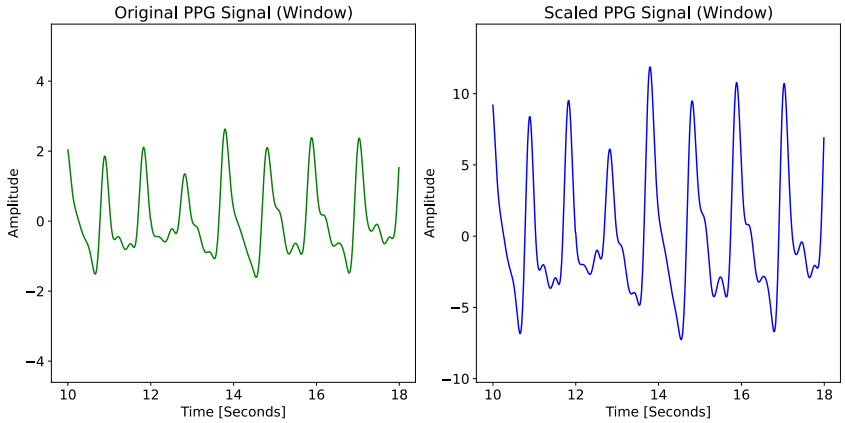


FIGURE 6.7: Signal Scaling of PPG Window. The figure illustrates signal scaling applied to the PPG data, which adjusts the amplitude uniformly across the entire signal.

Magnitude warping is applied to the PPG signals to replicate the non-linear morphological changes that can be introduced by shifts in physiological states or movement, as shown in Figure 6.8. This technique warps the signal in a controlled manner, creating realistic scenarios where the waveform is distorted, as it often happens in practical applications due to motion or pressure changes on the sensor.

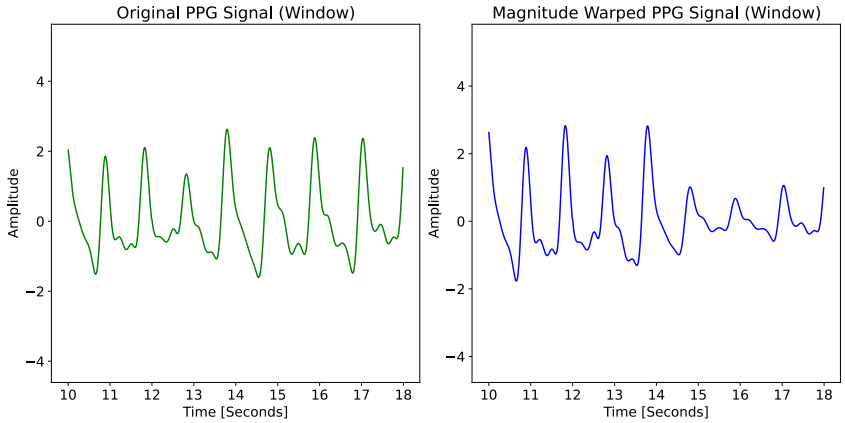


FIGURE 6.8: Magnitude Warping of PPG Window. The figure demonstrates magnitude warping applied to the PPG signal, which alters the amplitude differently across various sections. This effect is illustrated by the last three PPG waveforms.

These augmentation techniques enrich the training dataset with a broad spectrum of variations. These techniques, including jittering, scaling, and magnitude warping, are

evenly distributed in the augmented dataset. Random data windows undergo one of these augmentations, applied uniformly across all signals in a sample, encompassing all PPG channels, wavelengths, and accelerometer data. The specific augmentation parameters were set at 0.15 for jittering, 2.0 for scaling, and 0.5 for magnitude warping, enhancing the dataset's robustness and diversity.

6.2 Architecture

The architecture of a neural network essentially defines the space of possible solutions that the training process can explore via gradient descent. This space is influenced by the model's parameters. A well-designed network architecture incorporates prior knowledge about the data to guide the exploration towards optimal solutions that effectively capture the underlying relationships within the data [163]. In this study, certain assumptions are made about the nature of the data, which comprises one-dimensional time series data owing to its temporal nature and is characterised as multivariate due to its multi-wavelength and multi-channel composition.

Motion artefacts pose a significant challenge within wrist-worn PPG signal processing, necessitating a substantial portion of the network's computational efforts to discern signal from noise. A sensor fusion strategy incorporating motion references is utilised to aid with this. The network's primary objective is to output a continuous value representing the HR, framing the task as a regression task. Key features of PPG signals for heart rate estimation can be broadly categorised into two types: local and global. Local features focus on specific characteristics within a signal window, such as the presence and location of systolic peaks. Global features, on the other hand, capture overall signal properties within the window, such as the number of systolic peaks detected. The architectural framework is outlined in the following section based on these assumptions.

6.2.1 Sensor Fusion

Sensor fusion combines several sensing modalities to gain richer information about the individual parts and their emergent behaviour [219]. Classical approaches to the fusion of heterogeneous sensing modalities rely on feature engineering to extract independent features from each sensing modality, which are fused. This approach of extracting different features from individual sensors disregards features that use multiple sensors' data to capture information that neither has in isolation [220]. In many applications, DNNs have been adopted instead due to their ability to learn features during training [220–222], showing improved performance in applications such as gait

recognition [220], human activity recognition [220–222], car tracking [221], dynamic gas mixtures estimations and cuffless blood pressure monitoring [222].

For wrist-worn PPG sensing, motion artefacts have overlapping frequency bands with the cardiac signal, making the removal of such artefacts challenging. Motion references such as accelerometers, gyroscopes, and longer wavelength PPG have been commonly used in effective motion artefact reduction methods [42, 72–74, 128]. There are several strategies for data fusion, namely early, intermediate and late fusion [219]. In this research, channel-wise fusion occurs early in the network for each sensing modality. Then, intermediate fusion combines latent features from each sensing modality, enhancing the overall informative representation of heterogeneous sensory data.

6.2.2 One-dimensional Convolutions

In deep learning, convolutional operations excel at extracting local features from non-linear and multi-dimensional data. For bio-signals like PPG signals, which have a single temporal dimension, 1D convolutions are employed. These convolutions involve kernels that slide along the input sequence, generating an output where each value represents a weighted sum of its neighbouring input values. When dealing with multi-channel data (e.g., incorporating additional sensor data), each channel is processed independently, and the resulting activations are summed to produce a single output value.

The number of filters within a convolutional layer dictates the variety and complexity of features the network can learn. Each filter acts as a feature detector, emphasising specific patterns or aspects within the input data. Zero-padding is a common technique used to preserve the input dimensionality after the convolution operation.

The convolution operation is typically followed by adding a bias term b and applying an activation function σ. The mathematical expression for the forward pass of a 1D convolution layer is given by:

$$Y[i] = \sigma(\sum_{j=0}^{k-1} X[i+j] \cdot W[j] + b) \tag{6.1}$$

Here, $Y[i]$ is the output at position i in the output sequence, $X[i+j]$ is the input value at position $i+j$, $W[j]$ is the weight at position j, b is the bias term, and k is the size of the filter. The activation function, σ, introduces non-linearity, allowing it to learn intricate patterns. Common activation functions include Rectified Linear Unit (ReLU), Exponential Linear Unit (ELU), and Leaky Rectified Linear Unit (Leaky ReLU). Depending on the chosen activation function, distinct weight initialisation schemes

are recommended. For instance, ReLU activation favours He Initialisation, while ELU activation aligns well with LeCun initialisation. These initialisation strategies contribute to the stable and efficient training of convolutional neural networks by providing suitable starting points for optimisation.

Convolutional layers are often accompanied by pooling operations to achieve two key goals: dimensionality reduction and enhanced translation in-variance. Pooling works by summarising information within localised regions of the data. When applied sequentially, pooling performs a hierarchical aggregation of information. This allows the network to capture increasingly abstract representations of the input data. Essentially, pooling enables the network to recognise patterns and features at different scales, ranging from local details to global characteristics. In the context of PPG-based heart rate estimation, this hierarchical processing is key. It facilitates the transition from capturing localised features like individual systolic peaks to broader features like the total number of peaks within a signal segment.

The proposed architecture primarily relies on convolutional pooling layers for feature extraction, playing a vital role in processing the signals for accurate heart rate estimation. However, the effectiveness of convolutional and pooling layers is not isolated. Their success hinges on the integration of complementary techniques, such as normalisation, which standardises the input data to ensure a consistent range for efficient processing, and regularisation strategies that help prevent overfitting—a phenomenon where the model memorises training data specifics and performs poorly on unseen data. Additionally, a continuous output mechanism is key, as the final layer of the network needs to generate a continuous numerical value that accurately represents the heart rate.

6.2.3 Normalisation

Batch normalisation is pivotal for stabilising and improving the efficiency of neural network training by normalising the input of each layer during the training process. A "batch" refers to a subset of the training data processed through the neural network simultaneously. During training, the neural network does not process the entire dataset in one go but instead works on these smaller batches of data. Each batch goes through the normalisation process separately.

The input to each layer is normalised by subtracting the mean and dividing by the standard deviation, introducing a small constant (ϵ) to prevent division by zero. These normalised inputs are scaled and shifted using learnable parameters, γ and β. This affords the model the adaptability needed to fine-tune these inputs according to the learning dynamics within each batch [223].

6.2.4 Regularisation

Deep learning models are susceptible to overfitting, a phenomenon where the model learns not only the underlying patterns in the training data but also the noise and specific characteristics of that data. This leads to poor performance on unseen data. Regularisation techniques play a key role in mitigating overfitting and enhancing the model's generalisability.

One such technique is dropout. During training, dropout randomly deactivates a subset of neurons within the network. This forces the model to learn robust features that are not overly reliant on any specific neuron or group of neurons. By introducing beneficial noise, dropout discourages the model from simply memorising the training data and instead compels it to learn generalisable representations that perform well on unseen data [224].

In addition to dropout, weight regularisation techniques are employed to further prevent overfitting. These techniques penalise the model for having overly complex weights, encouraging simpler and more generalisable models. Two common weight regularisers are L1 (Lasso) and L2 (Ridge). L1 regularisation promotes sparsity by driving some weights towards zero, effectively removing them from the model. This can be particularly useful when dealing with high-dimensional data where many features might be irrelevant [225].

In contrast, L2 regularisation penalises large weights, even if they remain non-zero. This helps to improve the stability of the model and prevent it from becoming overly reliant on specific features in the training data [225]. Elastic Net combines both L1 and L2 regularisation, offering a balance between sparsity and weight value control [225]. This study utilises both dropout and Elastic Net regularisation to provide a robust approach to overfitting prevention, ultimately enhancing the generalisability of the neural network.

6.2.5 Global Pooling and Output

In a regression framework for heart rate estimation, the final layer uses a linear activation function, suitable for predicting the positive and continuous nature of heart rates. Global pooling is applied after the final convolutional layer to reduce dimensionality by summarising each feature map into a single value. This method establishes direct, weighted connections between the output neuron and each feature map, offering computational efficiency over flattening the temporal dimension, which would increase model complexity.

Global pooling also directs the model's focus towards the most critical features within each map, enhancing robustness and generalisability by mitigating overfitting. This approach simplifies the model while prioritising essential features, leading to a more effective heart rate estimation model. The final architecture, arrived at after hyperparameter optimisation (see section 6.4) is shown in Figure 6.9.

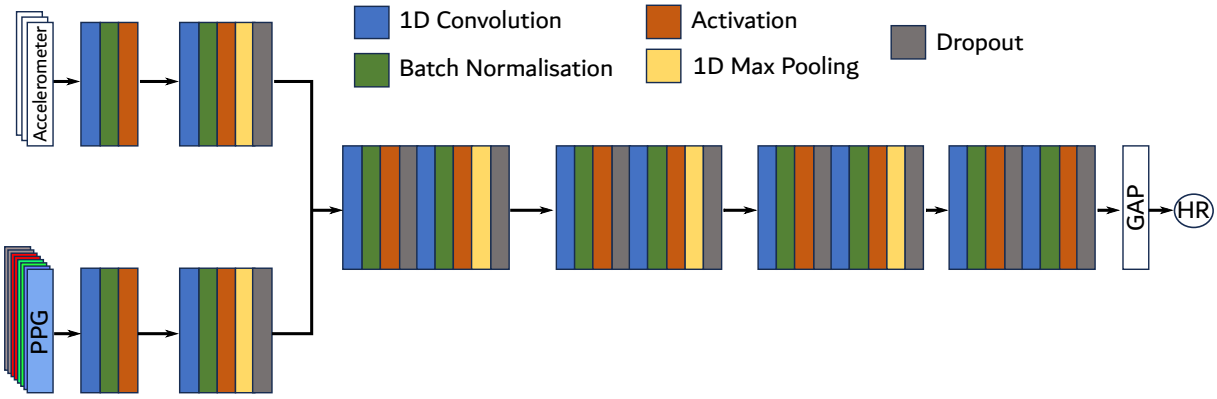


FIGURE 6.9: Schematic Representation of the Proposed Convolutional Neural Network Architecture for PPG Heart Rate Estimation. The architecture incorporates early channel-wise fusion for each sensing modality, followed by intermediate fusion of latent features from each sensing modality. It employs one-dimensional convolutional layers with zero-padding for feature extraction, batch normalisation for input stabilisation, and dropout for regularisation. The network uses global pooling to reduce dimensionality and focus on critical features. The final layer utilises a linear activation function for continuous heart rate output. This architecture, resulting from hyperparameter optimisation, contains 730,000 parameters and consumes 2.78 megabytes.

6.3 Training and Validation

6.3.1 Loss Functions

A loss function serves to quantify the distance between the current prediction generated by a neural network and the expected output. It functions as a mechanism for assessing how effectively the network captures the underlying patterns in the data. Furthermore, the loss function is pivotal in refining the network's parameters during the training process.

Common loss functions in cases with continuous predictions and ground truth include Mean Squared Error (MSE), MAE, and Huber Loss, which blends MSE and MAE characteristics. For this chapter, MSE is selected as the loss function due to its inherent capacity to penalise substantial errors with a quadratic progression. MSE computes

the average squared difference between the truth value and the predicted value and is calculated using the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6.2}$$

where y_i is the truth value for the i^{th} sample in the dataset and $\hat{y}_i$ is the predicted value for the i^{th} sample.

6.3.2 Backpropagation and Optimiser

Adjusting network parameters to minimise the loss function relies on two pivotal algorithms: backpropagation and an optimiser. Backpropagation is an iterative procedure characterised by a forward pass, during which input data traverses the network to generate predictions. Subsequently, a backward pass computes the gradients of the loss function with respect to the network parameters [228]. These gradients provide essential information for the optimiser, which utilises them to adjust the network parameters and minimise the loss function. Batch processing is typically employed for several reasons, including the stochastic nature that aids in escaping local minima and ultimately enhances model generalisability.

Leveraging the gradients computed by backpropagation, the optimiser iteratively adjusts the network parameters to minimise the loss function. This iterative refinement process, repeated multiple times throughout training, progressively improves the model's ability to generate accurate predictions. Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (ADAM) are common choices for optimisers in this context. SGD is a foundational optimisation algorithm with simplicity and efficiency, making it a widely used approach [229]. The SGD update rule is expressed as:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) \tag{6.3}$$

Here, θ_t represents the current parameters, η is the learning rate, and $\nabla L(\theta_t)$ signifies the gradient of the loss function with respect to the parameters. However, SGD is sensitive to the choice of the learning rate, potentially leading to slow convergence or oscillations around the optimum.

The ADAM optimiser extends the principles of SGD by adapting the learning rate for each weight of the neural network using estimations of the first and second moments of the gradient. This adaptability facilitates convergence in scenarios with varying

gradient magnitudes [229]. The ADAM update rule is given by:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\nu_t} + \epsilon} \cdot m_t \tag{6.4}$$

Here, η is the learning rate, m_t is the first-moment estimate, ν_t is the second-moment estimate, and ϵ is a small constant. Despite its popularity, ADAM may exhibit sensitivity to hyperparameter choices, and its inherent complexity may result in increased computational requirements.

NADAM, a hybrid of Nesterov accelerated gradient and ADAM, seeks to harness the advantages of both adaptive learning rates and momentum [229]. The NADAM update rule is a modification of ADAM, encompassing Nesterov momentum:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\nu_t} + \epsilon} \cdot (m_t + \beta \cdot m_{t-1}) \tag{6.5}$$

In this expression, β is a decay factor for the past gradient. While NADAM often demonstrates accelerated convergence, its performance can be contingent on the specific characteristics of the optimisation landscape.

Two hyperparameters are key in the optimiser: learning rate and weight decay. The learning rate, denoted as η, governs the magnitude of the network's parameter updates in each iteration. An inappropriate learning rate can hinder convergence or overshoot the optimal solution. Weight decay, a regularisation technique, adds a penalty based on the magnitude of model parameters. This discourages overly complex models and contributes to improved generalisation. The weight decay term is typically expressed as $\lambda||\theta||^2$, where λ is the weight decay coefficient and $||\theta||^2$ is the L2 norm of the model parameters.

6.3.3 Training Parameters and Callbacks

The performance of a neural network is not only influenced by the choice of optimisers but also by various training parameters. Training is carried out in batches, introducing stochasticity to the optimisation process. Each batch provides a different subset of the data, leading to variations in the gradients computed during each iteration. This stochasticity can help the optimiser escape local minima and explore a broader region of the optimisation landscape, potentially leading to improved convergence and better generalisation to unseen data. This research sets the batch size to 64 to balance the training time and sensitivity to individual inputs.

An epoch is complete when all batches undergo one full training iteration, and the number of epochs is a key training parameter. Sufficient epochs ensure the model learns the underlying pattern of the data without under-fitting, yet excessive epochs can lead to overfitting on training data, resulting in poor generalisation. to ensure sufficient training time. However, this upper limit is unlikely to be reached because the training process is controlled by callbacks. Two training callbacks are employed to counter overfitting. Model check-pointing saves weights from the lowest validation loss, a key performance indicator on unseen data. Early stopping halts training if validation loss does not improve over 25 epochs, assuming convergence and avoiding unnecessary computational costs.

As discussed in Section 6.3.2, the learning rate is a key hyperparameter initially set at 0.001. Stagnation in 'learning' can occur, indicative of overshooting the global minima due to large steps in the loss landscape. A 'reduce learning rate on plateau' training callback is implemented to address this. The learning rate is reduced if validation loss does not improve over five consecutive epochs. This adjustment reduces the magnitude of the step in network parameter updates, aiding the search for the global minima and promoting convergence.

6.3.4 Data Splitting and Cross Validation

During the training of a neural network, it is conventional to partition the dataset into distinct subsets, namely the train, validation, and test sets. The train set is employed for iteratively updating the network parameters through optimisation. Following each training epoch, the validation set evaluates the network's performance on data it has not been explicitly trained on. Although the validation set does not contribute to the parameter updates, it plays a pivotal role in assessing the training process. A separate test set gauges the trained network's performance on unseen data.

How data is divided can vary based on the chosen CV scheme, with common approaches involving random splitting or by subject. In this process, subsets are iteratively drawn from the dataset, and each subset is subsequently utilised for training, validation, and testing of the model. This approach allows for evaluating test error across multiple iterations, providing a more robust assessment of model performance.

In deep learning wrist-worn PPG heart rate estimation methods, two commonly used cross-validation (CV) schemes are k-fold and Leave-One-Subject-Out (LOSO), as shown in Figure 6.10. However, as discussed in Section 2.2.4, the practical application of these methods requires generalisability to new individuals not included in the training set.

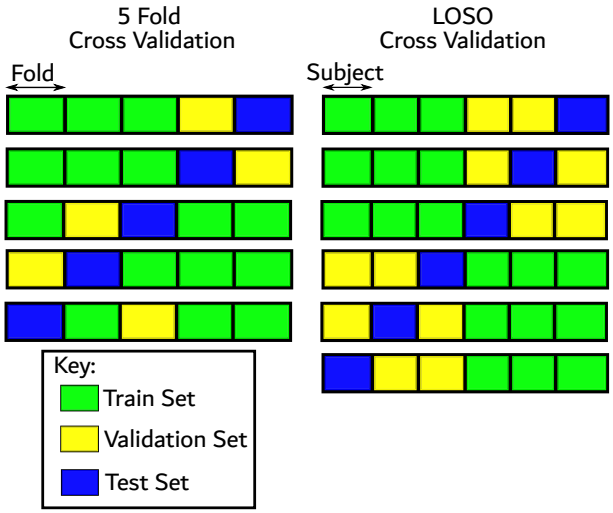


FIGURE 6.10: Comparison of Cross Validation Schemes.

K-fold CV, which assesses generalizability on a sample basis, may fall short in this regard, especially when data from all subjects are included in every data split.

Given that bio-signals can vary significantly between subjects and even across different sessions for the same individual—due to factors like sensor placement and environmental conditions—it is key to evaluate models on a subject basis. This makes LOSO CV a more appropriate method for assessing generalisability in real-world scenarios, ensuring that the model performs well across different individuals.

CV Scheme	IEEE Train [126]	IEEE Test [126]	BAMI 1 [127]	BAMI 2 [127]	PPG DaLiA [130]	MW PPG HR (This Work)
LOSO	5.8 ± 7.4	15.0 ± 11.8	3.6 ± 2.1	1.6 ± 0.6	4.9 ± 3.1	7.9 ± 2.6
5-Fold	**1.4 ± 0.7**	**2.8 ± 1.2**	**1.9 ± 0.5**	**1.2 ± 0.3**	**2.0 ± 0.4**	**2.1 ± 0.6**
p-value	0.004	0.001	<.00001	0.008	0.004	<.00001

All Values are MAE in BPM. Statistical Tests used the Mann-Whitney U test.

TABLE 6.1: Performance Comparison of Cross-Validation Schemes Across All Utilised Datasets. The table examines the effect of different cross-validation schemes on accuracy. While 5-fold cross-validation demonstrates higher accuracy, it is not applicable in real-world scenarios, which is why Leave-One-Subject-Out (LOSO) validation is utilised. Statistical testing reveals significant differences between the schemes for all datasets. **Bold** values indicate the lowest MAE distribution.

A comparative analysis was undertaken to validate the impact of distinct CV schemes on generalisation error. Table 6.1 outlines notable and statistically significant differences in MAE observed across all datasets. The conventional five-fold CV scheme yielded optimistic generalisation errors, registering 2.8 ± 1.2 BPM on the IEEE Test dataset, characterised by a predominantly arm movement-based protocol. In contrast, adopting a more practical approach with LOSO CV revealed more realistic generalisation errors, with a performance of 15.0 ± 11.8 BPM on the same IEEE Test dataset. This discrepancy

underscores the necessity for CV methods that consider both the substantial variability inherent in bio-signals across different subjects and the intended use case of the method.

6.4 Hyperparameter Optimisation and Ablation Study

6.4.1 Hyperparameter Optimisation

Each component in the network has configurable parameters that affect performance, including learning rate, dropout rate, and convolutional layer kernel size—referred to as hyperparameters. These hyperparameters interact in complex ways, making it ineffective to assess the impact of a single hyperparameter in isolation. The set of all potential hyperparameter combinations constitutes the search space, and hyperparameter optimisation algorithms navigate this space to identify the configuration yielding the smallest error.

Grid search exhaustively traverses the search space to find the optimal set of hyperparameters. While practical for small search spaces, it becomes computationally demanding as the space expands [226]. In this network, the search space encompasses nearly a million potential hyperparameter configurations, rendering grid search impractical.

Bayesian optimisation is preferred over grid search due to its sequential model-based approach, which efficiently identifies the global optimum with fewer trials. Bayesian optimisation balances exploration and exploitation to avoid local optima. A Bayesian probability surrogate model characterises the objective function, and an acquisition function determines the next sampling point [226].

This study uses a Gaussian process (GP) as a probability surrogate model of objectives. The GP models the method's performance, and the Lower Confidence Bound is selected as the acquisition function to guide the selection of the next hyperparameter configuration [227]. The optimisation was performed for 50 iterations, each encompassing a complete LOSO CV evaluation, with the test error serving as the optimisation metric. The optimisation used the IEEE Train dataset [126] due to its size. Table 6.2 displays the selected optimal hyperparameters. The optimised configuration results in an architecture with 730,000 parameters, consuming 2.78 megabytes, as shown in Figure 6.9.

6.4.2 Ablation Study

Following hyperparameter optimisation, an ablation study was conducted to examine the individual effects of various architectural components. This section reports on the

Hyperparameter	**Search Space**	**Selected Optimal**
Convolutional Kernel Size	[8,16,32]	16
Number of Convolutional Filters	[16,32,64]	64
Sensor Layers	[1,2,3]	2
Fusion Layers	[3,4,5]	4
Activation	[ReLU, Leaky ReLU, ELU]	Leaky ReLU
Global Pooling	[Max, Average]	Average
Elastic Net (L1)	1e-5 - 1e-2	1e-5
Elastic Net (L2)	1e-5 - 1e-2	1e-4
Dropout Rate	0.1 - 0.4	0.1
Weight Decay	1e-2 - 0.1	0.01
Learning Rate	1e-4 - 0.01	1e-3
Optimiser	[NADAM, ADAM, SGD]	SGD
Batch Size	[32,64,128]	64

TABLE 6.2: Overview of the Hyperparameter Optimisation Search Space.

impact of batch normalisation and optimiser choice.

Effect of Batch Normalisation

Batch Norm	**IEEE Train [126]**	**p-value**	**IEEE Test [126]**	**p-value**
Included	**5.8 ± 7.4**	0.010	**15.0 ± 11.8**	0.032
Excluded	9.9 ± 8.6		21.2 ± 10.5	

All Values are MAE in BPM. Statistical Tests used the Mann-Whitney U test.

TABLE 6.3: Effect of Including Batch Normalisation in the Architecture on Heart Rate Estimation Performance for IEEE Train and IEEE Test Datasets [126]. **Bold** values indicate the lowest MAE distribution. Ablation experiment carried out after hyperparameter optimisation.

Analysis of batch normalisation's impact on network performance, as shown in Table 6.3, reveals a statistically significant decrease in MAE values for both IEEE Train and Test datasets [126] when including batch normalisation in the network architecture. This demonstrates batch normalisation's important role in stabilising gradient flow, reducing internal covariate shift, and improving overall network performance.

Effect of Optimiser Choice

The hyperparameter optimisation process included the choice of optimiser in the search space, with SGD emerging as the optimal selection. Further analysis of optimiser performance reveals additional insights. As shown in Table 6.4, SGD achieved MAE values approximately 2 BPM lower than alternative optimisers for both IEEE Train and Test datasets [126]. Momentum-based optimisers accumulate updates by incorporating previous gradients, which can smooth convergence but may also introduce inaccuracies

Optimiser	IEEE Train [126]	p-value	IEEE Test [126]	p-value
SGD	**5.8 ± 7.4**	—	**15.0 ± 11.8**	—
ADAM	7.7 ± 6.7	0.100	16.5 ± 9.2	0.192
NADAM	7.7 ± 8.5	0.051	16.4 ± 9.6	0.174

All Values are MAE in BPM. Statistical Tests used the Mann-Whitney U test.

TABLE 6.4: Effect of Optimiser Choice in Model Training on Heart Rate Estimation Performance for IEEE Train and IEEE Test Datasets [126]. Statistical tests compare each optimiser individually to SGD. **Bold** values indicate the lowest MAE distribution. Ablation experiment carried out after hyperparameter optimisation.

from less accurate gradients. However, in our analysis, the impact of these inaccuracies was statistically insignificant, indicating minimal effect on overall performance.

6.5 Heart Rate Estimation Performance

6.5.1 Comparison of Wavelength Selection

This section addresses the primary research question 1: To what extent does the robustness and generalisability of wrist-worn PPG heart rate estimations vary across different wavelengths or combinations of wavelengths, compared to the green light conventionally used in consumer wrist-worn devices?

Analysis of the individual wavelengths indicates that longer wavelengths are associated with higher absolute errors. Specifically, the use of red and IR wavelengths resulted in MAE that were 7.7 BPM and 7.2 BPM higher, respectively, compared to the conventional green light. These differences are statistically significant, with p-values of $\leq 1e-5$. In contrast, blue light exhibited an MAE only 0.1 BPM greater than green light, and this difference was not statistically significant (p = 0.19) (see Figure 6.11 and Table 6.5).

When examining multi-wavelength combinations relative to green light, statistically significant differences were noted for the Blue-Green, Blue-Green-Red, and Blue-Green-Red-IR combinations. Of these, only the Blue-Green-Red-IR combination showed a notable improvement in MAE, reducing it by 0.4 BPM compared to green light (see Table 6.5).

During periods of active rest, which include both aperiodic and periodic motion types, the blue wavelength demonstrated a higher MAE compared to green light, with a difference of 0.2 BPM. However, only the Blue-Green-Red-IR combination showed an improvement, with a 0.2 BPM lower MAE and a statistically significant difference in error distribution ($p \leq 0.01$) (see Figure 6.12 and Table 6.5).

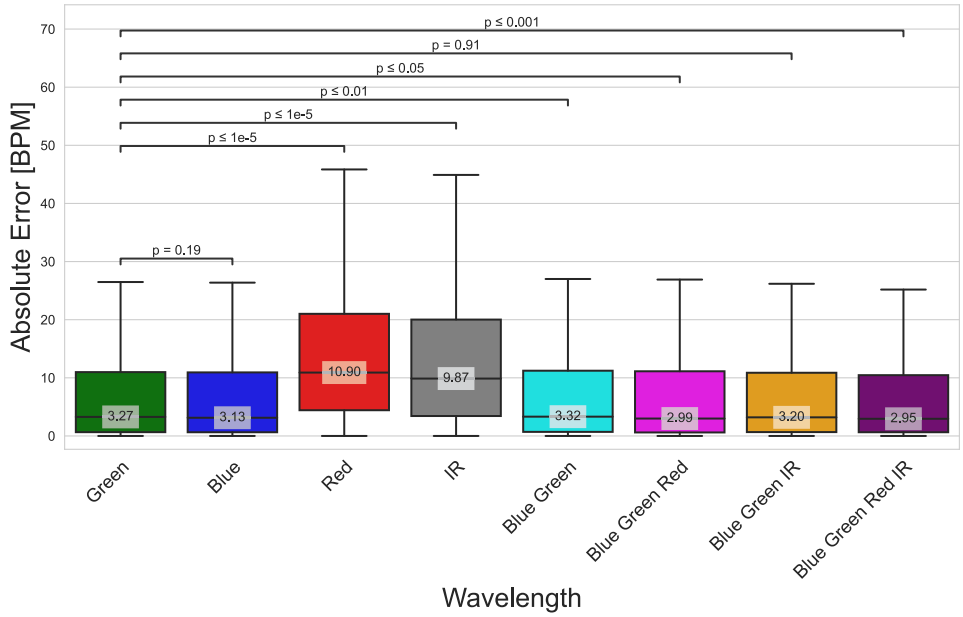


FIGURE 6.11: Comparison of Distributions of Absolute Error by Wavelength for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

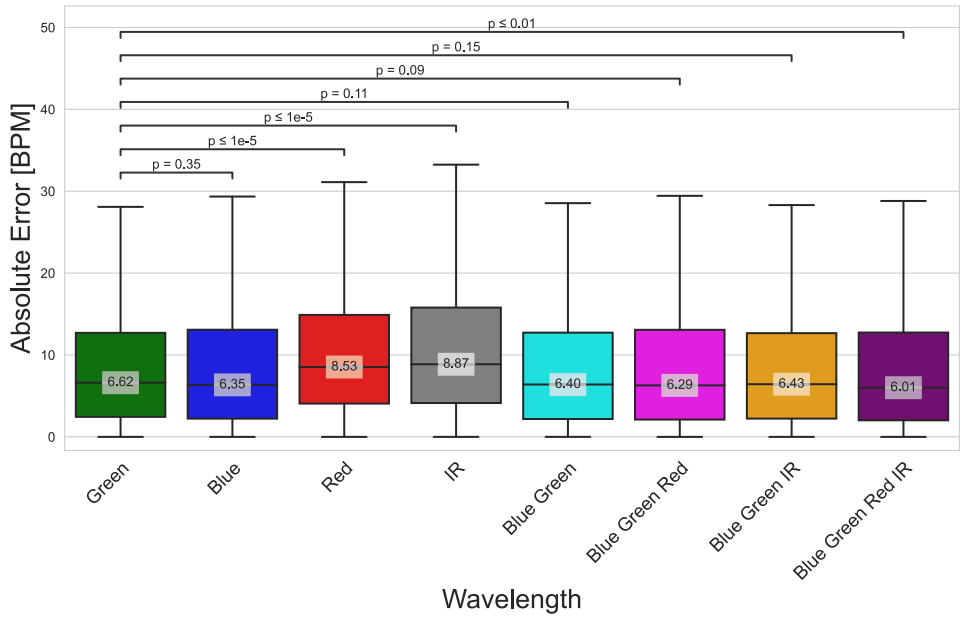


FIGURE 6.12: Comparison of Distributions of Absolute Error by Wavelength for Active Rest for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

In running, characterised by periodic motion, blue light was the best-performing single

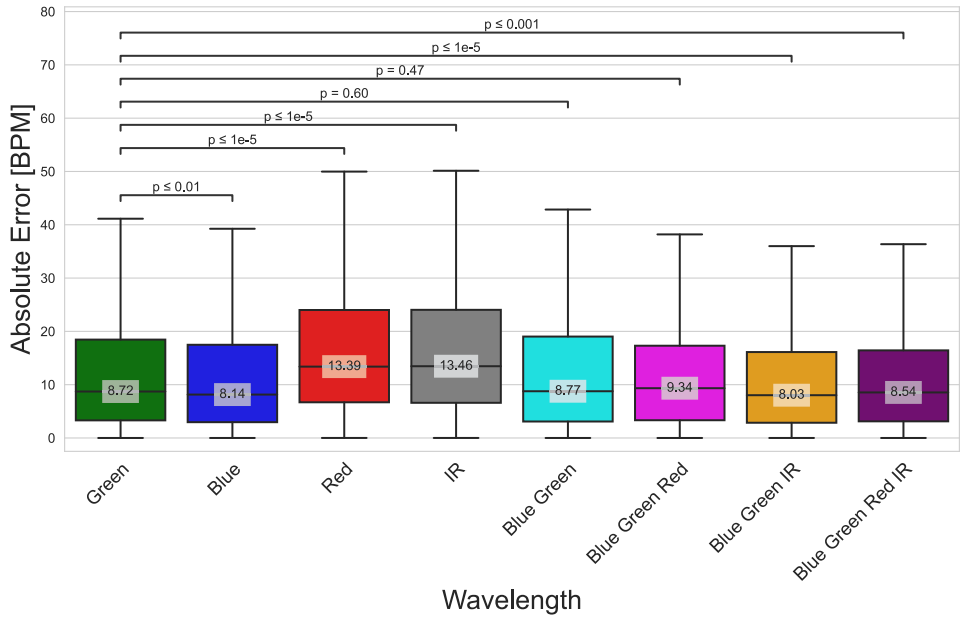


FIGURE 6.13: Comparison of Distributions of Absolute Error by Wavelength for Running for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

wavelength, with a 0.1 BPM lower MAE compared to green light, which was statistically significant ($p \leq 0.01$). Among multi-wavelength combinations, Blue-Green-IR achieved the lowest MAE, showing a reduction of 1.3 BPM compared to green light, which was statistically significant ($p \leq 1e-5$). Both Blue-Green-IR and Blue-Green-Red-IR also outperformed green light, demonstrating MAE reductions of 0.7 BPM and 0.9 BPM, respectively (see Figure 6.13 and Table 6.5).

During rest, characterised by minimal movement, green light consistently showed the lowest MAE across all tested wavelengths and combinations. The blue light exhibited a similar performance with only a 0.1 BPM increase in MAE. Among multi-wavelength combinations, Blue-Green-Red demonstrated an increase in MAE of 0.7 BPM compared to green light (see Figure 6.14 and Table 6.5).

In cycling, where subjects maintain hand contact with handlebars and motion is minimal, blue light was the best-performing single wavelength, with an MAE 0.1 BPM lower than green light. Among multi-wavelength combinations, Blue-Green-Red-IR achieved the lowest MAE, reducing it by 1.2 BPM compared to green light. However, this difference was not statistically significant (see Figure 6.15 and Table 6.5).

This section evaluated the accuracy of wrist-worn PPG heart rate estimations across various wavelengths and combinations, compared to the conventional green light.

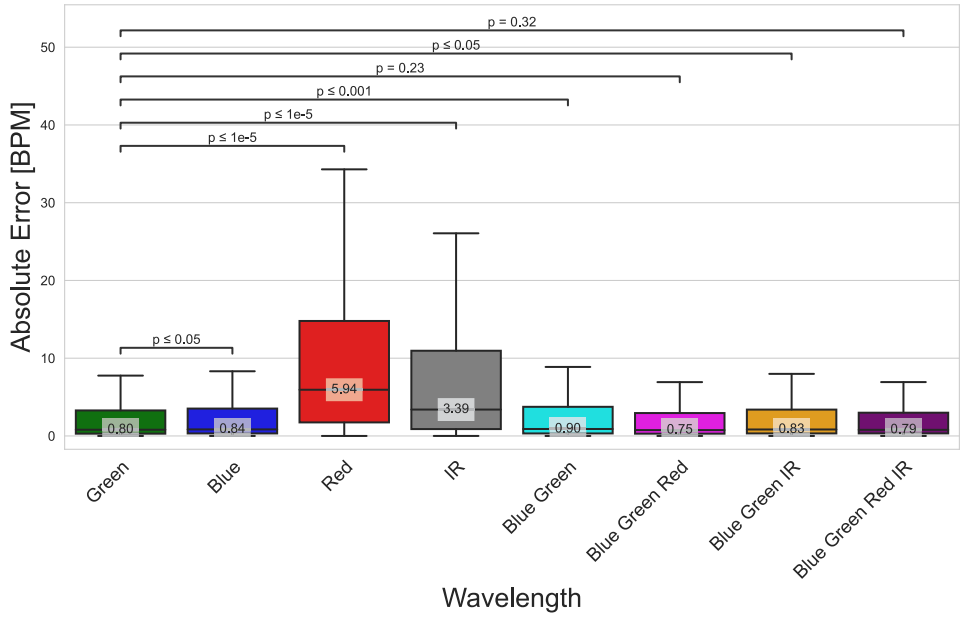


FIGURE 6.14: Comparison of Distributions of Absolute Error by Wavelength for Rest for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

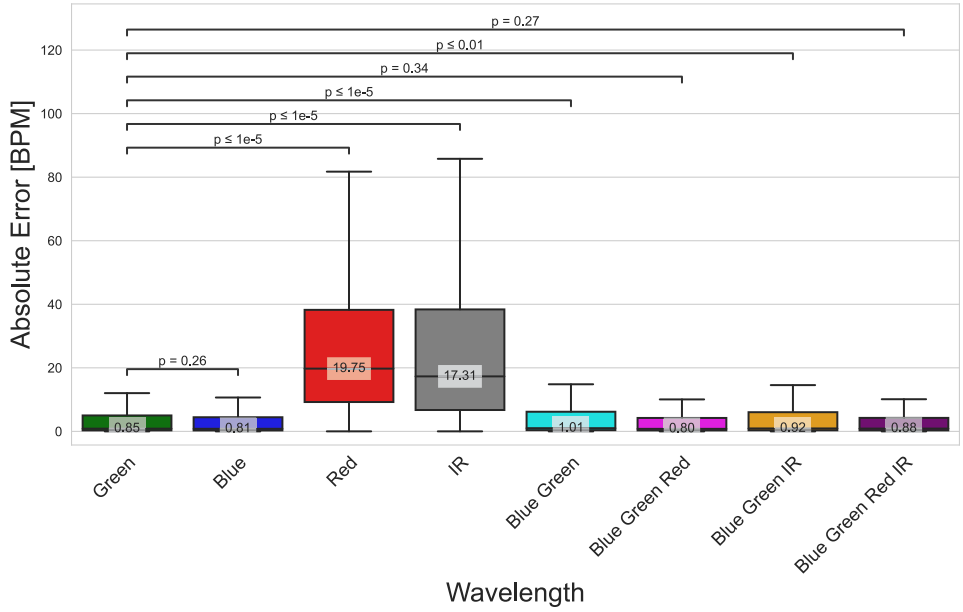


FIGURE 6.15: Comparison of Distributions of Absolute Error by Wavelength for Cycling for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

Longer wavelengths, such as red and IR, were associated with significantly higher MAE, whereas blue light showed comparable performance to green light. Multi-wavelength

	Green	Blue	Red	IR	Blue Green	Blue Green Red	Blue Green IR	Blue Green Red IR
Overall	8.1	8.2	15.8	15.3	8.6	8.1	8.3	**7.7**
Active Rest	9.0	9.2	10.8	11.7	9.2	9.0	9.0	**8.7**
Running	13.0	12.9	17.4	18.4	13.2	12.3	**11.8**	12.1
Rest	**3.4**	3.5	9.7	7.4	4.5	4.0	4.8	4.1
Cycling	7.3	7.2	25.9	24.9	7.8	7.6	7.8	**6.1**

All Values are MAE in BPM.

TABLE 6.5: Comparison of the Mean Absolute Errors of Wavelength and Wavelength Combinations over different Activities for MW PPG HR (This Work). **Bold** values indicate the lowest MAE value

approaches displayed varied results, with the Blue-Green-Red-IR combination providing the only significant improvement in MAE. The robustness assessment revealed that Blue-Green-Red-IR outperformed green light for activities involving motion, while green light remained the most accurate during rest.

6.5.2 The Influence of Demographic Variations

This section addresses objective 6 and the primary research question 2: What is the impact on performance based on variations in skin melanin content and biological sex in wrist-worn PPG heart rate estimation?

Analysis shown in Figure 6.16 reveals that individuals with higher skin melanin content experience a significant increase in absolute error, regardless of whether single or multi-wavelength methods are used. For single-wavelength methods, both blue and green wavelengths show similar error rates, but green (median absolute error = 4.02 BPM) performs slightly better than blue (median absolute error = 4.03 BPM) for individuals with higher melanin content. In multi-wavelength methods, the blue-green-red combination results in the lowest median absolute error for those with higher melanin content (median absolute error = 3.6 BPM), while the blue-green-red-IR combination is more effective for those with lower melanin content (median absolute error = 2.4 BPM).

When considering biological sex, the results are similar to those observed with skin melanin content, though the differences are less pronounced, as shown in Figure 6.17. For single-wavelength methods, the blue PPG sensor shows a lower median absolute error (3.2 BPM) compared to green (3.5 BPM) for females, although both errors are significantly higher than those for males, where blue (3.1 BPM) and green (2.9 BPM) have lower median absolute errors. In multi-wavelength methods, the blue-green-IR combination does not show a statistically significant difference between sexes. However,

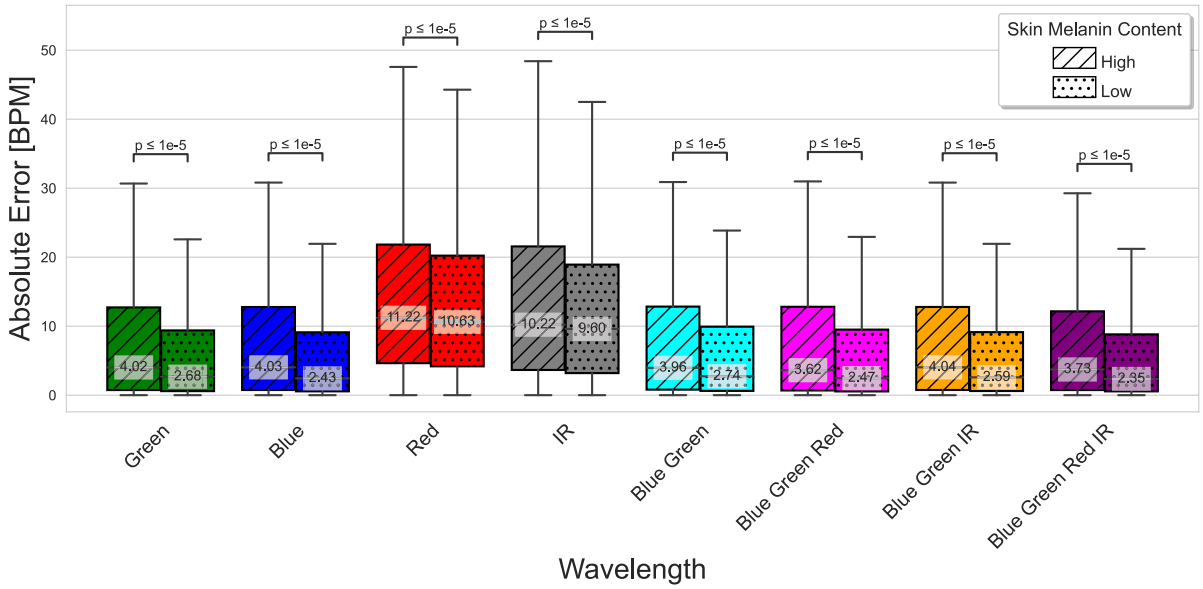


FIGURE 6.16: Comparison of Distributions of Absolute Error by Wavelength and Skin Melanin Content for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

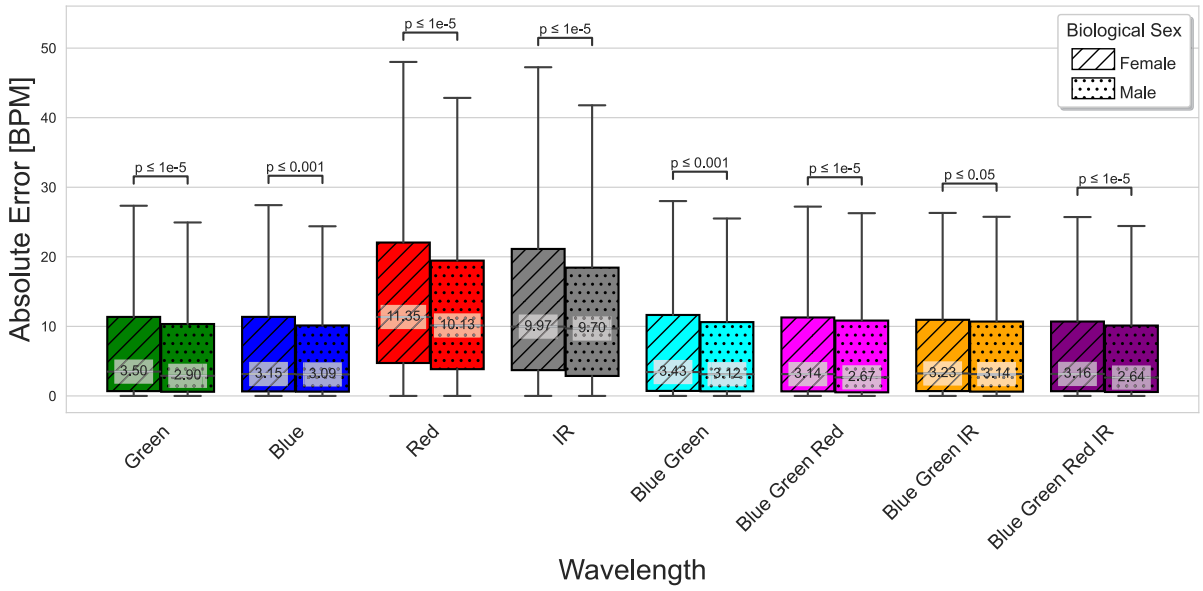


FIGURE 6.17: Comparison of Absolute Error Distributions by Wavelength and Biological Sex for MW PPG HR (This Work). The figure shows box plots illustrating the distribution of absolute errors, with the median, IQR, and 1.5 IQR whiskers displayed. Statistical analysis was performed using Mann-Whitney U tests.

its median absolute values (females = 3.2 BPM, males = 3.1 BPM) are comparable to the blue-green-red-IR combination (females = 3.2 BPM, males = 2.6 BPM), which provides the lowest median error for males. Conversely, the blue-green-red combination is most

effective for females (3.1 BPM).

Overall, the analysis confirms that both skin melanin content and biological sex impact the performance of wrist-worn PPG heart rate estimation. Higher melanin content is consistently associated with increased error rates across both single and multi-wavelength methods. Biological sex also affects performance, with different wavelength combinations showing varying effectiveness for males and females. These findings highlight the importance of considering demographic factors in developing wrist-worn PPG heart rate estimation methodologies. However, due to the limited sample size, these results are preliminary and further research with a larger sample is needed to better understand the effects of skin melanin content and biological sex.

6.5.3 Evaluation of Performance on Existing Single-wavelength Datasets

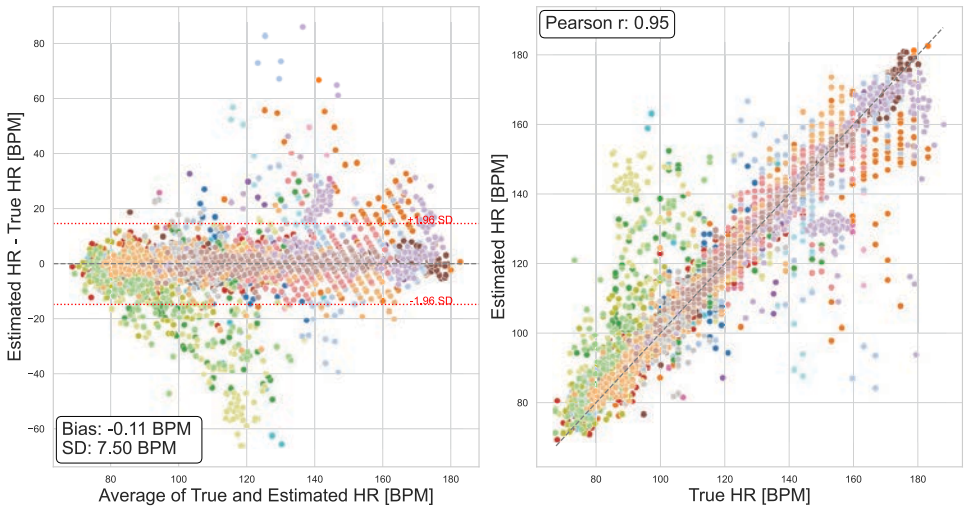


FIGURE 6.18: Bland-Altman and correlation analysis of estimated versus true heart rate measurements for the BAMI 1 dataset [127]. Each color represents a different subject. The Bland-Altman plot displays the bias and standard deviation (SD) of the differences, while the correlation plot shows Pearson's correlation coefficient (r).

This section addresses the objective 4. As mentioned in section 6.3.4, a LOSO CV scheme was used to assess the method's performance on unseen subjects. To evaluate the generalisability and robustness of the proposed method, the training process was repeated on existing datasets commonly used in PPG heart rate estimation method validation. Each dataset employed a different protocol and sensor configuration, as Section 2.2.1 described.

The datasets examined — IEEE Train [126], IEEE Test [126], BAMI 1 [127], BAMI 2 [127], and PPG DaLiA [130]—show a range of performance outcomes in heart rate estimation.

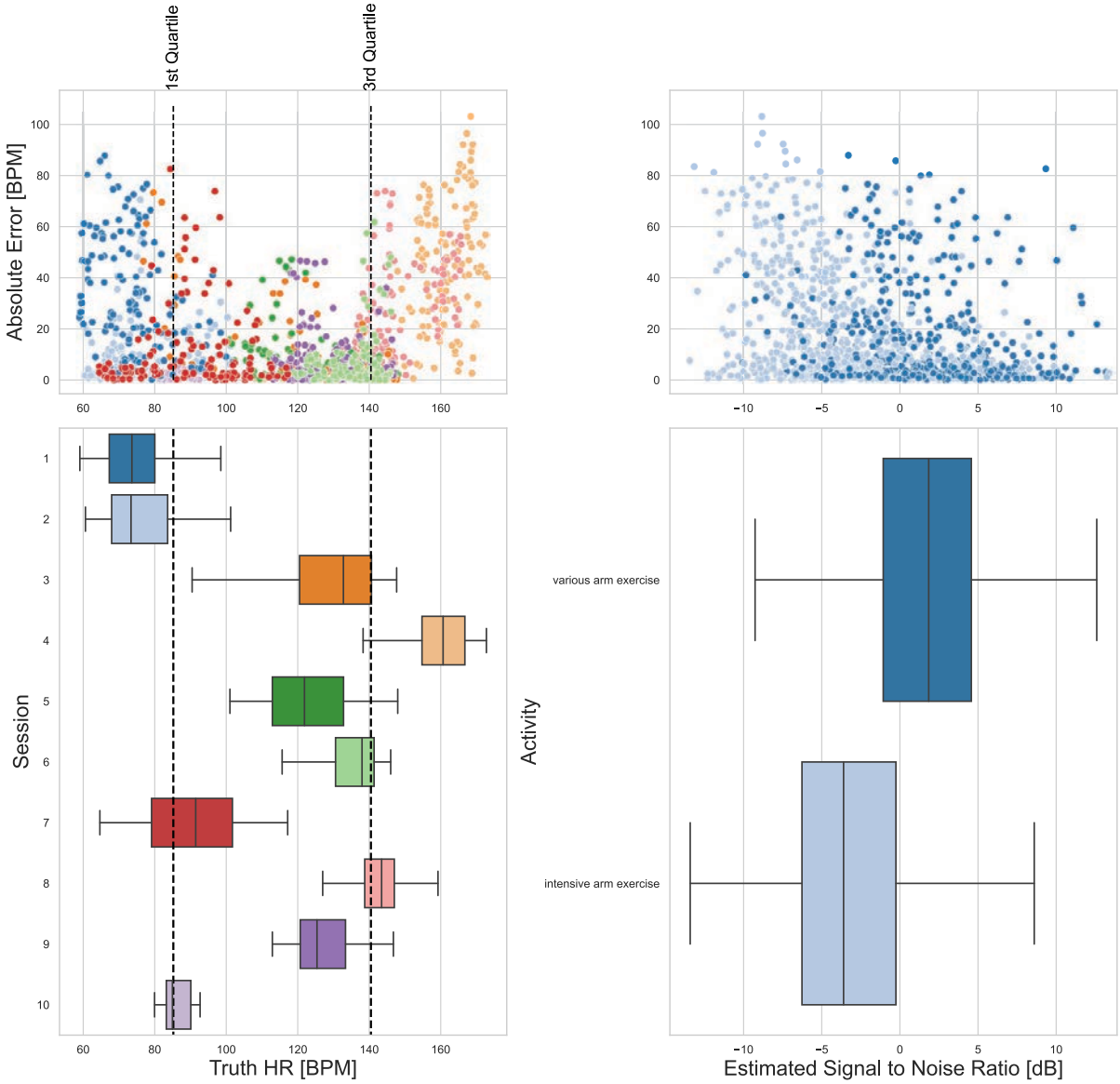


FIGURE 6.19: Comparative analysis of heart rate estimation accuracy across different sessions and activities, assessed against the true heart rate (HR) and ECG derived signal-to-noise ratio in the IEEE test dataset [126]. The left panel displays box-plot of subjects grouped by true HR, with the upper panel showing the corresponding absolute error. The right panel illustrates the ECG-derived SNR by activity, with the upper panel depicting the absolute error in relation to ECG SNR. Each box-plot represents the median, IQR, and 1.5 IQR whiskers.

The IEEE Train dataset achieved a moderate overall MAE of 5.8 ± 7.4 BPM with a correlation of $r = 0.89$ and bias of 1.4 BPM. While this dataset generally met the AAMI standard, variability was evident, particularly in subjects 4 (MAE = 10.2 BPM) and 11 (MAE = 29.7 BPM). In contrast, the IEEE Test dataset showed a marked drop in performance, with an MAE of 15.0 ± 11.8 BPM, a correlation of $r = 0.59$, and a bias of 1.8 BPM, failing to meet the AAMI standard and exhibiting high subject-specific variability.

The BAMI 1 dataset demonstrated promising performance, achieving an MAE of 3.6 ± 2.1 BPM with a high correlation of $r = 0.95$ and a bias of -0.1 BPM. A Bland-Altman plot for BAMI 1 illustrates the close agreement between estimated and true heart rates, reflecting its accuracy (see Figure 6.18). BAMI 2 outperformed all others with an exceptional MAE of 1.6 ± 0.6 BPM, a correlation of $r = 0.99$, and a bias of 0.2 BPM, indicating remarkable model generalisation and robustness across various activities. The PPG DaLiA dataset also performed well, meeting the AAMI standard with an MAE of 4.9 ± 3.1 BPM, a correlation of $r = 0.88$, and a bias of -0.04 BPM. However, stair climbing presented a challenge with an MAE of 14.1 BPM.

Across all datasets, two notable trends emerged. Firstly, a decrease in signal-to-noise ratio (SNR) generally led to higher absolute errors, highlighting the critical importance of signal clarity for accurate heart rate measurement. Secondly, larger datasets typically showed better performance than smaller ones. Specifically, for the IEEE Test dataset, Figure 6.19 illustrates that subjects with true heart rate values falling outside the IQR — represented by the dotted lines — tended to exhibit higher errors compared to those within the IQR.

6.5.4 Comparison with Conventional Heart Rate Estimation Methods

This sub-section addresses the primary research question 3: In PPG heart rate estimation, does deep learning demonstrate superior performance in terms of generalisability and robustness compared to conventional methods?

The comparison between the proposed method and conventional PPG beat detectors highlights the clear advantage of the proposed method in motion scenarios. In the PPG DaLiA dataset, the proposed method outperformed two leading conventional beat detectors (as referenced in [124]). Specifically, during cycling, the proposed method achieved a median absolute error of 4.2 BPM, compared to 13.0 BPM for MSPTD and 7.0 BPM for QPPG. In table soccer, the proposed method had a median absolute error of 6.7 BPM, whereas MSPTD and QPPG recorded 13.9 BPM and 19.1 BPM, respectively. However, during periods of minimal motion, MSPTD performed better than the proposed method, reducing the mean absolute percentage error by 1.0%.

Table 6.7 further demonstrates the superior performance of the proposed method in heart rate estimation across various activities in the MW PPG HR dataset. The proposed method shows a significantly lower median absolute error compared to conventional beat detection methods during low-motion activities. During periods of motion, it still outperforms the beat detectors, achieving 6.0 BPM in Active Rest compared to 12.9

	Cycling	Driving	Lunch	Sitting	Stairs	Table soccer	Walking	Working
This Method	**4.2**	**3.6**	**3.7**	3.5	**11.9**	**6.7**	**8.2**	**2.9**
MSPTD	13.0	5.7	6.7	**2.5**	20.1	13.9	19.1	4.3
qppg	7.0	7.8	8.2	4.3	15.1	19.1	13.7	8.0

All Values are MAPE (%).

TABLE 6.6: Performance Comparison of Proposed Method Against Conventional PPG Beat Detectors on PPG DaLiA [130] Activities. Results of other methods obtained from [124]. **Bold** values indicate the lowest MAPE value.

BPM with WFD using green light, and 8.5 BPM in Running compared to 29.7 BPM with QPPG using blue light.

	Active Rest		Running		Rest		Cycling	
	Method	Median AE	Method	Median AE	Method	Median AE	Method	Median AE
Beat Detector	Green WFD	12.9	Blue QPPG	29.7	Blue ERMA	11.5	Green QPPG	17.2
This Method	—	**6.0**	—	**8.5**	—	**0.8**	—	**0.9**

TABLE 6.7: Performance Comparison of Proposed Method Against Conventional PPG Beat Detectors on MW PPG HR (This Work) Activities. Results of other methods obtained from Chapter 5. **Bold** values indicate the lowest Median Absolute Error.

When comparing the proposed method to established heart rate estimation techniques on the IEEE Train dataset, the conventional methods generally outperformed the proposed method across most subjects (see Table 6.8). All conventional methods, except for one case, demonstrated higher accuracy. Notably, the method by Schaeck 2017 was the most effective, achieving a MAE of 2.9 BPM, nearly half the MAE of the proposed method, which had a mean MAE of 5.8 BPM. Despite these results, the proposed method did show some advantages. It outperformed the SpaMA method by an MAE margin of 8 BPM, indicating that while the proposed method may have limitations with smaller datasets, it still offers certain benefits over specific conventional methods like SpaMA in particular scenarios.

For the IEEE Test dataset, the proposed heart rate estimation method generally underperformed compared to conventional methods, excelling in only one subject (see Table 6.9). Among the conventional methods, SpaMA was the most accurate, achieving a MAE of 9.2 BPM, which is 5.8 BPM lower than the proposed method. Notably, the proposed method did surpass the Schaeck (2017) method by 9.7 BPM, highlighting a specific strength in this comparison.

	1	2	3	4	5	6	7	8	9	10	11	12	13	Mean ± STD
This Method	7.8	4.4	2.0	10.2	1.3	1.3	2.6	2.1	3.4	**0.6**	29.7	4.7	5.3	5.8 ± 7.4
SpaMA [161]	51.1	19.2	1.7	1.9	5.1	3.1	3.8	2.1	2.0	60.8	2.3	3.9	/	13.1 ± 20.7
SpaMA Plus [130]	**3.2**	10.0	1.6	2.7	1.5	2.8	1.0	2.2	0.4	21.3	1.8	2.4	/	4.3 ± 5.9
Schaeck 2017 [162]	16	**2.6**	**0.6**	**1.4**	**0.9**	**1.2**	**0.9**	**0.7**	**0.9**	7.6	**0.9**	**1.2**	/	**2.9 ± 4.6**

All Values are MAE in BPM.

TABLE 6.8: Performance Comparison of Proposed Method Against Conventional PPG Heart Rate Estimators on IEEE Train Subjects [126]. Results of other methods obtained from [130]. **Bold** values indicate the lowest MAE for each subject.

	1	2	3	4	5	6	7	8	9	10	Mean ± STD
This Method	32.0	10.3	**8.1**	40.6	8.3	6.3	11.0	21.8	9.4	1.9	15.0 ± 11.8
SpaMA [161]	18.3	**2.4**	9.8	**37.5**	**5.4**	2.3	**2.6**	**10.8**	**2.5**	**0.4**	**9.2 ± 11.4**
SpaMA Plus [130]	**17.5**	3.4	13.7	53.3	9.1	**2.1**	2.7	13.1	7.4	0.9	12.3 ± 15.5
Schaeck 2017 [162]	38.4	50.3	13.8	77.6	9.3	2.5	14.1	25.5	9.3	5.8	24.7 ± 24

All Values are MAE in BPM.

TABLE 6.9: Performance Comparison of Proposed Method Against Conventional PPG Heart Rate Estimators on IEEE Test Subjects [126]. Results of other methods obtained from [130]. **Bold** values indicate the lowest MAE for each subject.

The proposed heart rate estimation method demonstrated a significant improvement over conventional methods when applied to the PPG DaLiA dataset (see Table 6.10). It achieved an overall MAE of 4.9 ± 3.1 BPM, nearly halving the MAE of the closest conventional method and outperforming the least accurate conventional method by nearly fourfold. The proposed method delivered the most accurate results for 13 out of 15 subjects, showcasing its substantial advantage in handling the data and scenarios typical of the PPG DaLiA dataset.

Analysis across different datasets reveals a clear trend: the proposed heart rate estimation method generally performs better with larger datasets. While conventional methods excelled over the proposed method in most cases within the smaller IEEE Train and Test datasets, the proposed method showed substantial improvements with the larger PPG DaLiA dataset. Here, it not only surpassed conventional methods significantly but also achieved the lowest MAE in nearly all subjects. This pattern highlights the

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean ± STD
This Method	**4.1**	**2.6**	**4.5**	**5.8**	**1.7**	**2.5**	**3.0**	**3.2**	**2.1**	**5.4**	**10.9**	**4.2**	**2.0**	12.8	8.5	**4.9 ± 3.1**
SpaMA [161]	11.9	14.8	9.5	17.2	39.3	16.8	15.9	15.2	17.2	9.1	21.6	12.6	9.5	10.7	12.2	15.6 ± 7.5
SpaMA Plus [130]	8.9	9.7	6.4	14.1	24.1	11.3	6.3	11.3	16.0	6.2	15.2	12.0	8.5	**7.8**	**8.3**	11.1 ± 4.8
Schaeck 2017 [162]	33.1	27.8	18.5	28.8	12.6	8.7	20.7	21.8	22.3	12.6	21.1	22.7	27.7	12.1	16.4	20.5 ± 7.1

All Values are MAE in BPM.

TABLE 6.10: Performance Comparison of Proposed Method Against Conventional PPG Heart Rate Estimators on PPG DaLiA Subjects [130]. Results of other methods obtained from [130]. **Bold** values indicate the lowest MAE for each subject.

method's improved accuracy with larger datasets, suggesting that more extensive data provides a more diverse and representative sample of heart rate information allowing the method to generalise more effectively.

6.6 Summary

This chapter addresses Objective 4 by detailing a deep learning method for heart rate estimation, from pre-processing to hyperparameter optimisation. It identifies the blue-green-red-IR wavelength combination as the most effective, reducing MAE by 0.4 BPM compared to conventional green light. In contrast, longer wavelengths such as red and IR show higher MAE, increasing by 7.7 BPM and 7.2 BPM respectively. Blue light's performance is similar to green light, with a negligible 0.1 BPM difference.

The chapter also covers Objective 5 and Research Question 1 by evaluating performance across different activities. The blue-green-red-IR combination outperforms green light during motion-based activities, with the blue-green-IR combination showing a 1.3 BPM improvement over green light in running. Overall, blue-green-red-IR provides significant benefits in motion-based conditions, while green light remains most accurate during rest periods.

Analysis reveals that higher skin melanin content leads to increased error rates. Single-wavelength methods show green (median AE = 4.02 BPM) performing slightly better than blue (median AE = 4.03 BPM) for high melanin. Multi-wavelength methods are more effective, with blue-green-red showing median AE = 3.6 BPM for high melanin, and blue-green-red-IR showing median AE = 2.4 BPM for low melanin. Additionally, demographic factors affect performance, with increased errors associated with higher melanin content and biological females. These findings address Research Question 2

and Objective 6, highlighting the need for further research due to the preliminary nature of these results.

The performance on single-wavelength PPG heart rate estimation datasets — IEEE Train [126], IEEE Test [126], BAMI 1 [127], BAMI 2 [127], and PPG DaLiA [130] — showed varied performance. IEEE Train had a moderate MAE of 5.8 ± 7.4 BPM ($r = 0.89$), with high variability in certain subjects. IEEE Test's MAE was 15.0 ± 11.8 BPM ($r = 0.59$), not meeting the AAMI standard. BAMI 1 and BAMI 2 showed strong performance, with BAMI 2 achieving an exceptional MAE of 1.6 ± 0.6 BPM ($r = 0.99$). PPG DaLiA met the AAMI standard with an MAE of 4.9 ± 3.1 BPM ($r = 0.88$) but struggled with stair climbing. Notably, decreased SNR led to higher errors, and larger datasets performed better.

This chapter highlights that the proposed method generally outperforms conventional techniques on larger datasets, as demonstrated by a 5% reduction in MAPE and a notable decrease in MAE on the PPG DaLiA dataset. Specifically, it achieved a MAE of 4.9 ± 3.1 BPM, which is nearly half that of the closest conventional method and nearly fourfold better than the least accurate conventional method. This performance contrasts with its relatively weaker results on the smaller IEEE Train and Test datasets, where conventional methods, including Schaeck 2017 [162] and SpaMA [161], demonstrated higher accuracy. For instance, the proposed method's mean MAE of 5.8 BPM on the IEEE Train dataset was higher than the 2.9 BPM achieved by Schaeck 2017, and its MAE of 15.0 BPM on the IEEE Test dataset was 5 BPM higher than SpaMA's 9.2 BPM. This analysis meets Objective 9 and Research Question 3, illustrating the method's enhanced accuracy and robustness with larger datasets and its superior performance across various activities and signal conditions.

Chapter 7

Uncertainty Quantification Techniques for Convolutional Neural Networks in Heart Rate Estimation

The previous chapter detailed designing and implementing a convolutional neural network for heart rate estimation from wrist-worn PPG signals, examining the influence of wavelength, skin melanin, and biological sex. This chapter shifts focus to quantifying uncertainty in Deep Neural Networks, analysing its role in enhancing accuracy, robustness, and reliability in photoplethysmography (PPG) heart rate estimation. A critical evaluation of various uncertainty quantification methodologies is presented, emphasising their calibration and impact on heart rate estimation performance. The chapter further investigates the complexity of uncertainty types and concludes by discussing the integration of uncertainty estimates in post-processing and contrasting proposed methods with existing deep-learning approaches. The aim is to highlight the importance of uncertainty quantification in DNNs for the robustness and reliability of wrist-worn PPG heart rate estimation systems in practical scenarios. The detailed versions of the software used are outlined in Section 3.5

7.1 Uncertainty Quantification in Deep Learning

A significant drawback to using Deep Neural Networks (DNN) is a lack of trust in the predicted values due to the high complexity and un-interpretability of the generated DNNs, mainly from deep and non-linear structures [222]. To increase the reliability, usability, and interpretability of DNNs, researchers have investigated ways to represent uncertainty or quantify the confidence of a given prediction within DNNs [230]. According to Gawlikowski et al., uncertainty within DNNs is caused by five main factors:

1. **Real-world Variability:** Distribution shifts, arising from changes in real-world scenarios compared to training data, can markedly impact DNN performance due to its sensitivity to such shifts.
2. **Measurement System Errors:** These arise from limited data, measurement noise, or label inaccuracies, affecting the 'truth values' used in DNN training.
3. **Architectural Errors:** The DNN's structure, encompassing its components and hyperparameters, directly influences its performance and uncertainty.
4. **Training Procedure Errors:** Factors like batch size, optimiser, learning rate, and stochastic decisions, including weight initialisation, shape the DNN's training, performance, and uncertainty.
5. **Unknown Data Errors:** Pertains to valid samples outside the trained DNN's domain or task.

To quantify the effects of these factors on the DNNs' predicted values, two uncertainty terms are widely used, namely 'aleatoric' and 'epistemic' uncertainty [222, 230, 231]. Aleatoric uncertainty captures the inherent data uncertainty (Factor 2), while epistemic uncertainty pertains to model uncertainty, reducible by addressing Factors 1, 3, 4 & 5 [230]. Estimating these uncertainty terms in addition to the predicted value is highly desirable in increasing the reliability, usability and interpretability of the DNN.

7.2 Aleatoric Uncertainty Quantification

Aleatoric uncertainty is a fundamental type of uncertainty in data, reflecting the inherent ambiguity in the relationship between truth values and input data [230, 232]. This uncertainty often emerges from irreducible noise in the data, such as sensor-specific disturbances or motion-related variability. Aleatoric uncertainty is divided into two categories: homoscedastic, which remains constant across various inputs, and heteroscedastic, which fluctuates with the input [230, 232]. To measure this uncertainty, several techniques have been used, including non-linear quantile regression with pinball loss [233], implicit generative models [234], and employing predictive interval quality metrics as learning goals [235]. A commonly employed strategy involves learning the conditional distribution of a target variable based on the input [236].

This study focuses on the method of learning the conditional distribution of a target variable, selected for its widespread application, minimal alterations required in network designs, and effectiveness in modelling heteroscedastic variance. This approach is particularly relevant for PPG sensing, given the variability in noise levels within the

PPG signals. Consequently, the network architecture and training process, as outlined in Chapter 6, are modified in two key ways. First, the model is adapted to produce two outputs: the mean $\mu(x)$ and variance $\sigma^2(x)$ of a distribution, which represents the heart rate estimate (μ) and the corresponding aleatoric uncertainty (σ), illustrated in Figure 7.1 [236].

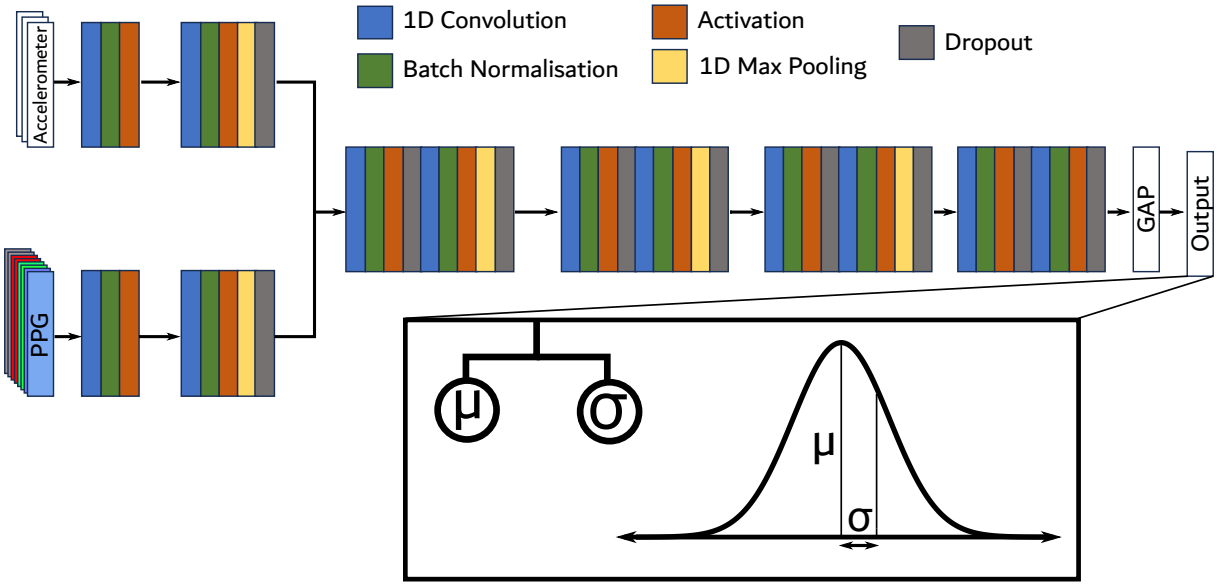


FIGURE 7.1: Modified Network Architecture for Aleatoric Uncertainty Quantification. This figure illustrates the adapted convolutional neural network architecture designed to quantify aleatoric uncertainty. The modification includes an output layer with two units, representing the mean (μ) and standard deviation (σ) of a predicted distribution. The loss function has been updated to the negative log-likelihood, shifting the model's objective from predicting a single value to predicting a distribution.

This adjustment in the network's output necessitates a modification of the training objective, as detailed in section 6.3.1. Instead of aiming to predict a single value that minimises the distance from the actual truth value, the goal shifts to predicting a distribution that maximises the likelihood of observing the truth value, given the input data. This change leads to a corresponding alteration in the loss function to align with the new objective. The Negative Log Likelihood (NLL) is chosen for its numerical stability and compatibility with the minimisation focus of optimisation algorithms [236].

Given the inherent variability in noise associated with wrist-worn PPG heart rate estimates, the choice of distribution for this model is critical. In this context, the logistic distribution is preferred over the more commonly used normal distribution. The logistic distribution is characterised by broader tails, which provides a higher tolerance for outliers, making it more suitable for dealing with the highly variable noise in PPG heart rate measurements. The NLL for the logistic distribution is derived from the log

probability density function, aggregated across all data points. For a dataset with n observations, the NLL is the negative sum of the log probability density function for each observation:

$$NLL(\mu, \sigma) = -\sum_{i=1}^{n}[-log(\sigma) - \frac{x_i - \mu}{\sigma} - 2 \cdot log(1 + e^{\frac{x_i - \mu}{\sigma}})] \tag{7.1}$$

where x_i is the truth value for the i^{th} sample, μ is the predicted mean, σ is the predicted standard deviation and n is the number of samples. As highlighted in section 7.3, sampling techniques are employed to quantify epistemic uncertainty; therefore, aleatoric uncertainty for input x_i is defined as $u_a(x_i) = \frac{1}{T}\sum_{t=1}^{T}\sigma_{i,t}^2$, where T is the number of samples.

7.3 Epistemic Uncertainty Quantification

Epistemic uncertainty stems from incomplete knowledge or information about the system being modelled [230, 232]. This type of uncertainty is distinct from aleatoric uncertainty, which is inherent and irreducible. In contrast, epistemic uncertainty can be mitigated or reduced by gathering more data, refining models, or deepening the understanding of the underlying processes [230, 232].

A common approach to epistemic uncertainty quantification is to apply Bayesian inference to derive a distribution over the network parameters using Bayes' theorem:

$$p(W|X, Y) = \frac{p(Y|X, W)p(W)}{p(Y|X)} \tag{7.2}$$

Where $p(W|X, Y)$ is the posterior distribution, which signifies the updated understanding or belief about the model parameters W after considering the observed data X and Y, $p(Y|X, W)$ is the likelihood indicating the probability of observing the data Y given the inputs X and a specific set of parameters W, $p(W)$ is the prior distribution that reflects the initial assumptions or knowledge about the parameters W before observing any data, $p(Y|X)$ is the evidence or marginal likelihood, serves to normalise the equation, ensuring that the posterior is a valid probability distribution, W is the network parameters and X, Y is the dataset. The posterior distribution represents the belief about the parameters after the data is observed, and to predict the distribution of output $y*$ corresponding to some new input $x*$, marginalisation over the posterior is employed to obtain the predictive posterior:

$$p(y^*|x^*, X, Y) = \int p(y^*|x^*, W)p(W|X, Y)dW \tag{7.3}$$

Where,$p(y^*|x^*, X, Y)$ posterior predictive distribution which gives the probability of a new data point y^* given a new input x^*, and the observed data X (inputs) and Y (outputs),$p(y^*|x^*, W)$ the likelihood of the new data point y^* given the new input x^* and a specific value of the parameters W and $p(W|X, Y)$ is the posterior distribution of the parameters W given the observed data X and Y. As this marginalisation is analytically intractable for deep non-linear models, it is common to sample W from $p(W|X, Y)$ to approximate the posterior distribution [230, 232]. This section will examine three of these sampling methods. Finally, to measure epistemic uncertainty in this study, the standard deviation of the approximated posterior distribution is taken.

7.3.1 Monte Carlo Dropout

Dropout operates by randomly deactivating network parameters during the training phase. This process prevents the network from becoming overly reliant on specific features, thereby encouraging a more robust feature detection capability and the prevention of overfitting [224]. Specifically, dropout layers multiply inputs element-wise by a binary dropout mask, Z, where $z_i \sim Bernoulli(p)$ and p is a fixed probability of deactivating inputs in the previous layer [224]. During the prediction phase, the dropout layers are typically disabled, allowing the network to achieve its full capacity.

Gal et al. proposed dropout variational inference as a posterior sampling method to quantify epistemic uncertainty [237]. This method involves applying Dropout not just during training but also during inference. This process, known as Monte Carlo dropout, allows for the generation of different "versions" of the model on each forward pass during inference, illustrated in Figure 7.2. These different versions are essentially samples from the approximate posterior distribution of the model's weights. From a theoretical standpoint, this approach is akin to performing approximate variational inference. The goal of variational inference is to find a simpler distribution (denoted as $q^*\theta(W)$) within a tractable family that minimises the Kullback-Leibler (KL) divergence to the true posterior distribution $p(W|X, Y)$ of the model weights given the data [236].

Monte Carlo Dropout acts as a variational Bayesian approximation. Each dropout mask applied during the forward passes can be seen as drawing from this simpler distribution $q^*\theta(W)$. This process approximates the true posterior by sampling from a range of possible model configurations, thereby capturing uncertainty in the model's predictions. This approach has several advantages, particularly in providing a practical and computationally efficient method to estimate uncertainty in deep learning models.

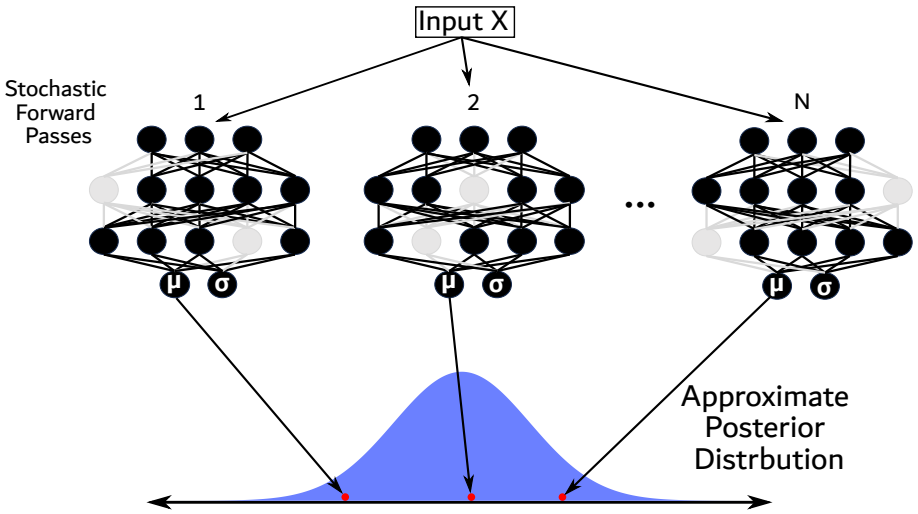


FIGURE 7.2: Monte Carlo Dropout Sampling of Posterior Distribution. This figure shows Monte Carlo dropout used to sample from the posterior distribution of a DNN's weights. Dropout masks randomly deactivate network parameters during both training and testing, generating multiple model 'versions' on each forward pass. These samples approximate the true posterior distribution, capturing epistemic uncertainty in the model's predictions.

It allows models to express uncertainty in their predictions, which is key in many applications like medical diagnosis or autonomous driving, where being confident in a prediction is as important as the prediction itself.

7.3.2 Concrete Dropout

A notable limitation of Monte Carlo dropout lies in fine-tuning the dropout rate to achieve well-calibrated uncertainty estimates. This tuning process requires a rigorous grid search scheme to find the optimal dropout rate. Such a scheme is computationally expensive and infeasible for larger models with multiple dropout layers due to the magnitude of the search space [238].

To address this issue, Gal et al. proposed Concrete Dropout, a continuous relaxation of the standard Dropout technique [238]. In the conventional dropout framework, the dropout vector z is binary, rendering the network loss function non-differentiable with respect to the dropout probability p. This binary nature of z prohibits the optimisation of p via backpropagation. Concrete Dropout offers a solution by sampling the elements of the dropout vector z from a continuous Concrete distribution [238]. This distribution is defined as follows:

$$\bar{z}_i = sigmoid\left(\frac{log(p) - log(1-p) + log(u) - log(1-u)}{t}\right) \tag{7.4}$$

where u is drawn from a uniform distribution, $u \sim Uniform(0,1)$. The smooth relationship between $\bar{z}$ and u is differentiable with respect to p, enabling the dropout probabilities to be treated as optimisable parameters within the neural network, illustrated in Figure 7.3.

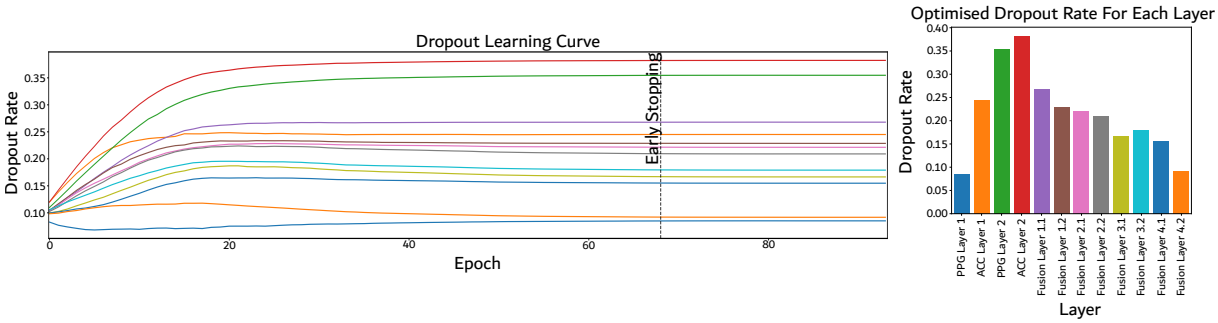


FIGURE 7.3: Concrete Dropout Learning Curve and Optimised Dropout Rates for Each Layer for Subject 13 of the PPG DaLiA Dataset [127]. This figure illustrates the optimisation process of dropout probabilities for each dropout layer in the network during training, with the optimised dropout rates displayed on the right. It highlights how these rates are fine-tuned to improve the model's uncertainty quantification capabilities.

This modification brings a twofold advantage. First, it eliminates the need for a grid search by allowing gradient-based optimisation methods to tune the dropout probability directly as part of the training process. Second, it ensures that the dropout rate adapts dynamically, reflecting the amount and variability of the training data. These improvements make Concrete Dropout particularly suited for large models and continuous learning scenarios, where maintaining calibrated uncertainty estimates is key [238]. Similarly to MC Dropout, during inference, the Dropout remains on and does N forward passes to obtain the μ and σ of the approximate posterior distribution to obtain the heart rate estimate and epistemic uncertainty estimate.

7.3.3 Deep Ensemble

Ensemble methods, which involve creating a committee of models with varying learned parameters and loss trajectories, have proven effective in enhancing predictive performance across various domains. The strength of ensembles lies in their ability to foster a diverse understanding of inputs. Ensembles generally yield more accurate results than a single network by averaging or using bootstrap aggregating (bagging) techniques for predictions. In PPG heart rate estimations, for instance, ensembles have demonstrated superior performance compared to single-network models [130].

The diversity in these ensemble models typically does not stem from different network architectures. Instead, it arises from the random initialisations of weights and the

randomised order in which training data is presented to the models. This approach ensures that each model in the ensemble develops a slightly different perspective on the data, leading to a richer understanding of the inputs.

Lakshminarayanan et al. proposed deep ensembles to quantify epistemic uncertainty by treating ensembles as a uniformly weighted mixture model, where predictions from individual models are combined. In this formulation, the prediction is combined as $p(y^*|x^*) = \frac{1}{M}\sum_{i=1}^{M} pW_m(y^*|x^*, W_m)$ where W_m corresponds to the network parameters of model m and M is the number of models in the ensemble [239].

In regression tasks, the prediction is often assumed to be a mixture of Gaussian distributions. For simplicity and ease of inference, the mean and variance of this mixture are assumed to be the mean and variance of a single Gaussian distribution. The mean and variance of the ensemble's output are calculated by averaging the means and variances of the predictions from each model [239].

Deep ensembles have been found to outperform MC dropout in quantifying uncertainty across various datasets and tasks in both regression and classification settings. They are particularly effective in out-of-distribution scenarios, such as data perturbations, or when encountering new classes not seen during training [239]. For practical implementation, Lakshminarayanan et al. recommend using a sample size of 5 models in the ensemble [239]. This recommendation balances the need for diversity in predictions with computational efficiency, ensuring a robust ensemble without excessively taxing computational resources.

7.4 Evaluation of Uncertainty Quantification Methods

This section addresses objective 7: 'Compare and evaluate aleatoric and epistemic uncertainty methods in deep learning, focusing on calibration, their distinctness or entanglement, and their relation to error rates and signal quality.' as well as primary research question 4: Which method for quantifying epistemic uncertainty yields the best calibrated metrics in the context of PPG heart rate estimations?

7.4.1 Aleatoric Uncertainty Quantification

This section provides an analysis of the aleatoric uncertainty quantification method detailed in Section 7.2. The analysis is designed to illuminate the behaviour of aleatoric uncertainty under various conditions and assess its impact on the predictive performance of the model (See Figure 7.4). A deeper understanding of how aleatoric uncertainty

is quantified can significantly improve the accuracy and reliability of predictions, ensuring the model's utility in practical scenarios. The evaluation of this uncertainty type is essential for models expected to function in the highly variable and often unpredictable domain of biological data.

Across the datasets used—IEEE Train [126], IEEE Test [126], BAMI 1 [127], BAMI 2 [127], and PPG DaLiA [130] and MW PPG HR (This Work) —the performance of aleatoric uncertainty quantification varied. For the IEEE Train dataset, the aleatoric uncertainty was under-confident, with a miscalibration area of 0.2, a correlation coefficient with SNR of 0.0, and an absolute error of 0.21. In the IEEE Test dataset, it was over-confident, showing a miscalibration area of 0.13, a correlation coefficient with SNR of -0.41, and an absolute error of 0.04. The BAMI-1 dataset also exhibited under-confidence, with a miscalibration area of 0.13, a correlation coefficient with SNR of -0.22, and an absolute error of 0.29. Similarly, in the BAMI-2 dataset, aleatoric uncertainty was under-confident, with a miscalibration area of 0.25, a correlation coefficient with SNR of -0.38, and an absolute error of 0.31. The PPG DaLiA dataset showed under-confidence as well, with a miscalibration area of 0.07, a correlation coefficient with SNR of -0.64, and an absolute error of 0.59. Lastly, in the MW PPG HR dataset (as shown in Figure 7.4), the aleatoric uncertainty was under-confidence in lower aleatoric uncertainty values and overconfidence in higher ones, with a miscalibration area of 0.09, a correlation coefficient with SNR of -0.32, and an absolute error of 0.44. Aleatoric uncertainty was expected to have a strong negative correlation with ECG-derived signal-to-noise ratio (SNR). However, the results did not fully meet these expectations, though some signs of effectiveness were observed.

Figure 7.5 presents a comparison of aleatoric uncertainty across different demographic groups for the MW PPG Dataset, specifically focusing on skin melanin content and biological sex, during various activities. For skin melanin content, a statistically significant difference in aleatoric uncertainty distributions was observed only during running, with a median aleatoric uncertainty of 6.1 BPM for individuals with low melanin content and 6.4 BPM for those with high melanin content. In contrast, when examining biological sex, statistically significant differences in aleatoric uncertainty distributions were observed during active rest, rest, and cycling. During active rest, the median aleatoric uncertainty was 5.4 BPM for females and 5.5 BPM for males. In the rest condition, females had a median aleatoric uncertainty of 3.8 BPM compared to 4.1 BPM for males. During cycling, females exhibited a median aleatoric uncertainty of 6.5 BPM, while males had a median of 6.0 BPM.These findings suggest that biological sex consistently affects aleatoric uncertainty across activities more than skin melanin content. However, the results are not entirely as expected, implying that aleatoric uncertainty may not

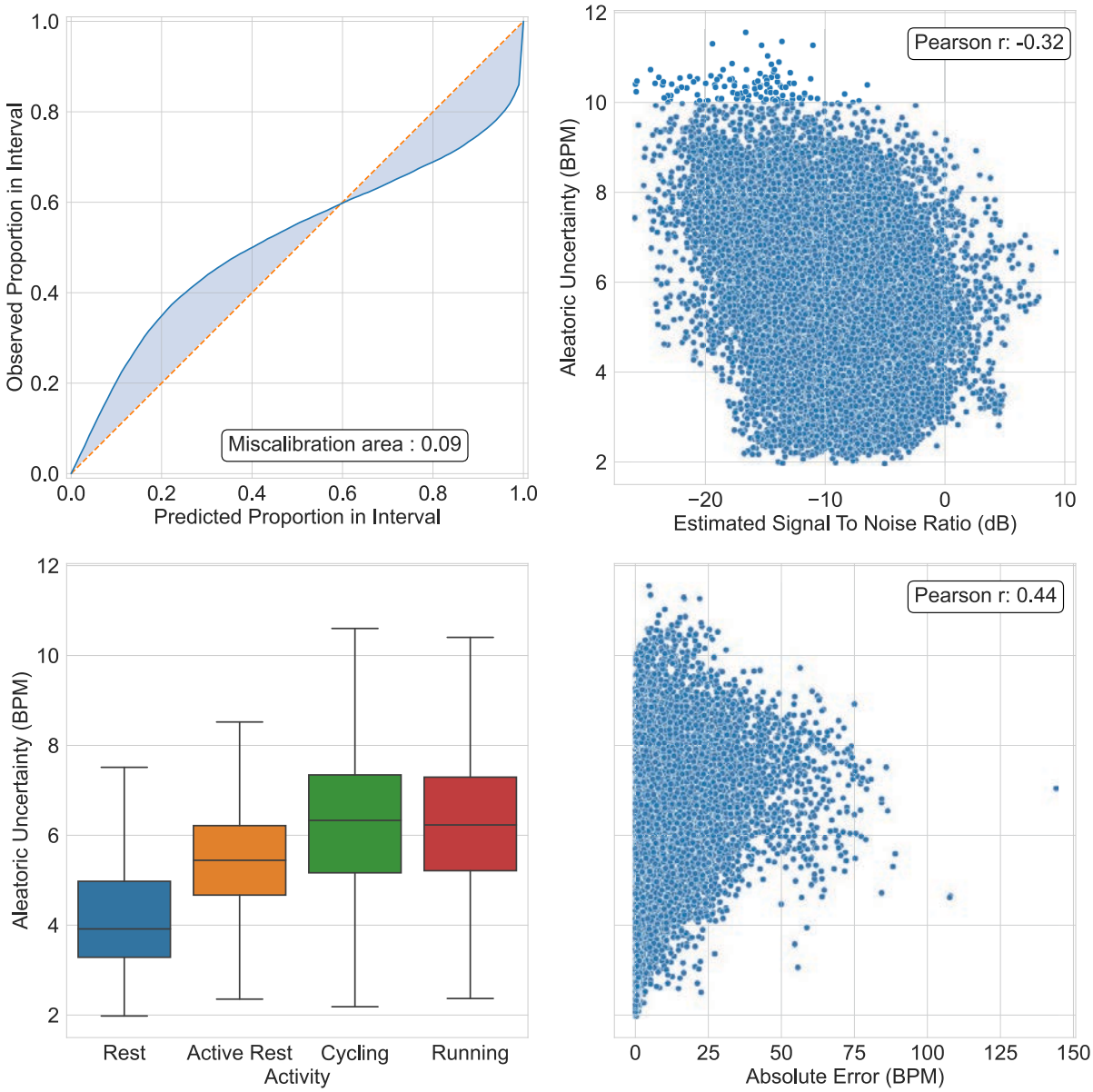


FIGURE 7.4: Aleatoric Uncertainty Calibration and Performance Analysis for MW PPG HR Dataset (This Work). The figure includes a calibration plot showing under-confidence in lower aleatoric uncertainty values and over-confidence in higher ones (top left). This can be understood in the context of regression calibration, where predictions are considered calibrated if, for example, a 50% confidence level holds true in 50% of cases, represented by the orange line. A weak positive correlation between electrocardiogram (ECG) derived signal-to-noise ratio (SNR) and aleatoric uncertainty is observed (top right). Interestingly, the distribution of aleatoric uncertainty across activities does not match expected noise levels (bottom left). Although rest shows low uncertainty as expected, cycling, which was also anticipated to have low uncertainty, does not. The box plots display the median, IQR, and 1.5 IQR whiskers. A moderately positive correlation between absolute error and aleatoric uncertainty was also observed (bottom right).

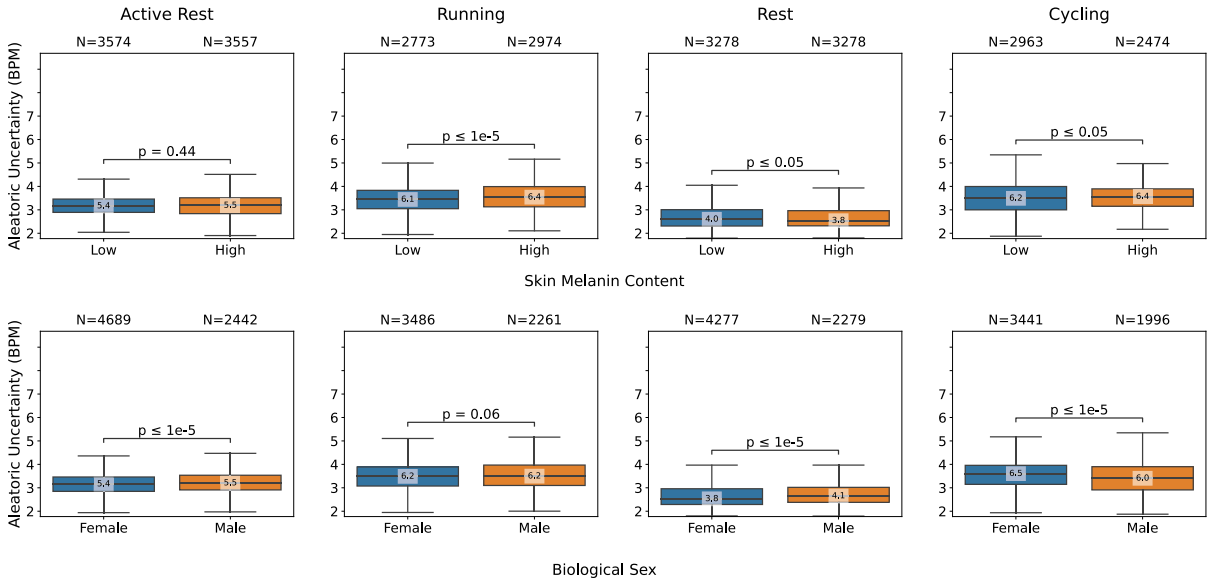


FIGURE 7.5: Effect of Skin Melanin Content and Biological Sex on Aleatoric Uncertainty Across Activities for the MW PPG Dataset (This Work). The figure presents box plots of aleatoric uncertainty distributions (via Distributional Estimation) across skin melanin content (top) and biological sex (bottom). The plots show the median, IQR, and 1.5 IQR whiskers. "N=" denotes the number of 8-second window samples for each activity and demographic variation. A weak trend in activity-related changes is observed, with lower uncertainty for rest, but cycling shows distributions similar to running and active rest, challenging initial assumptions of aleatoric uncertainty. Mann-Whitney U tests were used for statistical analysis.

solely reflect data uncertainty. The variations across demographic groups and activities suggest other factors may be influencing the model's predictions, highlighting a limitation in using aleatoric uncertainty as the sole measure of data uncertainty in PPG-based heart rate estimation methods.

Building on the previous analysis, Figure 7.6 explores the relationship between aleatoric and epistemic uncertainty by introducing varying levels of random noise to the input data. As anticipated, aleatoric uncertainty increases with added noise, consistent with its role in capturing data-inherent randomness. However, epistemic uncertainty, quantified through concrete dropout, exhibits a more pronounced increase in response to noise. This suggests an interplay between the two types of uncertainty, where epistemic uncertainty is particularly sensitive to external perturbations. The greater impact on epistemic uncertainty highlights a challenge in accurately calibrating and interpreting the model's predictions when data noise is present, as these intertwined uncertainties can complicate the distinction between data noise and model confidence.

In summary, this section thoroughly evaluated aleatoric uncertainty quantification across various datasets, demographic groups, and activities. The findings indicate that

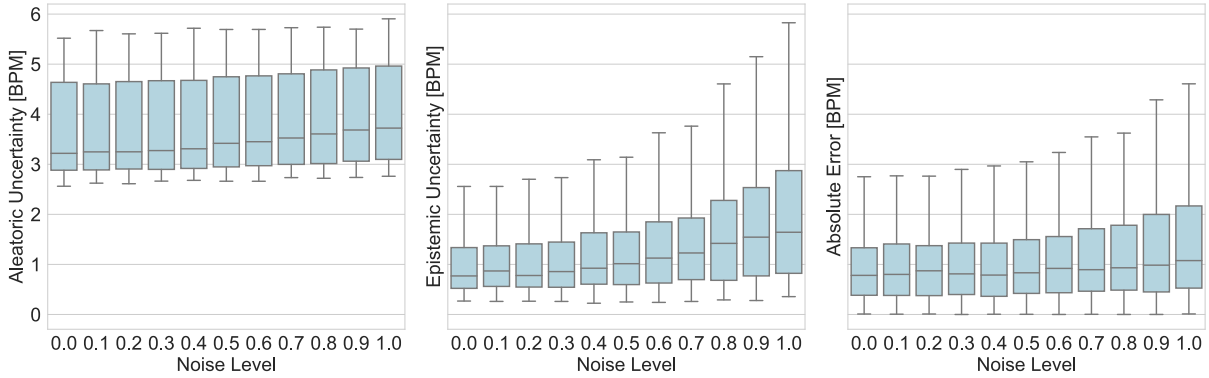


FIGURE 7.6: Impact of Varied Random Noise Levels on Aleatoric Uncertainty, Epistemic Uncertainty, and Absolute Error in the BAMI 2 Dataset [127]. The figure shows the effect of adding random noise with varying standard deviations - 'Noise Levels'. The box-plots display the median, IQR, and 1.5 IQR whiskers. The impact is more pronounced on epistemic uncertainty and absolute error, challenging the assumption that aleatoric uncertainty solely reflects data uncertainty.

while aleatoric uncertainty generally aligns with expectations, it does not consistently correlate with data uncertainty alone, particularly across different demographic factors and noise levels. The observed interplay between aleatoric and epistemic uncertainties underscores the complexity of accurately quantifying and interpreting uncertainty in wrist-worn PPG heart rate estimation methods, especially in diverse and noisy data environments.

7.4.2 Epistemic Uncertainty Quantification

This section provides a comparative analysis of the three epistemic uncertainty quantification methods detailed in Section 7.3. A key aspect of this analysis is determining the optimal number of samples for each method to ensure the most accurate measurement of epistemic uncertainty. As detailed in section 7.3.3, the ensemble method recommends 5 samples, however, the samples for MC Dropout and Concrete Dropout need to be determined. The IEEE Train dataset, chosen for its smaller size, serves as the basis for this analysis. The objective is to assess the sample size's influence on the miscalibration area and MAE.

Figure 7.7 demonstrates that using 25 samples achieves an optimal balance between MAE and the miscalibration area. At this point, the miscalibration area starts to increase while the MAE continues to decrease. Furthermore, the time required for each prediction with 25 samples is just under one second, offering a practical compromise between computational efficiency and prediction accuracy. This is particularly important since

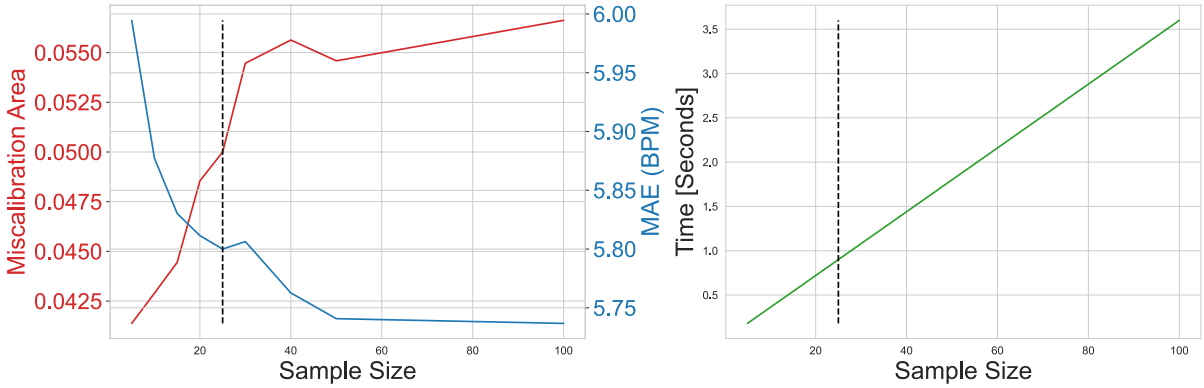


FIGURE 7.7: Analysis of Epistemic Sample Size Impact on Miscalibration Area, Mean Absolute Error, and Prediction Time for the IEEE Train Dataset [126]. The figure highlights the trade-offs associated with the Concrete Dropout sampling method for estimating epistemic uncertainty. It demonstrates the challenge of completing n forward passes within the 2-second window of the data slide, emphasising the balance between estimation accuracy, uncertainty accuracy and prediction time.

any estimation time exceeding two seconds would be infeasible, given that the slide of the windowing is two seconds.

Method	IEEE Train [126]	IEEE Test [126]	BAMI 1 [127]	BAMI 2 [127]	PPG DaLiA [130]	MW PPG HR (This Work)
MC Dropout	0.20	0.35	0.28	0.14	0.26	0.23
Concrete Dropout	**0.05**	**0.20**	**0.17**	**0.08**	**0.22**	**0.12**
Ensemble	0.22	0.31	0.31	0.28	0.32	0.24

TABLE 7.1: Comparison of Miscalibration Area by Epistemic Uncertainty Method across All Utilised Datasets. The miscalibration area is computed for all subjects within each dataset using a LOSO CV scheme. **Bold** values indicate the lowest miscalibration area for each dataset.

The evaluation of epistemic uncertainty quantification methods, detailed in Table 7.1, reveals Concrete Dropout as the most effective across all datasets. It consistently shows the lowest miscalibration areas, with values of 0.05 for IEEE Train, 0.20 for IEEE Test, 0.17 for BAMI 1, 0.08 for BAMI 2, 0.22 for PPG DaLiA, and 0.12 for MW PPG HR. This consistent performance underscores Concrete Dropout's reliability and accuracy in estimating uncertainty, making it particularly suitable for robust wrist-worn PPG heart rate estimation models. In response to Research Question 4, the analysis will now focus solely on Concrete Dropout to further explore its effectiveness in uncertainty quantification.

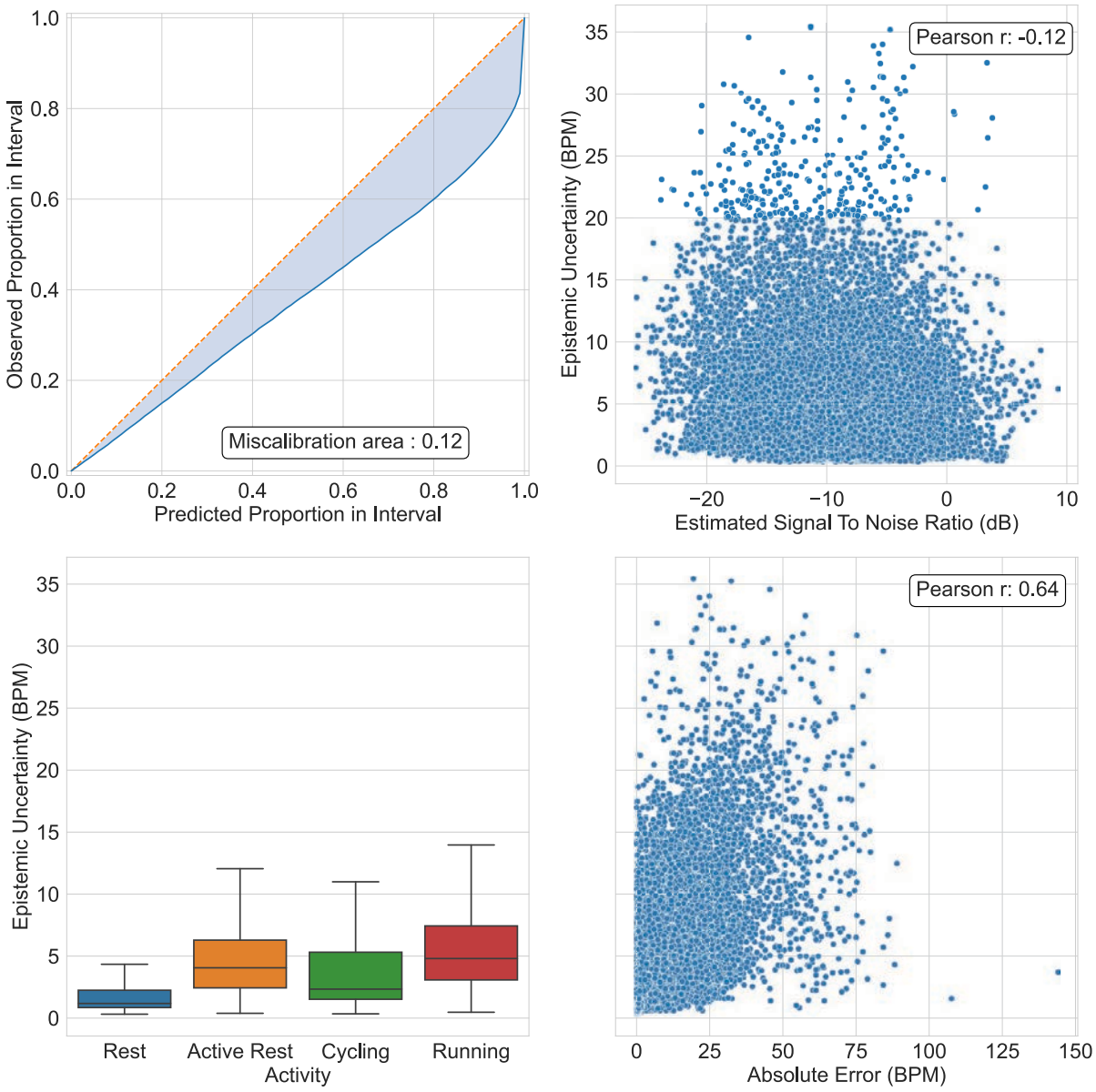


FIGURE 7.8: Epistemic Uncertainty Calibration and Performance Analysis for MW PPG HR Dataset (This Work). The figure includes a calibration plot revealing overconfidence in epistemic uncertainty values (top left). This can be understood in the context of regression calibration, where predictions are considered calibrated if, for example, a 50% confidence level holds true in 50% of cases, represented by the orange line. A weak negative correlation between ECG-derived SNR and epistemic uncertainty is observed (top right). Notably, the distribution of epistemic uncertainty across activities aligns with expected noise levels, with rest and cycling showing lower uncertainty than active rest and running (bottom left). The box-plots display the median, IQR, and 1.5 IQR whiskers. The was also a moderately positive correlation observed between absolute error and epistemic uncertainty (bottom right).

Across the datasets used—IEEE Train [126], IEEE Test [126], BAMI 1 [127], BAMI 2 [127], PPG DaLiA [130], and MW PPG HR (This Work)—the performance of epistemic uncertainty quantification (via concrete dropout) varied but was more consistent than that of aleatoric uncertainty. For the IEEE Train dataset, epistemic uncertainty was under-confident, with a miscalibration area of 0.05, a correlation coefficient with SNR of -0.3, and an absolute error of 0.78. In the IEEE Test dataset, it was over-confident, with a miscalibration area of 0.2, a correlation coefficient with SNR of -0.22, and an absolute error of 0.45. The BAMI-1 dataset also exhibited over-confidence, showing a miscalibration area of 0.17, a correlation coefficient with SNR of -0.3, and an absolute error of 0.64. Similarly, in the BAMI-2 dataset, epistemic uncertainty was over-confident, with a miscalibration area of 0.08, a correlation coefficient with SNR of -0.37, and an absolute error of 0.65. The PPG DaLiA dataset also showed over-confidence, with a miscalibration area of 0.22, a correlation coefficient with SNR of -0.43, and an absolute error of 0.58. Lastly, in the MW PPG HR dataset (as shown in Figure 7.8), epistemic uncertainty exhibited over-confidence, with a miscalibration area of 0.12, a correlation coefficient with SNR of -0.12, and an absolute error of 0.64. While epistemic uncertainty was expected to have no correlation with ECG-derived signal-to-noise ratio (SNR), a correlation was observed. However, as anticipated, epistemic uncertainty showed a strong correlation with absolute error.

Figure 7.9 compares epistemic uncertainty across different demographic groups, focusing on skin melanin content and biological sex during various activities. For skin melanin content, significant differences in epistemic uncertainty were observed across all activities. The largest difference occurred during running, where individuals with low skin melanin had a median epistemic uncertainty of 4.2 BPM, while those with high melanin had 5.7 BPM. Notably, despite similar sample sizes for low and high melanin groups across activities, the uncertainty distributions differed. This suggests that, contrary to expectations that more data should equalise distributions, epistemic uncertainty varies with skin melanin content. In contrast, significant differences in epistemic uncertainty were only found for running and cycling when considering biological sex. During running, females had a median epistemic uncertainty of 5.0 BPM compared to 4.6 BPM for males. In cycling, females had a median of 2.7 BPM, whereas males had 1.8 BPM. Additionally, although there were nearly three times more samples for females in cycling, their uncertainty distribution remained higher.

This section evaluates epistemic uncertainty quantification methods, focusing on Concrete Dropout, MC Dropout, and the ensemble method, across various datasets. Concrete Dropout consistently outperforms other methods, demonstrating the lowest miscalibration areas and thus proving to be the most reliable for estimating uncertainty

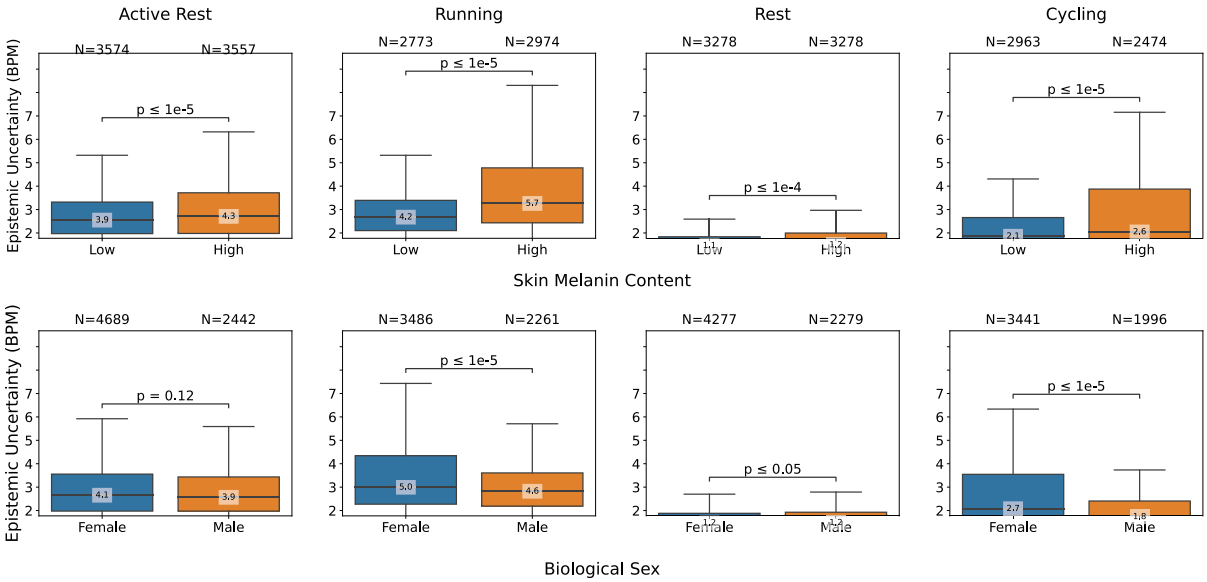


FIGURE 7.9: Effect of Skin Melanin Content and Biological Sex on Epistemic Uncertainty Across Activities for MW PPG Dataset (This Work). This figure shows box plots of epistemic uncertainty distributions (via Concrete Dropout) across skin melanin content (top) and biological sex (bottom). The plots display the median, IQR, and 1.5 IQR whiskers. 'N=' indicates the number of 8-second window samples for each activity and demographic variation. While epistemic uncertainty is theoretically reducible by adding data, the figure shows that more data does not necessarily reduce uncertainty, and a trend of increased uncertainty with motion is evident. Mann-Whitney U tests were used for statistical analysis.

in wrist-worn PPG heart rate estimation methods. Analysis of sample size reveals that 25 samples strike a practical balance between mean absolute error (MAE) and miscalibration area, while maintaining computational efficiency. Although epistemic uncertainty showed variability across datasets, it was generally more consistent than aleatoric uncertainty. Notably, Concrete Dropout exhibited robust performance, with miscalibration areas ranging from 0.05 to 0.22 across different datasets. Furthermore, epistemic uncertainty varied with demographic factors, showing higher values for individuals with high skin melanin content and females, suggesting that factors beyond sample size impact uncertainty. This analysis highlights Concrete Dropout's effectiveness and its strong correlation with absolute error, while also pointing to the interplay between aleatoric and epistemic uncertainties in the evaluated methods.

7.4.3 Effect of Uncertainty Quantification Methods on Heart Rate Estimation Performance

The assessment of epistemic uncertainty quantification methods—MC Dropout, Concrete Dropout, and the ensemble method—reveals their distinct impacts on heart rate

estimation accuracy, as shown in Table 7.2.

UQ Method	IEEE Train [126]	P Value	IEEE Test [126]	P Value	BAMI 1 [127]	P Value	BAMI 2 [127]	P Value	PPG DaLiA [130]	P Value	MW PPG HR (This Work)	P Value
None	5.8 ± 7.4	—	15.0 ± 11.8	—	3.6 ± 2.1	—	1.6 ± 0.6	—	4.9 ± 3.2	—	7.5 ± 3.2	—
MC Dropout	7.8 ± 9.3	0.100	15.8 ± 11.0	0.312	3.9 ± 2.3	0.334	1.8 ± 1.0	0.448	5.1 ± 3.5	0.452	7.9 ± 2.8	0.245
Concrete Dropout	**5.7 ± 7.6**	0.417	14.7 ± 10.6	0.484	3.5 ± 2.1	0.341	1.8 ± 1.2	0.413	4.9 ± 3.1	0.433	**7.3 ± 2.5**	0.496
Ensemble	6.1 ± 11.1	0.251	**14.5 ± 11.8**	0.364	**3.3 ± 2.1**	0.192	**1.4 ± 0.9**	0.041	**4.8 ± 3.3**	0.417	7.5 ± 2.7	0.409

All Values are MAE in BPM. Statistical Tests used the Mann-Whitney U test.

TABLE 7.2: Comparison of Heart Rate Estimation Performance by Epistemic Uncertainty Method For All Utilised Datasets. Statistical tests compare each method individually to the base method using no UQ method (None). **Bold** values indicate the lowest MAE distribution.

MC Dropout consistently results in the highest mean absolute errors (MAE), with values such as 7.8 ± 9.3 BPM on the IEEE Train dataset and 7.9 ± 2.8 BPM on the MW PPG HR dataset. This indicates that, although MC Dropout is effective at quantifying uncertainty, it does not improve heart rate estimation performance. In contrast, Concrete Dropout generally shows better performance, achieving lower MAE values on most datasets, including 5.7 ± 7.6 BPM on the IEEE Train dataset and 7.3 ± 2.5 BPM on the MW PPG HR dataset. However, its performance in BAMI-2 did not show significant differences compared to the base model without uncertainty quantification in Chapter 6. The ensemble method, meanwhile, achieved the lowest MAE across several datasets: 14.5 ± 11.8 BPM on IEEE Test, 3.3 ± 2.1 BPM on BAMI 1, 1.4 ± 0.9 BPM on BAMI 2, and 4.8 ± 3.3 BPM on PPG DaLiA. Notably, it demonstrated statistically significant improvement in MAE for the BAMI 2 dataset compared to the base model in Chapter 6.

Based on these findings and as discussed in section 7.4.2, Concrete Dropout emerges as the most effective method for uncertainty quantification, offering a balance between well-calibrated uncertainty estimates and enhanced predictive accuracy. This analysis highlights the importance of method selection for accurate heart rate estimation, influenced by the characteristics of the dataset and the specific performance of each method.

7.5 Uncertainty-Aware Post-processing

Post-processing is key in enhancing the accuracy of heart rate predictions made by conventional and deep learning PPG techniques [95]. This process involves ensuring that the predicted heart rate values are plausible based on the context provided by preceding predictions. To achieve this, post-processing methods work by rejecting or

modifying predictions deemed infeasible. In practical applications, it is important to note that post-processing cannot be individually calibrated for each subject. Additionally, this process is limited to considering only the predictions made before the current one, without the ability to anticipate future values.

Various post-processing approaches have been utilised, ranging from history tracking and thresholding as well as Viterbi decoding and Finite State Machines [95]. For instance, the SpaMA algorithm utilises a threshold-based scheme. This approach operates under the assumption that heart rate values are unlikely to vary more than 10 BPM within a 2-second interval, comparing the previous prediction to the current one [161].

This section presents a comparative analysis of two distinct thresholding techniques for post-processing in heart rate prediction: prediction-based and uncertainty-based methods. Both approaches aim to enhance the reliability of heart rate predictions by employing a rejection-based scheme. This scheme is designed to eliminate predictions deemed unfeasible or carry a high degree of uncertainty, yielding more reliable outcomes. To assess the effectiveness of these post-processing techniques, several key metrics will be utilised:

- Retention Rate: This metric represents the proportion of predictions that are retained after post-processing.
- Longest Prediction Gap: This is defined as the maximum duration for no 'valid' prediction.
- Retained Predictions Above AAMI Standard: This measures the percentage of predictions, deemed 'valid' by post-processing, yet were above the AAMI Standard.
- Removed Predictions Below AAMI Standard: This measures the percentage of predictions, deemed 'invalid' by post-processing, yet were below the AAMI Standard.

The prediction-based method is similar in principle to the SpaMA algorithm [161]. It operates on the premise that heart rate values are unlikely to experience fluctuations greater than 10 BPM within a 2-second interval [161]. Under this method, if a prediction deviates by more than 10 BPM from the preceding one, it is discarded. If the following prediction shows a deviation greater than 20 BPM from the last valid prediction, it, too, is rejected. This threshold incrementally increases by 10 BPM for each subsequent prediction until a valid prediction is identified.

Table 7.3 highlights the impact of the prediction-based post-processing scheme in improving accuracy across various datasets. For instance, the IEEE Test dataset showed a reduction in mean absolute error from 15.0 ± 11.8 BPM to 12.0 ± 11.1 BPM, and the

Dataset	Original MAE (BPM)	Post processed MAE (BPM)	Retention Rate (%)	Longest Prediction Gap (Seconds)	Retained Predictions Above AAMI Standard (%)	Removed Predictions Below AAMI Standard (%)
IEEE Train [126]	5.8 ± 7.4	5.1 ± 7.3	93.7 ± 8.6	6.0 ± 5.3	21.2 ± 22.8	75.6 ± 22.9
IEEE Test [126]	15.0 ± 11.8	12.0 ± 11.1	76.0 ± 14.6	9.6 ± 3.6	45.8 ± 26.9	49.0 ± 25.8
BAMI 1 [127]	3.6 ± 2.1	3.1 ± 1.7	96.8 ± 3.5	5.3 ± 2.9	14.1 ± 10.0	84.2 ± 10.9
BAMI 2 [127]	1.6 ± 0.6	1.5 ± 0.6	99.0 ± 1.1	2.9 ± 2.5	4.5 ± 4.6	94.8 ± 4.6
PPG DaLiA [130]	4.9 ± 3.2	3.9 ± 2.4	92.8 ± 5.0	9.5 ± 2.7	17.3 ± 10.6	79.7 ± 12.3
MW PPG HR (This Work)	7.5 ± 3.2	6.1 ± 2.7	91.6 ± 4.0	11.2 ± 3.7	34.4 ± 11.6	62.4 ± 12.0

All Values are Averages over Subjects, using LOSO CV.

TABLE 7.3: Comparative Evaluation of Prediction-based Post-processing Method across Datasets. This table shows the performance of the prediction-based post-processing method when applied to various PPG heart rate datasets. Key metrics include original/post-processed MAE, retention rate, prediction gap, and AAMI standard compliance. The results demonstrate the method's effectiveness in improving accuracy and reliability, with varying impact across diverse datasets.

MW PPG HR dataset saw a decrease from 7.5 ± 3.2 BPM to 6.1 ± 2.7 BPM. These improvements are reflected across other metrics, though the extent of the enhancement varies by dataset. Notably, the IEEE Test dataset was the only one with an average retention rate below 90%, indicating that a larger proportion of predictions were considered unfeasible or unreliable in this dataset. Additionally, the MW PPG HR dataset had the longest average prediction gap of 11.2 seconds, signifying extended periods without valid predictions. A concerning trend observed across all datasets is the high percentage of discarded predictions falling below the AAMI standard, with at least 50% of removed predictions below this benchmark in every dataset. This suggests that while the post-processing scheme enhances accuracy, it may also be overly conservative, potentially excluding a substantial number of accurate predictions.

Table 7.4 provides an in-depth analysis of how a prediction-based post-processing scheme affects the fairness of heart rate predictions across different demographic groups, focusing on skin melanin content and biological sex. For skin melanin content, the MAE before post-processing was 9.0 ± 3.1 BPM for individuals with high melanin and 6.0 ± 2.4 BPM for those with low melanin. After post-processing, these values improved to 7.3 ± 2.8 BPM for high melanin and 5.0 ± 2.1 BPM for low melanin. Regarding biological sex, the MAE before post-processing was 8.1 ± 3.2 BPM for males and 6.4 ± 2.8 BPM for females. After post-processing, the MAE decreased to 6.6 ± 2.8 BPM for males and 5.3 ±

Demographic	Original MAE (BPM)	P Value	Post processed MAE (BPM)	P Value	Retention Rate (%)	Longest Prediction Gap per Subject (Seconds)	Retained Predictions Above AAMI Standard (%)	Removed Predictions Below AAMI Standard (%)
Skin Melanin: High	9.0 ± 3.1	0.045	7.3 ± 2.8	0.064	89.6 ± 3.2	13.0 ± 3.5	37.9 ± 11.6	58.0 ± 11.2
Skin Melanin: Low	6.0 ± 2.4		5.0 ± 2.1		93.6 ± 3.7	9.4 ± 3.0	30.9 ± 10.6	66.8 ± 11.2
Biological Sex: Female	8.1 ± 3.2	0.351	6.6 ± 2.8	0.485	90.6 ± 3.7	11.1 ± 3.6	35.5 ± 11.3	60.9 ± 11.4
Biological Sex: Male	6.4 ± 2.8		5.3 ± 2.2		93.4 ± 3.9	11.4 ± 4.0	32.3 ± 12.0	65.2 ± 12.7

All Values are Averages over Subjects, using LOSO CV. Statistical Tests used the Mann-Whitney U test.

TABLE 7.4: Comparative Performance of Prediction-based Post-processing Method across Demographic Groups using MW PPG HR Dataset (This Work). This table evaluates the impact of the prediction-based post-processing method on heart rate prediction accuracy and reliability across different demographic characteristics, namely skin melanin content and biological sex. The analysis includes original/post-processed MAE, retention rate, prediction gap, and AAMI standard compliance, along with statistical significance testing. The results provide insights into the method's performance and potential biases across diverse user groups.

2.2 BPM for females. These results indicate that the post-processing scheme improves prediction accuracy for both skin melanin content and biological sex, enhancing fairness across these demographic groups.

The uncertainty-aware post-processing scheme uses a threshold-based approach, similar to existing methods, and relies on Dropout's well-calibrated uncertainty quantification, as detailed in section 7.4.2. It focuses on predictions with high uncertainty, rejecting those with an uncertainty estimate over 10 BPM. This threshold is chosen because the SpaMA algorithm also uses a 10 BPM threshold, based on the idea that heart rate changes are unlikely to exceed this amount within short periods [161]. By applying this threshold, the scheme aims to filter out potentially unreliable predictions, while considering both aleatoric and epistemic uncertainties, as well as their combined effect.

Table 7.5 provides a comprehensive analysis of how the incorporation of different types of uncertainty into post-processing impacts the accuracy of heart rate predictions across various datasets. It is observed that all types of uncertainty - aleatoric, epistemic (via concrete dropout), and predictive (combining aleatoric and epistemic) - enhance accuracy, but the degree of improvement varies. Aleatoric uncertainty, associated

Dataset	Uncertainty Type	Original MAE (BPM)	Post processed MAE (BPM)	Retention Rate (%)	Longest Prediction Gap per Subject (Seconds)	Retained Predictions Above AAMI Standard (%)	Removed Predictions Below AAMI Standard (%)
IEEE Train [126]	Aleatoric	5.7 ± 7.6	5.6 ± 8.0	91.3 ± 20.2	4.9 ± 10.6	22.0 ± 23.6	8.7 ± 19.3
	Epistemic		3.9 ± 3.8	90.8 ± 17.8	8.5 ± 13.6	20.2 ± 20.0	16.4 ± 27.8
	Predictive		1.3 ± 0.6	52.8 ± 27.7	39.4 ± 37.6	2.8 ± 4.1	61.6 ± 26.7
IEEE Test [126]	Aleatoric	14.7 ± 10.6	14.5 ± 10.4	96.8 ± 5.3	3.8 ± 5.9	57.5 ± 27.1	8.9 ± 18.2
	Epistemic		12.5 ± 10.3	81.0 ± 11.5	14.8 ± 12.3	52.3 ± 27.0	11.5 ± 13.2
	Predictive		6.6 ± 8.3	22.4 ± 29.2	59.4 ± 61.6	24.8 ± 25.8	31.7 ± 21.0
BAMI 1 [127]	Aleatoric	3.5 ± 2.1	3.5 ± 2.1	100.0 ± 0.0	0.0 ± 0.0	15.9 ± 11.9	0.0 ± 0.0
	Epistemic		2.9 ± 1.7	96.4 ± 4.4	7.1 ± 5.9	13.4 ± 10.3	13.9 ± 23.6
	Predictive		2.3 ± 1.1	87.6 ± 11.5	26.6 ± 24.3	8.8 ± 7.3	37.5 ± 21.1
BAMI 2 [127]	Aleatoric	1.8 ± 1.2	1.8 ± 1.2	100.0 ± 0.0	0.0 ± 0.0	6.0 ± 6.4	0.0 ± 0.0
	Epistemic		1.6 ± 0.9	99.1 ± 1.1	3.4 ± 3.7	5.3 ± 6.0	16.6 ± 31.1
	Predictive		1.2 ± 0.4	95.6 ± 4.5	12.5 ± 16.8	3.2 ± 3.6	33.0 ± 26.4
PPG DaLiA [130]	Aleatoric	4.9 ± 3.1	4.9 ± 3.0	99.9 ± 0.0	0.4 ± 1.5	22.5 ± 13.0	0.0 ± 0.0
	Epistemic		4.4 ± 2.5	97.0 ± 3.1	9.6 ± 6.2	20.7 ± 12.2	19.3 ± 7.8
	Predictive		2.5 ± 0.9	81.0 ± 12.9	80.1 ± 63.2	10.9 ± 6.3	29.1 ± 10.1
MW PPG HR (This Work)	Aleatoric	7.3 ± 2.5	7.2 ± 2.5	99.6 ± 1.6	4.3 ± 16.1	38.9 ± 11.6	6.9 ± 18.8
	Epistemic		5.8 ± 2.0	91.5 ± 5.4	34.4 ± 27.8	34.7 ± 10.8	10.4 ± 7.2
	Predictive		3.3 ± 1.1	60.9 ± 11.4	183.5 ± 116.5	19.1 ± 7.2	29.5 ± 11.4

All Values are Averages over Subjects, using LOSO CV.

TABLE 7.5: Evaluation of Uncertainty-Aware Post-processing Across Datasets and Uncertainty Types. This table provides a comprehensive assessment of the uncertainty-aware post-processing method for heart rate estimation. It compares metrics including original and post-processed Mean Absolute Error (MAE), retention rate, prediction gap, and AAMI standard compliance across three uncertainty types: aleatoric, epistemic (via concrete dropout), and predictive (combining aleatoric and epistemic). The results show consistent MAE reduction with post-processing, achieving the lowest MAE under predictive uncertainty, though this is associated with the lowest retention rate, highlighting an area for improvement.

with the inherent variability in the data, results in the smallest reductions in MAE across all datasets. In contrast, predictive uncertainty, which combines both aleatoric and epistemic uncertainties, achieves the largest reductions in MAE. This suggests that considering the complete spectrum of uncertainty yields more accurate heart rate predictions.

Despite these improvements in accuracy, the retention rate for predictions processed using the predictive uncertainty method shows a significant decrease. The average retention rates vary considerably, ranging from as low as 22.4% for the IEEE Test dataset to as high as 95.6% for the BAMI 2 dataset. This wide range indicates a notable trade-off between accuracy and the frequency of valid predictions. Additionally, the duration of the longest period without a valid prediction varies greatly across datasets, extending from 12.5 seconds in BAMI 2 to 3 minutes in the MW PPG HR dataset (This work). This variation highlights the trade-off between ensuring accuracy and maintaining a consistent frequency of valid predictions. In healthcare, accuracy can justify occasional prediction gaps, but further improvements are needed to reduce these gaps and enhance retention rates.

A key insight from the analysis is the performance of the uncertainty method concerning the AAMI standard. Both the percentage of retained predictions above the AAMI standard and the percentage of removed predictions below this standard are lower compared to the standard post-processing method. This indicates that the uncertainty-based method is more effective in identifying and discarding inaccurate predictions, further reinforcing its utility in enhancing the reliability of heart rate prediction systems.

The implementation of uncertainty-based post-processing marks a significant advancement in equitable heart rate prediction, as illustrated by the findings in Table 7.6. This approach successfully addresses and mitigates the biases evident in initial predictions and those processed by standard prediction-based methods. Specifically, for skin melanin content, initial model predictions exhibited a statistically significant difference in heart rate predictions ($p=0.045$), which the uncertainty-based post-processing method effectively neutralised ($p=0.910$), eliminating the disparity. In terms of biological sex, while the initial predictions showed no significant difference, the MAE before post-processing was 7.8 ± 2.1 BPM for males and 6.3 ± 2.8 BPM for females. After post-processing, the MAE decreased to 3.5 ± 1.1 BPM for males and 3.0 ± 1.0 BPM for females. These outcomes demonstrate that uncertainty-based post-processing not only consolidates the enhancements achieved by prediction-based methods but also delivers a superior level of fairness in heart rate prediction, offering more consistency and unbiased results across diverse demographic profiles.

Demographic	Original MAE (BPM)	P value	Post processed MAE (BPM)	P value	Retention Rate (%)	Longest Prediction Gap (Seconds)	Retained Predictions Above AAMI Standard (%)	Removed Predictions Below AAMI Standard (%)
Skin Melanin: High	8.4 ± 2.1	0.045	3.3 ± 0.9	0.910	57.7 ± 9.1	214.8 ± 109.3	18.6 ± 5.9	24.6 ± 7.2
Skin Melanin: Low	6.1 ± 2.2		3.3 ± 1.3		64.2 ± 12.5	152.2 ± 115.1	19.5 ± 8.2	34.3 ± 12.8
Biological Sex: Female	7.8 ± 2.1	0.183	3.5 ± 1.1	0.536	59.7 ± 10.6	176.0 ± 114.1	20.0 ± 7.0	27.5 ± 8.2
Biological Sex: Male	6.3 ± 2.8		3.0 ± 1.0		63.4 ± 12.4	197.4 ± 119.7	17.4 ± 7.2	33.1 ± 15.1

All Values are Averages over Subjects, using LOSO CV. Statistical Tests used the Mann-Whitney U test.

TABLE 7.6: Comparative Performance of Uncertainty-aware Post-processing Method across Demographic Groups using MW PPG HR Dataset (This Work). This table evaluates the impact of the uncertainty-aware post-processing method on heart rate prediction accuracy and reliability across different demographic groups, focusing on skin melanin content and biological sex. It compares original and post-processed MAE, retention rate, prediction gap, and AAMI standard compliance, with statistical significance testing. The results reveal that post-processing eliminated the statistically significant difference between high and low skin melanin content, with lower MAE values observed across all demographic groups.

In summary, both prediction-based and uncertainty-aware post-processing methods significantly enhance heart rate prediction accuracy and fairness. The prediction-based method, inspired by the SpaMA algorithm, reduces MAE across various datasets, though it can also discard many valid predictions. The uncertainty-aware approach; integrating aleatoric, and epistemic, predictive uncertainty achieves the greatest accuracy improvements but at the cost of lower prediction retention rates. It also enhances fairness by reducing biases related to skin melanin content and biological sex, offering more consistent and unbiased results. Overall, these methods provide valuable advancements in both the reliability and fairness of wrist-worn PPG heart rate estimation methods.

7.6 Comparison with Existing Deep Learning PPG Heart Rate Estimation Methods

This section presents a comprehensive comparison of various heart rate prediction methods, including the proposed method in its various versions, against existing approaches.

The comparison is based on performance metrics across different datasets, as shown in Table 7.7. For comparable results, methods that did not employ a LOSO CV scheme were excluded. BeliefPPG [27] emerges as a particularly notable method in this comparison. It achieves the highest accuracy across all datasets with the smallest number of parameters. This efficiency makes BeliefPPG an ideal candidate for deployment on edge devices, such as wrist-worn smartwatches, where computational resources are limited.

Method	**Version**	**# Params**	**IEEE Train** [126]	**IEEE Test** [126]	**BAMI 1** [127]	**BAMI 2** [127]	**PPG DaLiA** [130]	**MW PPG HR (This Work)**
CorNET [164]	Standard	250,000	4.7 ± 3.7	6.6 ± 5.4	—	—	—	—
DeepPPG [130]	Average	8,500,000	—	—	—	—	8.8 ± 3.8	—
DeepPPG [130]	Ensemble	60,000,000	4.0 ± 5.4	16.5 ± 16.1	—	—	7.7 ± 4.2	—
DeepPPG [130]	Constrained	26,000	—	—	—	—	10.0 ± 5.9	—
Binary CorNET [167]	Standard	250,000	6.2 ± 5.0	7.2 ± 6.1	—	—	—	—
Binary CorNET [167]	RTL	250,000	6.8 ± 5.3	8.0 ± 6.0	—	—		—
Wilkosz et al. [177]	Standard	60,000,000	—	—	—	—	6.3 ± 3.5	—
Kasnesis et al. [175]	Standard	132,000	5.0	16.5	—	—	4.4	—
Kasnesis et al. [175]	Post processing	132,000	4.4	13.5	—	—	4.0	—
BeliefPPG 2023 [27]	Standard	138,000	1.8 ± 0.8	3.8 ± 2.2	**2.0 ± 1.0**	1.5 ± 0.9	3.6 ± 1.4	—
BeliefPPG 2023 [27]	Viterbi	138,000	1.5 ± 0.6	**3.1 ± 1.9**	2.1 ± 1.0	1.5 ± 0.3	3.2 ± 1.3	—
Proposed Method	Standard	730,000	5.8 ± 7.4	15.0 ± 11.8	3.6 ± 2.1	1.6 ± 0.6	4.9 ± 3.2	7.5 ± 3.2
Proposed Method	Standard + Post processing	730,000	5.1 ± 7.3	12.0 ± 11.1	3.1 ± 1.7	1.5 ± 0.6	3.9 ± 2.4	6.1 ± 2.7
Proposed Method	Concrete Dropout	730,000	5.7 ± 7.6	14.7 ± 10.6	3.5 ± 2.1	1.8 ± 1.2	4.9 ± 3.1	7.3 ± 2.5
Proposed Method	Concrete Dropout + Uncertainty Aware Post Processing	730,000	**1.3 ± 0.6**	6.6 ± 8.3	2.3 ± 1.1	**1.2 ± 0.4**	**2.5 ± 0.9**	**3.3 ± 1.1**

All Values are MAE in BPM.

TABLE 7.7: Comparison of Heart Rate Estimation Performance with Existing Deep Learning Methods that used LOSO Cross Validation on All Utilised Dataset. **Bold** values indicate the lowest MAE distribution.

The proposed method, which integrates uncertainty-aware post-processing, demonstrates outstanding accuracy across various datasets. It achieves the lowest MAE compared to other deep learning methods using leave-one-subject-out cross-validation (LOSO CV). Specifically, it records MAE values of 1.3 ± 0.6 BPM on the IEEE Train dataset, 1.2 ± 0.4 BPM on BAMI 2, 2.5 ± 0.9 BPM on PPG DaLiA, and 3.3 ± 1.1 BPM on MW PPG HR. However, the method performs less effectively on the IEEE Test and

BAMI 2 datasets, with MAE values of 6.6 ± 8.3 BPM and 2.3 ± 1.1 BPM, respectively. In comparison, BeliefPPG achieved slightly higher accuracy with MAE values of 3.1 ± 1.9 BPM and 2.0 ± 1.0 BPM on these datasets.

It is important to note that while existing methods report accuracies based on all available samples, the proposed method selectively rejects uncertain samples during post-processing. This selective approach means that not all predictions are evaluated, as the method aims to enhance the reliability of the retained predictions. By discarding predictions with high uncertainty, the method improves the overall accuracy of the predictions that are kept, thus prioritising the reliability of heart rate estimates. In summary, this comparative analysis highlights the strengths of the proposed method, especially with its uncertainty-aware post-processing. It underscores the method's capability to deliver reliability, fairness and accuracy in heart rate predictions.

7.7 Summary

This chapter provides a comprehensive analysis of uncertainty quantification in wrist-worn PPG heart rate estimation deep learning methods and post processing methods answering both research questions 4: *What are the most effective methods for estimating uncertainty in deep learning methods for wrist-worn PPG heart rate estimation?* and 5: *How does incorporating uncertainty in post-processing improve the reliability of wrist-worn PPG heart rate estimation methodology?*

Addressing objectives 6 and 7, the chapter evaluated aleatoric uncertainty quantification across various datasets, demographic groups, and activities. The findings revealed significant differences in aleatoric uncertainty distributions by biological sex during active rest, rest, and cycling, as well as by skin melanin content during running. However, aleatoric uncertainty did not consistently correlate with signal quality alone, as demonstrated by a weak correlation with SNR of -0.32 in the MW PPG HR dataset. This inconsistency was particularly evident when random noise was added to the signals, which resulted in a much larger increase in epistemic uncertainty compared to aleatoric uncertainty. The observed interplay between these uncertainties highlights the complexity of accurately quantifying and interpreting uncertainty in wrist-worn PPG heart rate estimation methods, particularly in diverse populations and motion conditions.

Addressing objectives 6 and 7, the chapter also explored three epistemic uncertainty quantification methods: Monte Carlo dropout, Concrete dropout, and Ensemble. The analysis began by determining the optimal number of samples for both dropout techniques, finding that 25 samples provided the best trade-off between miscalibration area, MAE, and processing time, while the ensemble method was most effective with

5 samples. Concrete dropout emerged as the most effective method for producing well-calibrated uncertainty, achieving miscalibration areas of 0.05 on the IEEE Train dataset, 0.08 on BAMI 1, and 0.12 on MW PPG HR. Additionally, both concrete dropout and ensemble methods significantly improved estimation accuracy across all datasets. For instance, the base model described in Chapter 6 achieved a MAE of 3.6 ± 2.1 BPM on the BAMI 1 dataset, which was improved to 3.3 ± 2.1 BPM with concrete dropout; similarly, on the MW PPG HR dataset, the MAE improved from 7.5 ± 3.2 BPM to 7.3 ± 2.5 BPM. Concrete dropout showed a strong correlation with absolute error, with correlation coefficients of 0.78 on IEEE Train and 0.64 on both BAMI 1 and MW PPG HR. Contrary to expectations, a correlation between epistemic uncertainty and ECG-derived SNR was observed. The study also revealed significant differences in epistemic uncertainty distributions based on skin melanin content during active rest, running, rest, and cycling, and differences based on biological sex during running and cycling. The activity-based analysis of epistemic uncertainty showed unexpected trends, with rest and cycling having similar distributions, and active rest and running showing similar patterns, indicating that motion-based uncertainty was captured.

Addressing Objective 8, the chapter also examined two post-processing methods: a prediction-based approach and an uncertainty-based approach. The prediction-based method effectively reduced MAE across all datasets, such as improving MAE from 7.5 ± 3.2 BPM to 6.1 ± 2.7 BPM on the MW PPG HR dataset. However, the uncertainty-based method outperformed the prediction-based approach, particularly in enhancing the fairness of the model. For instance, before post-processing, the MAE for individuals with high skin melanin content was 8.4 ± 2.1 BPM compared to 6.1 ± 2.2 BPM for those with low melanin content—a statistically significant difference. After applying the uncertainty-based post-processing, the MAE was equalised to 3.3 ± 0.9 BPM for high melanin content and 3.3 ± 1.3 BPM for low melanin content, effectively eliminating the significant difference. Similar improvements were observed for biological sex. However, this approach led to lower prediction retention rates, highlighting a potential area for future enhancement.

Addressing Objective 9, the proposed method, which integrates uncertainty-aware post-processing, demonstrates competitive accuracy across various datasets. It achieves the lowest MAE compared to other deep learning methods using LOSO CV. Specifically, it records MAE values of 1.3 ± 0.6 BPM on the IEEE Train dataset, 1.2 ± 0.4 BPM on BAMI 2, 2.5 ± 0.9 BPM on PPG DaLiA, and 3.3 ± 1.1 BPM on MW PPG HR. However, the method performs less effectively on the IEEE Test and BAMI 2 datasets, with MAE values of 6.6 ± 8.3 BPM and 2.3 ± 1.1 BPM, respectively. In comparison, BeliefPPG achieved slightly higher accuracy with MAE values of 3.1 ± 1.9 BPM and 2.0 ± 1.0 BPM

on these datasets [27].

In conclusion, this chapter thoroughly explores uncertainty quantification in deep learning for wrist-worn PPG heart rate estimation, addressing key research questions and objectives. The analysis shows that aleatoric and epistemic uncertainties significantly affect model performance across demographic groups and activities, interacting in complex ways, especially with noise and signal quality variations. The post-processing methods examined highlight the potential of uncertainty-aware approaches to improve model fairness, particularly regarding skin melanin content and biological sex, though challenges in prediction retention remain. Overall, the proposed method demonstrates strong accuracy, laying a foundation for enhancing reliability and fairness in future wrist-worn PPG heart rate estimation methods.

Chapter 8

Conclusion

This chapter collates the key findings and contributions of the research carried out in this thesis on wrist-worn photoplethysmography (PPG) heart rate estimation. The chapter discusses the significance of the novel dataset, deep learning methods, and uncertainty quantification techniques, critically examining their implications for the field. The chapter analyses the limitations of the research and outlines promising directions for future work. This chapter aims to provide a comprehensive overview of the research outcomes and their potential impact on advancing wrist-worn PPG heart rate estimation methodology.

8.1 Discussion

This thesis has contributed to the field of wrist-worn PPG heart rate estimation by addressing several key challenges and advancing understanding of the method's capabilities and limitations. The thesis's comprehensive approach, from protocol design to deep learning implementation and uncertainty quantification, has yielded valuable insights that have practical implications.

One of the primary contributions of this research is the development of a novel, diverse dataset that addresses limitations in existing protocols. The inclusion of active rest and cycling phases, along with a balanced representation of biological sex and skin types, provides a robust foundation for evaluating PPG heart rate estimation methods. This dataset, featuring the largest representation of physical effort rates of 60% or higher, 26,442 samples with comprehensive multi-wavelength PPG data, fills a key gap in the field and enables more thorough validation of estimation methodology across various physiological states, demographics, and motion types. The dataset also includes challenging conditions like erratic wrist movements, cross-over effects, and motion-free periods, providing a robust basis for evaluating wrist-worn PPG heart rate estimation methods.

The analysis of signal quality indices (SQIs) revealed that the electrocardiogram (ECG)-derived signal-to-noise ratio (SNR) is the most reliable metric for assessing PPG signal quality in the context of wrist-worn PPG heart rate monitoring. This finding has important implications for future research and device development, as it provides a more accurate means of evaluating signal integrity, particularly in motion-rich conditions. The evaluation of PPG beat detectors across different activities, wavelengths, and demographics yielded valuable insights into the strengths and limitations of conventional methods. The observed variations in detector accuracy across different scenarios underscore the importance of considering contextual factors in PPG signal processing and highlight areas for potential improvement in beat detection algorithms.

The deep learning method developed for heart rate estimation demonstrated the potential of multi-wavelength approaches, with the blue-green-red-IR combination showing particular promise. This combination reduced MAE by 0.4 BPM compared to green light alone and improved accuracy by 1.3 BPM during motion-based activities like running. However, the thesis also revealed persistent challenges related to demographic factors, particularly skin melanin content and biological sex. The observed increase in error rates for individuals with higher melanin content (MAE of 8.4 ± 2.1 BPM) compared to those with lower melanin content (MAE of 6.1 ± 2.2 BPM) and for females highlights the need for more inclusive design and validation of PPG systems. These findings underscore the importance of diversity in research cohorts and the need for both fairness quantification and mitigation methods in algorithm development.

The exploration of uncertainty quantification in deep learning methods for wrist-worn PPG heart rate estimation represents a significant advancement in the field. The comparison of aleatoric and epistemic uncertainty quantification methods provides valuable insights into the sources and nature of estimation errors. The finding that Concrete dropout produced the best-calibrated epistemic uncertainty estimates offers a promising direction for improving the reliability of wrist-worn PPG heart rate estimation methodology. Concrete dropout was found to correlate strongly with absolute error and ECG-derived SNR across datasets, enhancing the method's reliability in variable conditions.

The post-processing methods examined, particularly the uncertainty-based approach, demonstrated potential for enhancing both the accuracy and fairness of heart rate estimation models. The ability to equalise performance across different skin melanin levels and biological sexes is a key step towards a more fair wrist-worn PPG heart rate estimation methodology. The proposed uncertainty-aware post-processing method achieved low MAE values across multiple datasets, including 3.3 ± 1.1 BPM on the newly collected dataset. However, the trade-off with heart rate estimation retention

rates highlights an area requiring further research and optimisation. When compared to existing methods on various datasets, the proposed approach showed competitive performance, particularly on larger datasets. This suggests that the method's strength lies in its ability to handle complex, real-world data with varying conditions and demographics.

In conclusion, this thesis has not only addressed key challenges in wrist-worn PPG heart rate estimation but has also paved the way for future advancements in this area. By developing a novel, diverse dataset and integrating deep learning techniques with uncertainty quantification, the thesis has provided both a robust framework for evaluating existing methods and a foundation for future innovations. The insights gained, particularly regarding the role of multi-wavelength approaches, demographic factors, and post-processing methods, underscore the importance of inclusive and context-aware designs in improving the reliability and fairness of wrist-worn PPG heart rate monitoring systems. As the field continues to evolve, the contributions of this thesis hope to serve as a valuable resource, guiding the development of more reliable, inclusive, and effective wrist-worn PPG heart rate estimation methods

8.2 Limitations and Future Research

This section critically examines the limitations of the research and outlines potential areas for future research. It acknowledges the constraints in the data, the limitations of the predictive models used, and the challenges in uncertainty quantification. While highlighting these areas, the section also suggest directions for future research to enhance data, refine model accuracy, and improve methods in uncertainty analysis, thereby contributing to the advancement of the field.

8.2.1 Data

A significant limitation of the study stems from the size of the dataset, particularly in terms of participant numbers. The recommended threshold of at least 60 participants, as suggested by Colvonen et al. for accounting variations across Fitzpatrick skin types, was not met [191]. This shortfall impacted the robustness of the conclusions, especially regarding the effects of skin tone and biological sex. While the introduction of the skin melanin content aspect partially mitigated this issue, a larger participant pool would have undoubtedly lent more concrete validity to the findings. Additionally, the diversity in the number and types of activities recorded was limited, potentially influencing the generalisability of the results.

Incorporating orange wavelengths (590-595 nm) could offer several advantages [50,114]. Orange light penetrates skin deeper than green light and is less affected by melanin absorption than shorter wavelengths [50,114]. This could potentially improve signal quality for darker skin tones and provide more resilience to motion artefacts.

Exploring multi-site sensor arrays on the wrist could provide valuable insights into optimal sensor configurations [80–82]. This approach, involving multiple PPG sensors positioned around the wrist's circumference, could help collect more accurate signals by accounting for variations in blood flow, muscle tissue, and subcutaneous fat distribution [80–82]. A multi-site array would adapt to individual anatomy and movement, simultaneously explore multiple measurement sites, and compensate for activity-related fluctuations in signal quality.

Additionally, investigating the impact of temperature variations on PPG signal quality is key [62,65,66]. Skin temperature affects blood flow which can diminish PPG signal quality [62,65,66]. Understanding how temperature changes affect wrist-worn PPG heart rate estimations during different activities and demographically variations could inform more robust methodology.

8.2.2 Deep Learning Heart Rate Estimation Method

The thesis's deep learning heart rate estimation method, while demonstrating competitive accuracy, has room for improvement. Future work could enhance the current architectural framework by introducing individual input branches for each PPG wavelength or by grouping short and long wavelengths into separate branches. Advanced feature extraction techniques, such as incorporating frequency domain features like FFT or wavelet transforms, and adding attention mechanisms within branches, could also be explored. Improved fusion strategies, such as weighted fusion of branch outputs or using learnable fusion layers like 1x1 convolutions, may further refine performance. Additionally, adapting the model to multi-task learning — such as predicting multiple physiological parameters simultaneously or incorporating auxiliary tasks like signal quality assessment and activity recognition — could broaden its capabilities.

Beyond refining the current architecture, exploring other architectures like WaveNet [240], Inception Blocks [241], Transformers [242, 243], Informer [244], and ResNets [245] could yield significant benefits. WaveNet's dilated causal convolutions capture long-range temporal dependencies [240], Inception Blocks facilitate multi-scale feature extraction [241], Transformers/Informer manage global dependencies in sequences with self-attention mechanisms [242,244], and ResNets, with deep architecture and skip connections, mitigate the vanishing gradient problem, enabling the training of very

deep networks [245]. For example, Zerveas et al. evaluated some of these methods on the IEEE datasets, achieving root mean square error of 23.9 BPM with an Inception network, 33.2 BPM with ResNet, and 25.0 BPM with a time series transformer [242]. However, these networks were tested on diverse datasets without optimisation for PPG signals and without using LOSO CV.

Additionally, combining time and frequency representations, such as continuous wavelet transform, could enhance signal representation [27]. Addressing the model's large parameter count through model compression techniques, such as quantisation, knowledge distillation, and model binarisation, would improve its suitability for deployment on edge devices, like wrist-worn wearables. Expanding data augmentation methods and exploring advanced pre-processing techniques, including multi-wavelength noise reduction, could further enhance heart rate estimation accuracy. Implementing a transfer learning approach, similar to the works of Davies et al., Naeini et al., and Meng et al. [243,246,247], where a model is pre-trained on all other wrist-worn PPG heart rate datasets and then fine-tuned on a specific dataset (excluded from pre-training), could also improve generalisability and performance.

8.2.3 Uncertainty Quantification

Future research should explore deterministic uncertainty quantification methods to enhance efficiency in wearable devices, moving away from sampling-based approaches that require multiple forward passes of the data. One promising method is 'Evidential Deep Regression', which deterministically predicts a Normal-Inverse-Gamma distribution, allowing for single-pass uncertainty estimation that captures both aleatoric and epistemic uncertainties [248]. Additionally, methods like 'Zig Zag' could also be considered for their ability to quantify these uncertainties deterministically [249].

An alternative approach could involve replacing aleatoric uncertainty quantification by integrating signal quality assessment with downstream tasks as detailed in '*The 2023 wearable photoplethysmography roadmap*' [113]. This could entail developing a supplementary deep learning model to predict ECG-derived SNR (dB), which would then serve as an input to the heart rate estimation model. By providing a direct measure of signal quality, this method could enhance the overall method.

Further, advancing uncertainty-aware post-processing techniques, such as adaptive or activity-aware thresholding, could improve the retention rate of valid heart rate estimations while effectively filtering out inaccurate ones, thereby increasing the model's reliability and applicability in real-world scenarios.

8.3 Summary

This chapter has collated the key contributions of the research on wrist-worn PPG heart rate estimation, highlighting the development of a novel, diverse dataset as well as the exploration of multi-wavelength deep learning methods and uncertainty quantification techniques. The research addressed critical challenges, such as demographic disparities and signal quality issues, providing a more comprehensive and inclusive approach to heart rate estimation. Despite limitations in dataset size and model complexity, the findings offer valuable insights for future work, emphasising the need for more diverse participant pools, refined model architectures, and more efficient uncertainty quantification methods. These contributions lay a solid foundation for further advancements in the reliability, fairness, and accuracy of wrist-worn PPG heart rate estimation systems.

Appendix A

Publications

A.1 A Review of Wearable Multi-Wavelength Photoplethysmography

Ray, D., Collins, T., Woolley, S., & Ponnapalli, P. (2023). A Review of Wearable Multi-Wavelength Photoplethysmography. IEEE Reviews in Biomedical Engineering, 16, 136–151. https://doi.org/10.1109/RBME.2021.3121476

A Review of Wearable Multi-wavelength Photoplethysmography

Daniel Ray *(Student Member, IEEE)*, Tim Collins *(Senior Member, IEEE)*, Sandra I. Woolley *(Senior Member, IEEE)*, and Prasad V. S. Ponnapalli *(Member, IEEE)*

***Abstract*— Optical pulse detection 'photoplethysmography' (PPG) provides a means of low cost and unobtrusive physiological monitoring that is popular in many wearable devices. However, the accuracy, robustness and generalizability of single-wavelength PPG sensing are sensitive to biological characteristics as well as sensor configuration and placement; this is significant given the increasing adoption of single-wavelength wrist-worn PPG devices in clinical studies and healthcare. Since different wavelengths interact with the skin to varying degrees, researchers have explored the use of multi-wavelength PPG to improve sensing accuracy, robustness and generalizability. This paper contributes a novel and comprehensive state-of-the-art review of wearable multi-wavelength PPG sensing, encompassing motion artifact reduction and estimation of physiological parameters. The paper also encompasses theoretical details about multi-wavelength PPG sensing and the effects of biological characteristics. The review findings highlight the promising developments in motion artifact reduction using multi-wavelength approaches, the effects of skin temperature on PPG sensing, the need for improved diversity in PPG sensing studies and the lack of studies that investigate the combined effects of factors. Recommendations are made for the standardization and completeness of reporting in terms of study design, sensing technology and participant characteristics.**

***Index Terms*— Multi-wavelength Photoplethysmography, Skin Optics, Skin Melanin, Skin Temperature, Motion Artifact Reduction, Physiological Monitoring.**

I. Introduction

THE current passive paradigm of late-stage treatment of chronic diseases is transitioning towards more proactive and preventative measures, such as cost-effective continuous monitoring tools to support early prediction, early prevention, early diagnosis and early treatment [1]. The World Health Organization (WHO) have recommended continuous monitoring as an effective means to assess physiological conditions, monitor the progression of diseases and support daily self-management of health [2].

Clinically performed electrocardiography (ECG), such as the conventional 12-lead ECG acquisition, is widely considered the 'gold standard' of non-invasive cardiovascular monitoring. ECG can identify cardiovascular abnormalities and detect irregularities in heart rhythms as well as performing physiological assessments of Heart Rate (HR) and Heart Rate Variability (HRV) by recording the depolarization of the heart's conductive pathway and the surrounding cardiac muscle tissues during each cardiac cycle. Although accurate, multi-lead clinical ECG is unsuited to continuous monitoring. It lacks portability, and convenience, and the bio-electrodes are obtrusive, cannot get wet and must be placed at specific locations on the body and connected to a recording device [3].

Consumer health monitoring devices have underpinned growth in the wearable devices market, a market expected to reach $30 billion by the end of 2023 [4], and the maturation of low cost, unobtrusive sensing devices incorporating optical pulse detection 'photoplethysmography' (PPG) sensors [3], [5]. A sensing method similar to PPG sensing was first devised in 1936 by two American research groups [6], but it was Alrick Hertzman who established PPG sensing in 1937 [7]. Consisting of a light source and photo-detector, light is emitted into the skin and the intensity of light transmitted into the photo-detector will vary depending on the volume of blood in the vascular bed of the measurement site, taking advantage of blood's absorbent qualities to visible and infrared (IR) light. During the contraction of the left ventricle, blood is ejected out of the heart and propagates along the circulatory system, corresponding to the initial positive slope of a PPG pulse (Figure 1). The systolic peak marks the maximum amount of blood present in the vascular bed at the measurement site. The pulse waveform then decreases in amplitude until a local minimum where it transitions into the diastolic phase. The local minimum or dicrotic notch marks the closure of the aortic valves. The end of the diastolic phase marks the closure mitral valve [8]. As well as the AC or pulsatile component of the signal, PPG sensing also collects the DC or non-pulsatile component which is shaped by respiration, sympathetic nervous system activity, and thermoregulation [3].

There are two modes of PPG sensing with different measurement sites (Figure 2). Transmission PPG sensors are usually sited on the fingertip or earlobe where the light source and detector are separated by tissue. Reflectance PPG sensors,

Manuscript received April 26, 2021; revised August 03, 2021; accepted September 22, 2021. This research received no external funding.

D. Ray is with the Department of Engineering, Manchester Metropolitan University, Manchester, UK (e-mail: Daniel.Ray@stu.mmu.ac.uk).

T. Collins is with the Department of Engineering, Manchester Metropolitan University, Manchester, UK (e-mail: T.Collins@mmu.ac.uk).

S. I. Woolley is with the School of Computing and Mathematics, Keele University, Staffordshire, UK (e-mail: S.I.Woolley@keele.ac.uk).

P. V. S. Ponnapalli is with the Department of Engineering, Manchester Metropolitan University, Manchester, UK (e-mail: P.Ponnapalli@mmu.ac.uk).

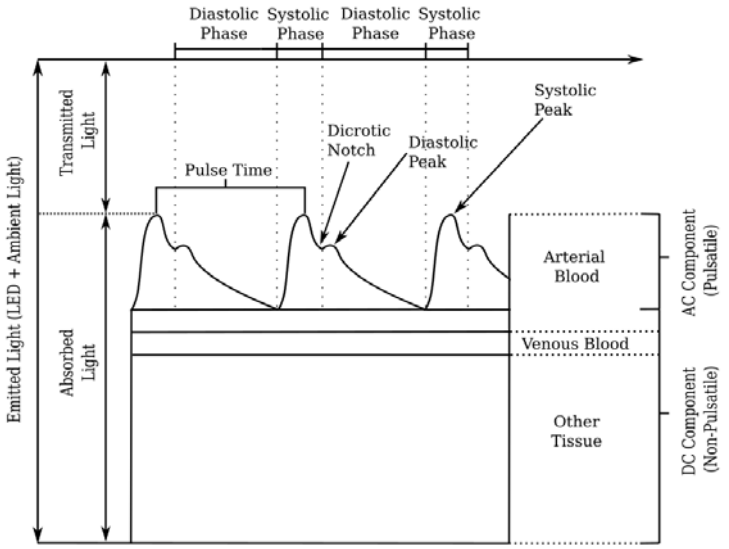


Fig. 1: A typical PPG waveform adapted from Lemay *et al.* [34].

which have both components positioned alongside each other on the same side of tissue, are commonly sited on the wrist, forehead or torso [3]. Both reflectance and transmission PPG sensing can provide physiological insights for HR [5], Blood Oxygen Saturation (SpO_2) [9], Respiration Rate [10], Vascular Aging and Atherosclerosis [3], [11], Blood Pressure (BP) [12] which is an indicator for Hypertension [13] and Atrial Fibrillation as well as HRV which itself provides indications about Coronary Heart Disease [14] and Autonomic Nervous System functionality [15].

Although transmission mode PPG sensing is widely used in clinical settings for pulse oximetry measurements, reflectance mode PPG and PPG sensing for other physiological measurements has not been widely adopted in clinical practice. One of the main factors affecting PPG sensing performance is its susceptibility to interference, predominantly from motion artifacts [16]. Other significant factors affecting the performance include the amount of blood flowing into the peripheral vascular bed, the varying optical properties of skin and blood, ambient light, and the wavelength used to illuminate the skin [5]. Addressing these factors would produce a low cost, simple and unobtrusive method to accurately, robustly and continuously measure the physiological status of patients having the potential to reduce premature mortality and the economic burden of disease and illness.

Beyond cardiovascular monitoring and general well-being, PPG sensing has seen several developments including the detection and monitoring of epileptic seizures [17], diagnosis of respiration diseases [18], mental stress and affect recognition [19], [20], monitoring of sleep conditions [21], [22], estimation of blood glucose [23], and drug delivery monitoring [24]–[26] showing its capacity to enhance healthcare systems around the world.

This review explores multi-wavelength PPG approaches for signal acquisition and improved resilience to motion artifacts and variations in skin melanin content and skin temperature with the aim of providing robust estimations of

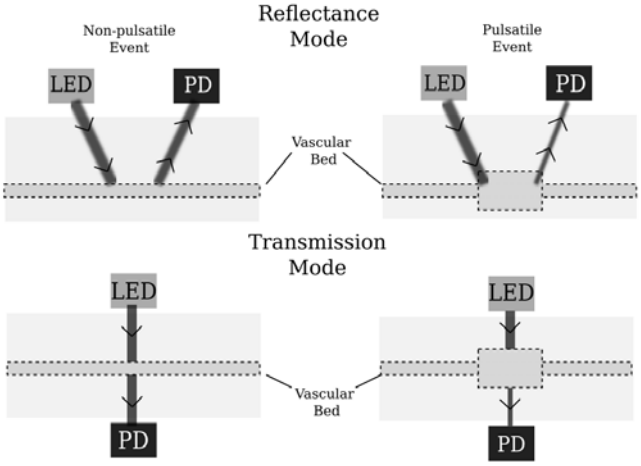


Fig. 2: The two modes of PPG sensing.

physiological parameters. In terms of methodology, the papers which constitute this review were retrieved from the following search engines and digital libraries: Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library, Google Scholar, Medline, ScienceDirect, and Wiley Online Library using the keywords: “Multi-wavelength”, “Photoplethysmography”, “PPG”, “Skin Tone”, “Skin Temperature”, “Skin Optics”, “Motion Artifacts”, “Physiological Monitoring”, “Blood Oxygen Saturation”, “Pulse Oximetry”, “Blood Pressure”, “Blood Glucose” and “Drug Delivery Monitoring”.

II. Skin Optics and PPG Sensing

Human skin is a complex heterogeneous medium consisting of three main layers: epidermis, dermis and hypodermis (or subcutaneous tissue), each with varying thicknesses dependent on body site that follow a general trend [27], [28]. The epidermis is the top-most layer of skin comprising of several sub-layers of living and non-living cells, all of which contain little or no blood flow. The stratum corneum or non-living

epidermis is typically 20μm thick and consists of only dead squamous cells [28]. Beneath is the living epidermis which is typically 100μm thick and contains most of the skin pigmentation compounds, such as pheomelanin and eumelanin, broadly referred to as melanin [27]–[30].

The dermis is found below the epidermis and consists of two main layers: the papillary dermis typically 150μm thick and the reticular dermis which has a thickness generally ranging from 1-4mm dependent on body site [28]. The papillary dermis is made up of loose connective tissue which is vascularized from a network of capillaries, small blood vessels typically ranging from 1-8μm in diameter [31] that exchange materials, such as oxygen and carbon dioxide, between blood and tissue. The reticular dermis is made up of dense connective tissue housing structures such as nerves, glands and hair follicles. Additionally, the reticular dermis contains arterioles and venules which are slightly larger blood vessels, typically ranging from 2-30μm in diameter [31], that connect the capillaries to the arteries and veins [27].

The deepest layer of the skin is the hypodermis which connects the skin to the bones and muscles and has a typical thickness in the range of 1-6mm dependent on body site [28]. The hypodermis contains larger blood vessels, arteries and veins, typically ranging from 500-5000μm in diameter [31], which transport blood around the body. The hypodermis is mainly used to store fat and primarily consists of loose connective tissue [27].

Due to the inhomogeneous distribution of blood, cells and pigments in the skin; measuring the optical properties is challenging. Usually, the main optical properties of skin are described as absorption, scattering and penetration depth as well as reflection, transmission and fluorescence [6], [28], [29], [32]–[36]. Researchers have employed several methods to model the optical properties of skin such as the radiative transport equation, the Beer-Lambert law, stochastic models such as the Monte Carlo and random walk as well as the adding doubling method with varying results [35], [37]. Summarized in Figure 3 are the wavelengths of light explored in this section and subsequent sections.

A. Optical Properties of Skin and Blood

The main light-absorbing components within the skin are water, hemoglobin and melanin; however, each absorb light differently depending on the wavelength of light and chemical bonding (Figure 4). Water, the main component of skin, is highly absorbent to IR light (900-1100nm) whilst possessing little to no absorbent qualities to visible light (390-780nm) [6], [28], [34], [36], [38]. Melanin protects the skin against the harmful ultraviolet (UV) radiation from the sun [29], its absorbing qualities increase as the wavelength of light decreases, being highly absorbent to shorter wavelengths ranging from UV to yellow light (200-600nm) [3], [6], [28]–[30], [32]–[34]. Similarly, hemoglobin's absorbing qualities decrease as the wavelength of light increases. However, when chemically bonded with oxygen, its absorbing qualities dramatically decrease when exposed to light in the range of 570-700nm and is more absorbent to longer wavelengths such as IR when compared to non-oxygenated hemoglobin [6], [28], [29], [32]–[34], [36], [38].

Scattering occurs as either a surface effect such as reflection and refraction or as an interaction with compounds in the skin that possess different optical properties. It has been estimated that 4-7% of light is reflected from the surface of the skin independent of wavelength [32]. Generally, within the skin the scattering coefficients decrease with an increase in the wavelength of light [28], [32], [33], [35], [36]. In the epidermis, large melanosomes exhibit mainly forward scattering whilst small "melanin dust" has an isotropic scattering profile. Collagen's fibrous structures define the scattering profile of the dermis whilst the main source of scattering in the hypodermis are spherical droplets of lipids [28]. Additionally, research suggests that the effects of scattering are greater on the breast, abdomen and forehead compared to the arm [36].

The path of light in the skin for reflectance mode PPG sensing is theorized to follow a "banana-like" shape [39] where the maximum depth of the path of light in the skin is called the penetration depth which is a function of its absorption and scattering coefficients [35]. In transmission mode PPG sensing, the path of light travels through the skin from the LED to the photodiode. Generally, the penetration depth for reflectance mode sensing increases as the wavelength of light increases in the range of visible and near-IR light (Figure 5) [3], [6], [28], [33], [34], [36], [38], [40]–[42] with the maximal penetration depth being 3-4mm for IR light (800-1100nm) [28], [36], [42], [43]. When the wavelength of light increases past 1250-1400nm the penetration depth significantly decreases [28], [36], [43]. Additionally, the penetration depth of light for reflectance mode sensing changes depending on the measurement site with the breast and abdomen possessing the largest penetration depths compared to the arm and forehead [36].

B. Effects of Biological Characteristics on PPG Sensing

As described in Section II(A), wavelengths of light interact with skin and blood to varying degrees due to their inhomogeneous nature. Researchers have explored the effects that different biological and skin characteristics have on PPG sensing, which is summarized in this section.

1) Skin Melanin Content: The accuracy and reliability of PPG sensing is sensitive to skin melanin content. First reports of potential inaccuracies arose from pulse oximetry studies which concluded that higher skin melanin content may influence the performance and accuracy [44]. Measurements of pulse oximetry on patients with higher skin melanin content and low blood oxygen saturation had up to 10% differences in estimates from different pulse oximeters [45] and blood oxygen saturation levels were over-estimated during hypoxia for patients with higher skin melanin content [46]. A larger scale study (1609 subjects: 1333 white patients and 276 black patients) also found black patients to have nearly three times the frequency of occult hypoxemia (an arterial oxygen saturation estimate of $<$88% despite an oxygen saturation of 92% - 96%) as white patients [47]. However, several other studies suggest that higher skin melanin content doesn't influence

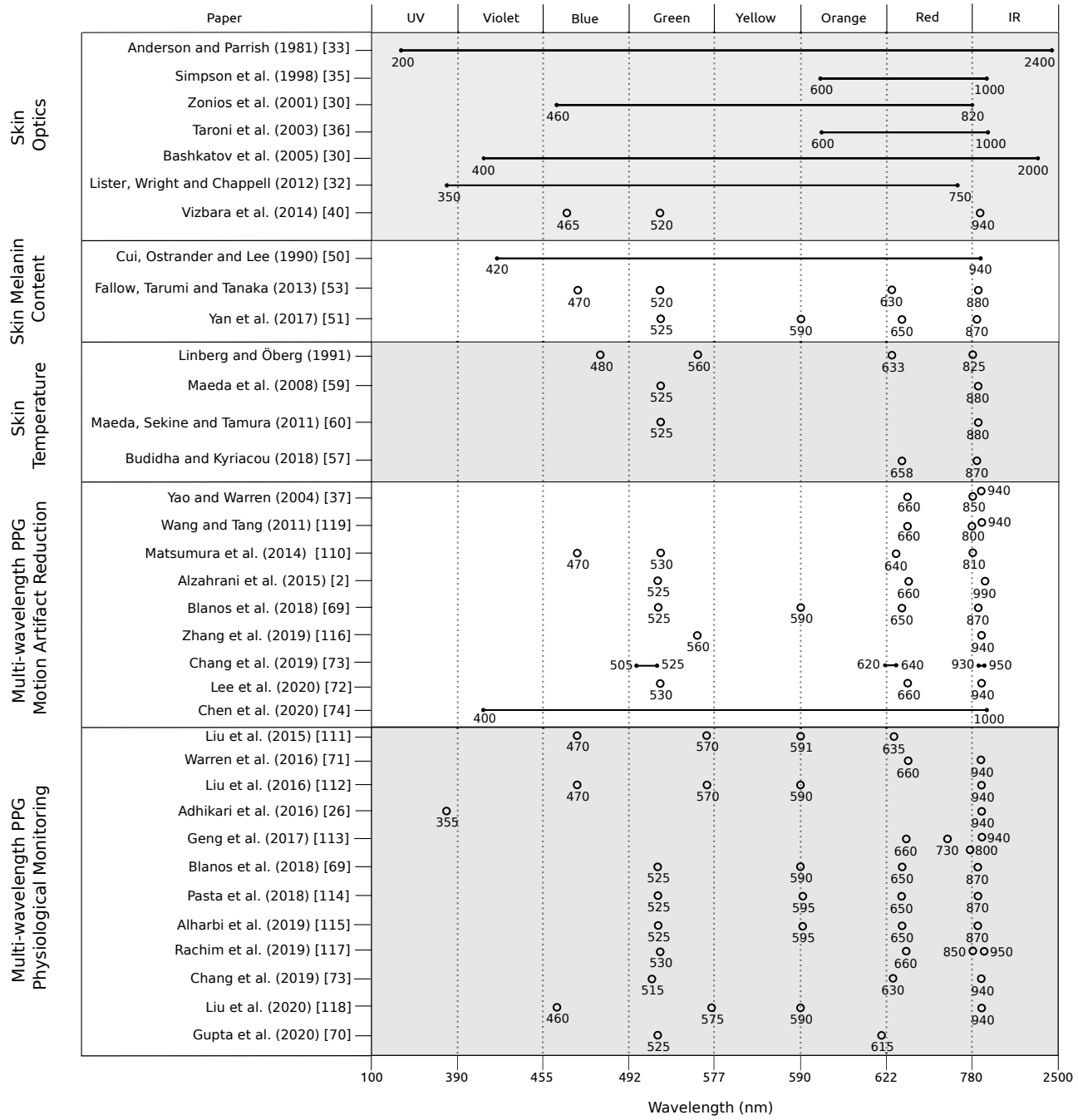


Fig. 3: A summary of the wavelengths of light explored for each multi-wavelength PPG study in each section of this review.

oximetry measurements [48], [49]. Most studies exploring higher skin melanin content and PPG sensing agree that melanin isn't a significant factor when at rest [48], [50]–[52]. The pulsatile component of the PPG signal is collected from the dermis and hypodermis. The epidermis, which contains melanin, absorbs a constant fraction of the signal without affecting the pulsatile component suggesting skin melanin content can be compensated by stronger intensity of light [50], [53]. During active states Bent *et al.* found no statistically significant differences in HR estimation accuracy across skin tones for commercially available wrist-worn reflectance mode PPG devices [52]. Fallow *et al.* found no significant interaction between skin tone and motion type for wrist-worn reflectance mode PPG sensing but did find a trend towards skin type having a significant effect yet no interaction was present at rest [53]. Yan *et al.* found skin melanin content to not have a significant factor on palm-worn reflectance mode PPG HR estimations as well as green light (525nm) producing the best modulation for all skin tones [51]. This agrees with Fallow *et al.* who found green light (520nm) to produce the best modulation for all skin tones at rest although blue (470nm) and green (520nm) light produced the best modulation for all skin

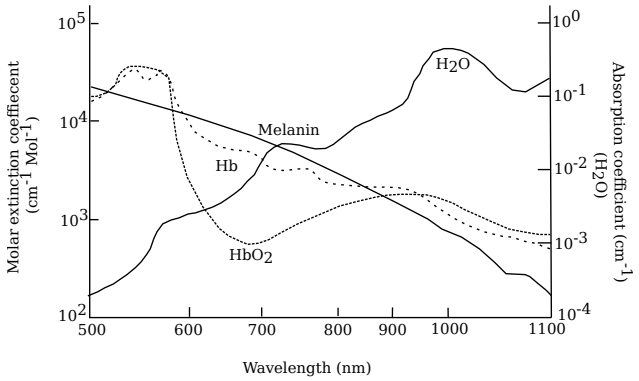


Fig. 4: The light absorption coefficients of biological compounds present in the epidermis-hypodermis layers of skin adapted from Lemay *et al.* [34].

tones during active state [53]. In contrast, Shcherbina *et al.* found co-variates such as higher skin melanin content, larger wrist circumference, and higher BMI to positively correlate with increased HR error rates across multiple wrist-worn reflectance mode PPG devices [54]. Additionally, Preejith *et al.* using a green light reflectance mode wrist-worn PPG sensor on 256 subjects found an mean absolute error of 1.04BPM for lighter skin tones compared to 10.90BPM for darker skin tones when computing HR estimations [55].

2) Skin Temperature: When the temperature of skin reduces, perfusion rates in the vascular bed dramatically decrease due to the Autonomic Nervous System constricting blood vessels in the dermis as a means to conserve body heat [56], [57]. Reductions in skin temperature typically affect the peripheral circulation more than the central areas of the body; for example, when the body is exposed to 10°C ambient temperature the blood flow through the hand drops to less than 1 ml/min [57].

All studies exploring temperature and PPG sensing agree that temperature influences the signal [38], [56]–[61] but to differing degrees. Ralston *et al.* suggested that errors resulting from varying skin temperatures are unlikely to be clinically significant for transmission mode PPG sensing [58]. However, Budidha *et al.* found that in some volunteers exposed to cold temperatures, the amplitude of the PPG signal significantly diminished to the extent of being unusable for ear-worn reflectance mode PPG sensing [57].

Maeda *et al.* found green light (525nm) reflectance mode PPG HR estimations to have a higher correlation to the ECG signal HR estimation than IR light (880nm) reflectance mode PPG HR estimations at temperatures below 15°C. As shorter wavelengths penetrate the skin to a lesser degree than longer wavelength they aren't subject to the optical processes that occur in deeper tissue which produce more complex signals. Green PPG signals include less information from various non-pulsatile media therefore affected by noise to a lesser degree than IR PPG signals [59]. In a subsequent study, Maeda *et al.* found that with cold exposure, the pulsatile component of both green and IR reflectance mode signals decreased significantly whilst the non-pulsatile component remained similar. With hot exposure, the pulsatile component of both green and IR signals as well as the non-pulsatile component the IR signals decreased. The decrease in amplitudes of both components of the IR signal during hot exposure is due to a larger amount of blood in the peripheral vascular bed [60].

3) Other Biological Factors: Although not examined with multi-wavelength PPG devices, there are several other known biological factors that affect PPG sensing. Higher body mass index (BMI) and obesity has been shown to produce less accurate HR estimations [54], [62], [63], which has been speculated to be a co-variate with larger wrist circumferences [64], however other studies reported BMI not to be an affecting factor [65], [66]. Shcherbina *et al.* also found sex to be an affecting factor with males getting higher error rates [54]. Additionally, Fine *et al.* explored several studies looking into subject age. As aging leads to various anatomical and physiological changes, the ability of PPG sensing to assess cardiovascular health varies [63].

Other research has suggested that sweat and hair follicle density may be adverse factors to PPG sensing [67] as well as research showing that in underwater conditions skin temperature significantly affects PPG sensing compared to dry conditions [61]. Finally, pre-existing conditions such as Raynaud's syndrome and Anemia may be affecting factors to the accuracy of PPG sensing.

III. Hardware and Data Collection

In this section, an examination of the current state-of-the-art research and commercial multi-wavelength PPG hardware solutions is given as well as a summary of the various data collection protocols that explored the use of multi-wavelength PPG devices.

A. Multi-wavelength PPG Research Hardware

Developments into multi-wavelength PPG sensing hardware has seen dramatic improvements of the past decade in research settings. Initial hardware was reliant on fiber optics [41], [68] which then progressed into Optical Electronic Patch Sensor (OEPS) development [51], [69] due to its low cost and simple form factor with researchers also exploring ear-worn, finger-worn, forehead-worn and wrist-worn PPG sensors [57], [70]–[72]. The latest development in hardware for multi-wavelength PPG sensing is an on-chip spectrometer approach based on plasmonic filters [73] which has been adapted for an all-wavelength PPG sensing device [74]. Summarized in Table I are various multi-wavelength PPG research hardware solutions.

The measurement site of PPG sensing is a key factor due to the varying qualities of tissue thickness, skin pigmentation, blood distribution in vascular bed and amount of movement present at the measurement site [76]–[81]. Researchers examined 52 measurement sites across the body finding fingers, palm, face, and ears to produce larger amplitudes of the pulsatile component of the signal when compared to other measurement sites [78]. These findings are consistent with other studies [76], [81]. However, when examining the effects of measurement site on motion artifacts it was found that

TABLE I: Multi-wavelength PPG Research Hardware Solutions

Paper	Sensor Type	Study Materials	Wavelengths	Comments
Spigulis *et al.* (2007) [41]	Finger-worn Reflectance Laser Sensor	**Input fiber:** 600μm silica core Z-Light, Ltd. Latvia **Round-to-line detection fiber bundle:** 7x 200μm silica core fibers Z-Light, Ltd. Latvia **Spectrometer:** AvaSpec 2048-2 Avantes BV, The Netherlands	**Violet:** 405nm **Green:** 532nm **Red:** 645nm **IR:** 807nm & 1064nm	Provides exact wavelengths of light making it suitable for clinical applications however it is unsuitable for continuous monitoring due to a lack of portability and obtrusive nature.
Leier *et al.* (2015) [75]	Wrist-worn Reflectance LED Sensor	**Four independent groups:** comprising of green, red and two infra-red LED emitters and a photodiode.	**Green:** 560nm **Red:** 660nm **IR:** 880nm & 940nm	All optical components are positioned on a flexible circuit board to allow for movement on the wrist. Light barriers are provided on the photodiodes to prevent light crossover and skin back-scattering. LEDs and photodiodes are in a matrix formation. Sensor is strapped to wrist to ensure sufficient contact force. It is unsuitable for continuous monitoring due to a large form factor of both the sensor and logic board as well as requiring a wired connection to a computer.
Warren *et al.* (2016) [71]	Forehead-worn Reflectance LED Sensor	Six photodiodes are positioned concentrically around two pairs of red and IR LEDs at an equidistant separation distance of 10 mm as well as a tri-axial accelerometer.	**Red:** 660nm **IR:** 940nm	Positioned on the forehead, signals collected are less susceptible to motion but may become obtrusive and inconvenient for daily monitoring. As there are 6 photodiodes the total active area is 15.9mm^2. An opaque ring was incorporated to minimize light crossover from LEDs and photodiodes.
Blanos *et al.* (2018) [69]	Reflectance LED OEPS	**PPG Sensor:** 4 channel board DISCO4, Dialog Devices Ltd., Reading, Berkshire, UK	**Green:** 525nm **Orange:** 590nm **Red:** 650nm **IR:** 870nm	Sensor has a small form factor with LEDs and photodiodes in a cross formation with the photodiode in the center. A layer of clear epoxy medical adhesive was used to protect the optical components. Patch sensors can be placed anywhere on the body however due to perspiration and general wear and tear it requires re-application making it of a disposable nature.
Budidha *et al.* (2018) [57]	Ear-worn Reflectance LED Sensor	**LED:** CR 50 IRH and CR 50 1M, Excelitas technologies, Massachusetts, USA **Photodiode:** SR 10 BP-BH, Excelitas technologies, Massachusetts, USA	**Red:** 658nm **IR:** 870nm	Ear-worn sensor has a small form factor with LEDs and photodiodes positioned next to each other. Ear-worn sensors are less susceptible to motion artifacts and are well suited to remote monitoring during specific activities. For 24 hour continuous monitoring ear worn sensors may become obtrusive and inconvenient.
Han *et al.* (2019) [42]	Reflectance LED Sensor	**PPG Sensor:** 2x AFE4404s Texas Instruments, Inc., Dallas, TX, USA	**Blue:** 460nm **Green:** 530nm **Red:** 660nm **IR:** 940nm	Sensor is in a circular formation with 2 layers of LEDs with the photodiode in the center. The sensor board was treated with a black coating to prevent light reflection as well as providing light barriers to prevent light crossover and skin back-scattering.
Chang *et al.* (2019) [73]	Finger-worn Reflectance LED Sensor	**PPG Sensor:** based on plasmonic filters which can be integrated onto a regular photo detector.	**Green:** 515nm **Red:** 630nm **IR:** 940nm	The sensor has a small form factor based on plasmonic filters, nanoscale structures on metal films. Providing a unique way to control polarization and wavelength of light passing through the structures. Fabrication cost of the plasmonic filters can be as low as a few dollars at volume.
Lee *et al.* (2020) [72]	Wrist-Worn Reflectance LED Sensor	**PPG Sensor:** 4X SFH7050 sensors OSRAM, Munich, Germany **Motion Sensor:** MPU9250, InvenSense, San Jose, CA, USA	**Green:** 530nm **Red:** 660nm **IR:** 940nm	The sensor consists of 4 integrated PPG sensing units positioned in a cross formation. Data collected from the sensor is streamed to a computer via Bluetooth, allowing for remote continuous monitoring. The sensor is attached to the wrist using a wrist sweatband which may not provide optimal contact force.
Gupta *et al.* (2020) [70]	Finger-worn Reflectance and Transmission mode LED Sensor	Two LEDs and a photodiode. Device allows for both transmission and reflectance type PPG signals. An Arduino micro-controller unit is used to control the whole system.	**Green:** 525nm **Red:** 615nm	Device provides both transmission and reflectance type PPG sensing. Transmission mode sensing typically requires obtrusive and inconvenient solutions for continuous monitoring due to positioning of device.
Chen *et al.* (2020) [74]	Wrist-worn Reflectance LED Sensor	**LEDs:** two white LEDs, a green LED (525nm), and a IR LED (940nm) **Photodiode:** chip-scale spectrometer, NSP32 (nanolambda, Daejeon, Korea) as well as a micro-controller and a Bluetooth Low Energy transceiver.	**All Wavelength:** 400-1000nm	The sensor has a small form factor based on plasmonic filters, nanoscale structures on metal films. Providing a unique way to control polarization and wavelength of light passing through the structures with a broad band of wavelengths that can be utilized. Fabrication cost of the plasmonic filters can be as low as a few dollars at volume.

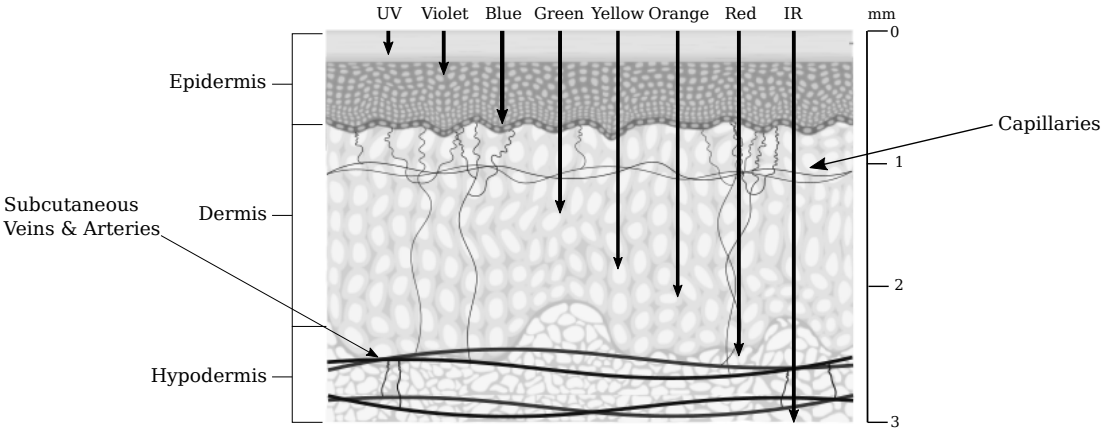


Fig. 5: Approximate maximum penetration depth of each wavelength of light in the skin using reflectance mode sensing.

motion had a large effect on the distribution of blood in the vascular bed at peripheral measurement sites such as fingers and wrist [76], [79].

Due to the preexisting widespread adoption of wrist-worn devices and their unobtrusive nature, the wrist is the most common measurement site for consumer-grade PPG sensing devices. Research shows that the wrist is the worst performing measurement site for extracting HR, pulse oximetry and respiration rate in periods of both rest and movement [81] highlighting the need for more robust methodology. Additionally, researchers have challenged the typical measurement site for wrist-worn PPG sensing device, suggesting the radial zone, side of wrist with the thumb, may produce better signals dependent on light wavelengths selected when compared to the central zone of the dorsal surface of the wrist [82], [83].

B. Multi-wavelength PPG Commercial Hardware

Polar Unite, Grit X and Vantage V2 are the only devices that currently use four wavelengths [84] whilst the other commercial devices have at most three; typically using green light for HR measurements and red and IR light for pulse oximetry measurements. “Research-grade” multi-wavelength wrist-worn PPG devices such as Empatica E4 and Biovotion Everion (now Biofourmis Biovitals) have the advantage of providing data-streamed raw signals, however, their HR estimation accuracy has been reported to be lower than consumer-grade wrist-worn PPG devices with Empatica E4 achieving a mean absolute error of 11.3BPM at rest and 12.8BPM during activity. Biovotion Everion achieved a mean absolute error of 16.5BPM at rest and 19.8BPM during activity whilst the Apple Watch achieved a mean absolute error of 4.4BPM at rest and 4.6BPM during activity [52]. This is in agreement with Rukasha *et al.* who found Empatica E4 heart rate estimate mean absolute percentage errors (MAPEs) between 7.2% and 29.2% whilst collecting data on a treadmill and heart rate estimate MAPEs between 5.3% and 13.5% during 12-hour continuous monitoring [85].

Concerns have arisen due to both consumer and research-grade devices being used in clinical trials, with Fitbit alone having 476 published studies and 449 studies registered on ClinicalTrials.gov [86] as well as Apple Watch having gained FDA 510(k) class II clearance for the ECG feature and ability to detect arrhythmias [64] and Empatica Embrace 2 gained FDA 510(k) clearance for epilepsy detection [87]. Fitbit, Garmin and Apple all state that their optical heart rate monitors should not be used as medical devices with intent to diagnose, treat, cure or prevent any disease [88]–[90].

C. Multi-wavelength PPG Sensing Units and Configurations

When creating multi-wavelength PPG devices, there are several design choices to be considered such as the number of and placement of LEDs and photo-detectors (PD), intensity of light from LEDs, sample rate, contact force and ambient light and electrical noise cancellation. Summarized in Table II are various multi-wavelength integrated PPG sensors that have been developed and optimized to eliminate having to make the design choices previously mentioned, however they lack the customizability and adaptability required for specific research cases. Analog Front Ends provide a means to develop a custom sensor module adapted to specific needs; a brief summary of multi-wavelength PPG Analog Front Ends is shown in Table III.

In order to develop a sensing unit with an Analog Front End that obtains a strong signal the correct placement of LEDs and PDs is necessary. Research suggests, to obtain the maximum AC/DC ratio, shorter wavelength LEDs such as green should be positioned 1.85mm from the PD whilst red and IR LEDs should be placed 2.35mm and 2.75mm from the PD, respectively [98]. At 9.75mm between the LED and PD, no pulsatile waveform was observed at any wavelength [42] and it was found that nearly double the driving current was needed to obtain a signal at similar distances apart for both red and IR LEDs [77]. Increasing the applied current and number of LEDs increases the radiation power [42] however much of the radiation power is not captured if the PD active area is small and may not produce any increase in signal amplitude [99]. Increasing the active area of the PD or number of PDs, however, produces a stronger signal with increases in average amplitude of 42% for wrist-worn red PPG signals and 73%

TABLE II: Multi-wavelength PPG Integrated Sensing Units

Device	Wavelength of LEDs				Features
	Blue	Green	Red	IR	
Analog Devices ADPD188GG [91]		2			2 Photodiodes I^2C & SPI Communication 2 external sensor inputs 3 LED drivers Ambient Light Rejection
Analog Devices ADPD144RI [92]			1	1	I^2C Communication external LED emitters Ambient Light Rejection
Maxim Integrated MAX30101 [93]		1	1	1	I^2C Communication Ambient Light Rejection
Maxim Integrated MAX86150 [94]			1	1	I^2C Communication Ambient Light Rejection Electrocardiogram
Maxim Integrated MAX86916 [95]	1	1	1	1	I^2C Communication Ambient Light Rejection
OSRAM SFH 7072 [96]		2	1	1	Light Barrier to block optical cross-talk Requires Analog Front End
OSRAM SFH 7050 [97]		1	1	1	Light Barrier to block optical cross-talk Requires Analog Front End

TABLE III: Multi-wavelength PPG Analog Front Ends

Device	Drivers	Features
Analog Devices ADPD4000/4001 [101] ADPD4100/4101 [102]	8 LED drivers 8 Inputs for PPG, ECG, EDA, impedance and temperature	I^2C & SPI Communication Ambient Light Rejection
Maxim Integrated MAX30110 [103]	2 LED 1 Photodiode	SPI Communication Ambient Light Rejection
Maxim Integrated MAXM86146 [104]	3 LED Two Integrated Photodiode	SPI Communication Ambient Light Rejection Integrated Micro Controller
Texas Instruments AFE4950 [105] AFE44S30 [106]	8 LED 4 Photodiode	1/2/3 Lead ECG (AFE4950) I^2C & SPI Communication Ambient Light Rejection
Texas Instruments AFE4900 [107]	4 LED 3 Photodiode	1 Lead ECG I^2C & SPI Communication Ambient Light Rejection
Texas Instruments AFE4404 [108]	3 LED 1 Photodiode	I^2C Communication Ambient Light Rejection

for IR. Additionally, increasing the number of PDs is more beneficial than increasing the number of LEDs as it provides reduced power consumption and heat generation [77], [99]. Finally, it is recommended that the wavelengths are collected in order of size, with the longest first, as the pulsatile event first occurs in the deeper blood vessels [42] as well as having a sample rate in the optimal range of 21–64Hz as to compress biomedical data and reduce storage needs [100].

A key factor in PPG sensing is contact force [77], [78], [83], [109]. As the contacting force of the sensor increases the amplitude of the pulsatile component of the signal increases until the difference between the external pressure and the intra-arterial pressure, called transmural pressure, reaches zero. At this point the amplitude of the pulsatile component of the signal starts to decrease as the external pressure increases until circulation is cut off due to the flattening of the arterial walls [78], [83], [109]. On the wrist, using reflectance mode, it has been suggested that a contact pressure of 80mmHg produces the highest amplitude for red light [83]. On the upper arm, using reflectance mode, an average contact pressure of 30mmHg produces the highest amplitude for green and IR light [78]. Finally, it was found that minimal contact pressure was needed on the forehead using reflectance mode [77].

D. Multi-wavelength PPG Data Collection

A summary of multi-wavelength PPG data collection protocols is given in Table IV. There are only three study protocols which account for skin tone when exploring multi-wavelength PPG sensing with larger numbers of lighter skin tones. However, there are single-wavelength data collection protocols that account for skin tone such as Bent *et al.* [52] and Preejith *et al.* [55] but fall outside of the scope of this paper. Additionally, there are only three protocols which explore skin temperature and multi-wavelength PPG sensing [57], [59], [60].

IV. Multi-wavelength PPG Motion Artifact Reduction

Motion artifacts are one of the main causes of inaccuracies in PPG sensing. Due to the placement of sensors and the varying penetration depths of light wavelengths, motion from the body distorts the PPG signal. Motion artifacts can be classified as either periodic or non-periodic and can possess amplitudes much larger than the pulsatile component of the signal [72], [116]. Blanos *et al.* showed that green (525nm) and orange (590nm) light were affected by artifacts from motion to a lesser degree than red light (650nm) [69]. Matsumura *et al.* agreed stating that the signal to noise ratio (SNR) for green (530nm) and blue (470nm) light was higher than red (640nm) when being subject to various motion types [110]. Shorter wavelengths can result in improved SNRs because their comparatively shorter path lengths and penetration depths make them less susceptible to noise from motion [72]. Shorter wavelengths also suffer from less attenuation from optical processes, such as absorption and scattering, due to their shorter path lengths. Additionally, the shallower penetration

TABLE IV: Multi-wavelength PPG Data Collection Study Protocols

Paper	Motivation	Cohort Metrics	Protocol	Wavelengths
Maeda *et al.* (2008) [59]	Skin Temperature	**Subjects:** 22 **EXP1:** **Age:** 22.4 ± 0.8 **Gender:** 8M 3F **EXP2:** **Age:** 23.6 ± 3.4 **Gender:** 8M 2F	**EXP1:** room temperature of 25°C Rest seated – 5 minutes Measurement taken **EXP2:** peripheral skin temperature of 15°C Rest seated – 5 minutes Measurement taken	**Green:** 525nm **IR:** 880nm
Maeda, Sekine and Tamura (2011) [60]	Skin Temperature	**Subjects:** 12 **Age:** 23.6 ± 1.5	**Seated** **Cold exposure:** Immersed the hand into the isothermal bath at 10°C Steady-state reached in 23.8±8.2 minutes Measurements - 1 minute **Hot exposure:** Immersed the hand into the isothermal bath at 45°C Steady-state reached in 20.7±7.3 minutes Measurements - 1 minute	**Green:** 525nm **IR:** 880nm
Fallow *et al.* (2013) [53]	Skin Tone	**Subjects:** 23 **Age:** 31 ± 12 **Gender:** 11M 12F **Height:** 172 ± 8cm **Weight:** 72 ± 14kg **Skin Type:** I & II = 8, III = 5, IV = 4, V = 6	Resting forearms Bicep Curl Flexion - 10s Grasping Dynamometer with force 5-10nm - 10s	**Blue:** 470nm **Green:** 520nm **Red:** 630nm **IR:** 880nm
Matsumura *et al.* (2014) [110]	Motion Artifacts	**Subjects:** 12 **Age:** 20.6 ± 0.76 **Gender:** 12M	**Adaptation period** - 5 minutes **Experimental period:** Horizontal motion – 20s, Rest – 10s, Vertical motion – 20s, Rest – 10s Baseline – 20s, Horizontal motion – 20s, Rest – 10s, Vertical motion – 20s, Rest – 10s, Baseline - 20s	**Blue:** 470nm **Green:** 530nm **Red:** 640nm **IR:** 810nm
Liu *et al.* (2015) [111]	Physiological Monitoring	**Subjects:** 10 **Age:** 22 - 60 **Gender:** 6M 4F	Rest – 1 minutes **Eight levels of cuff pressure:** 0 mmHg – 15s, 20mmHg – 15s, 40mmHg – 15s, Diastolic BP (DBP) – 15s, DBP+25% – 15s, DBP+50% – 15s, DBP+75% – 15s, DBP+100% – 15s **Deflated in the reverse order** Rest – 1 minute	**Blue:** 470nm **Green:** 570nm **Orange:** 591nm **Red:** 635nm
Alzahrani *et al.* (2015) [2]	Motion Artifacts	**Subjects:** 16 **Age:** 20 - 47 **Gender:** 15M 1F	Standing (30s), Sitting (30s), Walking - 3.0km/h (30s), Walking - 6.0km/h (30s), Cycling - 20.0km/h (60s), Cycling - 35.0km/h (60s), Running - 7.0km/h (30s), Running - 8.5km/h (30s)	**Green:** 525nm **Red:** 660nm **IR:** 990nm
Liu *et al.* (2016) [112]	Physiological Monitoring	**Subjects:** 20 (10 Healthy/10 Patients with Cardiovascular diseases (CVD)) **Average Healthy Age:** 26 **Average CVD Age:** 68	Subjects at rest in seated position Reference BP was measured on the middle finger and left upper arm. One-lead ECG electrodes on the left and right arms of the subjects. A custom made four-wavelength PPG device used to collect PPG signals	**Blue:** 470nm **Green:** 570nm **Yellow:** 590nm **IR:** 940nm
Warren *et al.* (2016) [71]	Motion Artifacts & Physiological Monitoring	**Subjects:** 15 **Age:** 23 – 30	Alternate between 3 min of rest and 5 min of bouncing on a exercise ball for a total of 19 min. Using a reference device Masimo-57 Radical (Masimo SET®, Masimo Corporation, CA, USA) finger type transmittance pulse oximeter that was kept motionless by resting the left hand on a table.	**Red:** 660nm **IR:** 940nm
Adhikari *et al.* (2016) [26]	Physiological Monitoring	**Subjects:** 5/3/3 Mice	A delivered dose was 5 mg/kg, which is the typical clinical dose. The PPG device is placed on the tail throughout the injection phase, then for short periods throughout the clearance phase. 6 samples for gold nanorods, 9 samples for quinine and 7-8 samples for amphotericin B	**UV:** 355nm **IR:** 805nm
Yan *et al.* (2017) [51]	Skin Tone	**Subjects:** 33 **Age:** 18 – 41 **Gender:** 33M **Skin Type:** I & II = 11, III = 10, IV = 7, V = 5	**Room temperature:** 23–26 °C, **Humidity:** 22–36% Resting, Walking (3km/h), Jogging (6km/h), Running (9km/h)	**Green:** 525nm **Orange:** 590nm **Red:** 650nm **IR:** 870nm
Geng *et al.* (2017) [113]	Physiological Monitoring	**Subjects:** 9 (6 Healthy/3 Diabetic)	3 lunch experiments & 2 supper experiments. Lunch was standardized to 90g of standard tortilla, while the supper was without specific requirements. Healthy volunteers did lunch experiment without wearing the dynamic glucometer. At 10min before the meal, finger stick glucose monitoring (Roche glucometer, ACCU-CHEK® Performa) was used for reference glucose then performed once every 30 mins.	**Red:** 660nm **IR:** 730nm, 800nm and 940nm.

continued on the next page

TABLE IV: Multi-wavelength PPG Data Collection Study Protocols

Paper	Motivation	Cohort Metrics	Protocol	Wavelengths
Budidha and Kyriacou (2018) [57]	Skin Temperature	**Subjects:** 15 **Age:** 28 ± 5 **Gender:** 9M 6F	**Basline Temperature** (24°C) – 2 minutes, **Cold Exposure** (10±1°C) – 10 minutes, **Baseline Temperature** (24°C) – 10 minutes	**Red:** 658nm **IR:** 870nm
Blanos *et al.* (2018) [69]	Physiological Monitoring	**Subjects:** 15 **Age:** 25 ± 5 **Height:** 178.9 ± 4.2cm **Weight:** 70.9 ± 7.9kg	Settle - 30s, Rest - 180s, Settle - 30s, Cycling - 180s, Settle - 30s, Run (3Km/h) - 180s, Settle - 30s, Run (6Km/h) - 180s	**Green:** 525nm **Orange:** 595nm **Red:** 650nm **IR:** 870nm
Pasta *et al.* (2018) [114]	Physiological Monitoring	**Subjects:** 25 **Age:** 28 ± 7 **Skin Tone:** I = 1, II = 9, III = 9, IV = 1, V = 2 VI = 1	**OEPS signal was measured at:** Fingertip, Rest - 2 minutes, Radial artery, Rest - 2 minutes, Wrist, Rest - 2 minutes	**Green:** 525nm **Orange:** 595nm **Red:** 650nm **IR:** 870nm
Alharbi *et al.* (2019) [115]	Physiological Monitoring	**Subjects:** 31 **Age:** 25±5 **Gender:** 31M **Height:** 179 ± 4cm	**Protocol 1:** Sitting with hand movements **Protocol 2:** Cycling and Walking	**Green:** 525nm **Orange:** 595nm **Red:** 650nm **IR:** 870nm
Zhang *et al.* (2019) [116]	Motion Artifacts	**Subjects:** 6 **Age:** 25-35	**Stationary** – 5 minutes **Motion:** Index finger tapping, Hand waving (horizontal), Hand shaking (vertical), Running arm swing, Fist opening and closing, Radial/ulnar deviation, Wrist extension/flexions	**Green:** 560nm **IR:** 940nm
Rachim *et al.* (2019) [117]	Physiological Monitoring	**Subjects:** 12 **Age:** 22–30, **Gender:** 2F 4M	Commercial ECG device AD8232 (Analog Devices Inc., USA) and PPG device RP520 (Laxtha Inc., Korea) were used as reference devices. Each subject sat in a chair. Data points are collected every 10 min, to find baseline value. After that, the subject ate a carbohydrate rich meals, then collected data every 20 min for a total 120 min	**Green:** 530nm **Red:** 660nm **IR:** 850 & 950nm
Chang *et al.* (2019) [73]	Motion Artifacts & Physiological Monitoring	**Subjects:** 10 **Age:** 20-60 **Gender:** 7M 3F **Height:** 155-180cm	Index finger is placed on the sensing devices A blood oximetry meter (TRUST, TD-8250A) and an upper arm blood pressure monitor (Omron, HEM-7121) are used as reference devices	**Green:** 505, 510, 515, 520 & 525nm **Red:** 620, 625, 630, 635 & 640nm **IR:** 930, 935, 940, 945 & 950nm
Lee *et al.* (2020) [72]	Motion Artifacts	**Subjects:** 7 **Age:** 27.1±5	Resting – 1 minute, Walking- 2 minutes (1m/s), Resting – 1 minute, Fast Walking – 2 minutes (1.8m/s), Resting – 1 minute, Running - 2 minutes (2.2m/s)	**Green:** 530nm **Red:** 660nm **IR:** 940nm
Liu *et al.* (2020) [118]	Physiological Monitoring	**Subjects:** 22 **Age:** 70.2 ± 5.4 **Gender:** 17M 5F	**Rest** – 5 minutes Measurements taken **Rehabiliation Exercise** – 2 hours lower limb strengthening and balance exercises with mobility and agility training. Rest – 5 minutes Measurements taken	**Blue:** 460nm **Green:** 575nm **Orange:** 590nm **IR:** 940nm
Chen *et al.* (2020) [74]	Motion Artifacts	**Subjects:** 6	Used the developed device to record the MRC-AW-PPG signals, and then used AFE4404EVM as a reference to record the green, red, and NIR PPG signals separately, all in the three different postures (hands down, hands forward & hands up) whilst seated	**All Wavelength:** 400-1000nm

depths for shorter wavelengths result in less physiological information from deeper tissue such as bone movement being collected [69]. However, some shorter wavelengths due to shallower penetration depths do not reveal much cardiac activity [72]. The typical frequency range of a PPG signal is 0-5Hz whilst motion artifacts fall within 0-10Hz making the removal of motion artifacts challenging. Most approaches that tackle motion artifacts involve the use of a motion reference signal, typically collected from an accelerometer or gyroscope [116]. Wang *et al.* used the isobestic (800nm) wavelength as a motion reference, implementing a normalized least mean squares adaptive noise canceling algorithm to reconstruct the clean PPG signal [119]. Zhang *et al.* proposed a similar method using an infrared (940nm) PPG signal as a motion reference, due to its comparatively deep penetration depths and susceptibility to motion implementing a continuous wavelet transform for signal cleaning and reconstruction reducing error in HR estimations for all motion types to less than 2BPM [116]. Yao *et al.* developed a method to separate motion artifacts from PPG signals using an algorithm based on the Beer-Lambert law which utilized red (660nm) and two infrared (850 and 940nm) wavelengths [37]. The Beer-Lambert law suggests that the sum of transmitted and absorbed or scattered light is equal to the incident light through homogeneous layers however human skin and blood are not homogeneous [34]. Chang *et al.* used 15 PPG signals and the maximal-ratio combined algorithm

as a means for motion artifact reduction showing a 50% variation reduction when compared with the single-wavelength reference sensor [73]. Similarly, Chen *et al.* used a maximal-ratio combined algorithm on an all-wavelength wrist-worn PPG device. The results showed the all-wavelength approach had an improved SNR when compared to the conventional green, red and IR PPG sensing [74]. Lee *et al.* developed a motion artifact reduction algorithm using 12-channel PPG signals comprising of green (530nm), red (660nm) and IR (940nm) wavelengths. An independent component analysis was first carried out to extract the independent components of the signals. The most pulsatile component of the signals were then selected using principal component analysis implemented with a truncated singular value decomposition showing a sensitivity of 82.49%, a positive predictive value of 99.83%, and a false detection rate of 0.17% in periods of high motion [72].

V. Multi-wavelength PPG Physiological Monitoring

In this section, an examination of the current state-of-the art research physiological measurement extraction methodologies is given ranging from cardiovascular measurements such as blood oxygen saturation, HR, blood pressure and blood glucose as well as other physiological measurements.

A. Blood Oxygen Saturation & Heart Rate

Currently, the most common application for multi-wavelength PPG sensing is pulse oximetry as it requires two wavelengths to calculate blood oxygen saturation levels. The blood oxygen saturation level can be estimated from the ratio of pulsatile and non-pulsatile components of each wavelength [9]. Typically, the wavelengths used are red (622–780nm) and IR (780-2400nm) [73] however, researchers have identified orange and green light to perform better due to their robustness against motion artifacts [69], [115].

Alharbi *et al.* found that green-orange pulse oximetry measurements from a reflectance mode OEPS device had a r=0.98 correlation with commercial pulse oximeter in periods of both rest and motion as well as a r=0.98 correlation with pulse oximetry measurements from red-IR light using the same OEPS device in periods of both rest and motion [115]. Additionally, Blanos *et al.* found no significant difference between green-orange pulse oximetry measurements from a reflectance mode OEPS device and a commercial pulse oximeters in periods of both rest and motion [69]. Blanos *et al.* also extracted HR estimations from a reflectance mode OEPS device at four different wavelengths in both periods of rest and motion. It was found that green light had a correlation of r=0.992 with the recorded ECG values. Orange light had a correlation of r=0.984, whilst red and IR light had a correlation of r=0.952 and r=0.97, respectively [69].

Warren *et al.* developed a multi-channel, multi-wavelength forehead-worn PPG reflectance sensor, using two red (660nm) and two IR (940nm) wavelengths, with a tri-axial accelerometer. They also developed an advanced multi-channel switching algorithm that chooses the channel least affected by motion artifacts to calculate HR estimates for each time instant. They found that for a wide variety of random motion, channels respond differently to motion artifacts. The multi-channel switching algorithm estimates produced improved results compared to the individual channel estimates because the multi-channel switching algorithm enabled automatic selection of the best signal fidelity channel at each time point among the multi-channel PPG data [71].

Green-orange pulse oximetry has shown to be a promising alternative due to their robustness against motion [69] and should be explored further in daily activity settings. Utilizing two wavelengths for HR estimations showed promising results [71]. Further exploration of methods using multiple wavelengths for HR estimations may uncover improved results.

B. Blood Pressure

Blood Pressure (BP) can be extracted using an ECG and peripheral PPG sensing to compute the Pulse Transit Time (PTT) which has a high correlation with systolic blood pressure (SBP) and diastolic blood pressure (DBP). Liu *et al.* developed a reflectance mode multi-wavelength light-skin interaction model based on the modified Beer-Lambert law. The model was calibrated for BP extraction using a cuff-based BP measuring device and ECG. Evaluating the dominance of different pulsation patterns based on absorption weighting factors showed a significantly improved BP tracking ability. The mean absolute difference between the reference and the estimated SBP varies from 5.7mmHg (for single-wavelength PPG) to 4.0mmHg (for two-wavelength PPG) and 2.9mmHg (for three-wavelength PPG) [112]. Blood pressure estimation methods require a cuff-based BP measuring device for calibration. When the cuff is inflated, pressure is exerted on the vascular bed causing the arterial properties to potentially be altered due to the smooth muscle relaxation, thereby increasing the PTT. Liu *et al.* examined the effect of cuff induced pressure and subsequent effects on the PTT at four different wavelengths of light using reflectance mode: blue (470nm), green (570nm), yellow (591nm) and red (635nm). The results showed that red PTT, yellow PTT and green PTT had a trend of increased PTTs after cuff pressurization while blue PTT nearly had no change. Indicating that PTTs calculated from cuff-based BP measuring devices and different wavelengths of PPG are influenced by smooth muscle relaxation to different degrees. Blue light has a relatively shallow penetration depth so blue PTT stays nearly unchanged after cuff pressurization. Yellow PTT had the most significant change which may be due to yellow light having deeper penetration depths into the skin compared to blue light therefore reaching muscle tissue that is influenced by smooth muscle relaxation from the inflated cuff to a greater degree. Additionally, yellow light has shallower penetration depths than red light so is unable to reach the larger blood vessels found in deeper tissue [111].

Efforts have been made to extract BP estimates without ECG devices using the time difference between different wavelengths of PPG signals, referred to as local PTT, due to its strong relationship to PTT showing promising results. Pasta *et al.* examined BP cuff-based measurements from reflectance

mode multi-wavelength PPG sensing at 3 different locations (fingertip, radial artery and dorsal surface of wrist) using four different wavelengths (green 525nm, orange 595nm, red 650nm and IR 870nm) without an ECG device for calibration. When the cuff was pressurized, the blood vessels were gradually blocked by the increasing pressure. As the systolic pressure was reached, the PPG signals became too weak for the sensor to pick it up. Upon the signal's disappearance, the cuff pressure was decompressed gradually. Then, the PPG pulse reappeared at a certain point. The algorithm was able to identify all the peaks and provided information such as the time of signal loss and re-acquisition, thus allowing for a correlation with the pressure inside the cuff. The results showed that the fingertip site provided the most accurate values amongst all wavelengths with an error of 8.07%, compared to the radial artery error of 13.17% and the wrist error of 17.44%. Green light recorded the best performance for every site, followed by the orange light with an error difference of 2%. Red light obtained the best results on the fingertip, with an error of 6.33% whilst IR had an error of 7.27%. Additionally, it was reported that smaller error rates were obtained from lighter skin tones compared to darker skin tones [114].

Chang *et al.* produced a method without an ECG device using 15 different reflectance mode wavelengths extracting local PTT using a cross-correlation method. The average of the 15 local PTTs was computed and used with regression coefficients of the linear models for SBP and DBP to estimate SBP and DBP values with correlations of r=0.79 and r=0.78, respectively [73]. Liu *et al.* proposed a method, similarly without an ECG device. Using shorter wavelengths, blue and green, to measure the capillary pulsation and longer wavelengths, red and IR, to measure the arterial pulsation using reflectance mode sensing. Principle component analysis was employed to extract the first principle component of the shorter wavelengths as the capillary pulse and the second principle component as the motion signal as well as the first principle component of the longer wavelengths as the arterial pulse. From these principle components, Fourier transforms are used to extract features such as phase shift which indicates arteriolar PTT and HR with heart period and pulse decay time being computed separately. These features are used to compute mean blood pressure and pulse pressure which are then transformed into SBP and DBP estimates yielding errors of 1.44 ± 6.89mmHg for SBP and -1.00 ± 6.71mmHg for DBP [118].

Utilizing multiple wavelengths for BP estimations showed improved error rates compared to single-wavelength estimations [73], [112], [118] especially when taking advantage of the differing interactions of wavelengths with skin and blood [118]. Methods that do not require an ECG device [73], [114], [118] have the advantage of not requiring multiple devices.

C. Blood Glucose

As well as Blood Pressure sensing, another application of wearable multi-wavelength PPG sensing is Blood Glucose (BG) estimation. Gupta *et al.* analyzed the mode of PPG sensing when extracting BG measurements. Two wavelengths of light, green (525nm) and red (615nm), were collected from the finger using both transmission and reflectance modes of PPG sensing. All collected signals were subject to a 10th order low pass Butterworth filter with a cutoff frequency of 8 Hz. The filtered signals were then used to extract 22 features which can be split into two parts: PPG based and general signal characteristics. The PPG features included HR, SpO_2 and breathing rate whilst the signal characteristics included zero-crossing rate, power spectral density, Teager–Kaiser energy and Qi-Zheng energy. These features were then used in a random forest regression algorithm. In transmission, the correlation between the estimated BG measurements and the reference device was r=0.72 pre-prandial (before food consumption) and r=0.91 post-prandial (after food consumption). In reflectance mode, the correlation pre-prandial was r=0.82 whilst the correlation post-prandial was r=0.62 [70].

Geng *et al.* developed a multi-site and multi-sensor system consisting of a wrist-worn device and an upper-arm worn device. The wrist-worn device contained a temperature sensor, a humidity sensor, a high frequency flexible electrode and one pole of a low-frequency electrode as well as a multi-wavelength reflectance PPG sensor (red - 660nm and IR - 730nm, 800nm and 940nm). The upper arm device was equipped with the other pole of the low-frequency electrode to detect the low-frequency impedance of the arm. All candidate features were calculated from the original signals and were screened according to the similarity between the feature and reference glucose profile. A single-feature model was constructed based on the candidate features using time series analysis. A weighted average method was used on the single-feature model-based glucose profiles to produce multi-feature fusion parameters. The glucose profile estimation model is made up of both the single-feature model-based glucose profiles and multi-feature fusion parameters. After the estimated glucose profile was obtained, the peak time of postprandial glucose can be obtained. The results show a correlation between the reference device and the estimated values of r=0.83 and a standard error of prediction (SEP) of 14.6mg/dL [113].

Rachim *et al.* analyzed four wavelengths of light (green, red and two IR) in extracting BG measurements from the wrist using reflectance mode sensing. A local maxima algorithm was used to detect the peaks in the collected signals which were then segmented into windows and averaged using an ensemble average algorithm. From the averaged signals, 24 features were extracted: 12 features from the difference of optical density between the pulsatile components and the amplitude of non-pulsatile component as well as 12 features from a Teager–Kaiser energy operator. The features were then used in a Partial Least Squares algorithm to find the relationship between the reference device and extracted features. Using only green (535nm) and red light (660nm) a SEP of 12.4mg/dL was found and a correlation of r=0.55 with the reference device. Using only IR light (850nm and 950nm) a SEP of 10.1mg/dL was found and a correlation of r=0.67 with the reference device. Finally, using all four wavelengths a SEP of 6.16mg/dL was found and a correlation of r=0.86 with the reference device [117].

Similarly to BP estimation, using multiple wavelengths for

BG estimations produced the lowest error rates [117]. Methods that required several sensors at multiple sites showed similar error rates [113] to methods with one device. Finally, deep learning as a method for feature extraction may be advantageous for BG estimation and should be explored further.

D. Drug Delivery Monitoring

Adhikari *et al.* developed a multi-wavelength transmission mode PPG method for the monitoring of drug delivery. They examined the use of Gold Nanorods, Quinine and Amphotericin B in mice possessing absorption peaks of 805nm, 355nm and 355nm, respectively. Blood samples were collected from the mice after each PPG reading. Estimates were calculated using the pulsatile and non-pulsatile components of the signal to determine the extinction change due to pulsation at each wavelength using the Beer-Lambert law. The results showed that Gold Nanorods had a correlation of r=0.94 with the blood samples, Quinine had a correlation of r=0.96 and Amphotericin B had a correlation of r=0.88. This methodology could be used to monitor the circulation of molecular drugs and therapeutic nanoparticles having variable circulation half-lives and could be applicable to a wide range of optically active drugs and nanoparticles [24]–[26].

VI. Discussion and Recommendations

Multi-wavelength PPG shows promising signs of becoming a viable method for remote physiological monitoring as well as an alternative to ECG for cardiovascular monitoring. The selection of wavelengths used in PPG sensing is a compromise. It appears that green light (492–577nm) produces the best generalized modulation, however, multi-wavelength approaches for HR, blood pressure and blood glucose estimations have been shown to out-perform single-wavelength approaches. Multi-wavelength approaches in clinical applications are not unique to PPG sensing. Narrow-band imaging for gastrointestinal endoscopy has seen improvements from white light endoscopy using blue and green light due to their varying interactions with blood and tissue. Blue light (400–430nm) penetrates to the depth of the capillaries in the superficial mucosa, while the green light (525–555nm) penetrates deeper into the mucosa [120]. With the use of multiple wavelengths, the accuracy, robustness and generalizability of PPG sensing could be dramatically increased.

The findings, albeit limited, are conclusive that skin temperature affects PPG sensing which is concerning since the number of PPG studies that include skin temperature as a factor is small. There has been no research to the best of the authors' knowledge exploring the combination of motion artifacts, cold skin temperatures and higher skin melanin content. Data and insights from studies with combined factors could support the development of more robust solutions for continuous PPG sensing.

With small study sizes, anecdotal evidence and conflicting findings, understanding of the magnitude and scope of the potential inaccuracies of current PPG sensing due to skin melanin content is unclear. This is a concerning problem given that 80% of the world population are individuals with pigmented skin [121] and it has been projected that by 2035 half of the black population in USA will be affected by CVDs [122]. In order to address this problem researchers and industry professionals need to increase the diversity of subjects in validation studies to have proportional representation.

Current studies exploring skin melanin content tend to have smaller numbers of participants with darker skin tones [51], [53], [114] raising concerns of misleading conclusions [64], [86]. Bent *et al.* was the only study, to the best of the authors' knowledge, to have proportional distribution of skin tones [52], [123]. In addition, the current practice of classifying skin tone using the Fitzpatrick Skin Type Scale or the Von Luschan's chromatic scale, is a subjective process that may vary based on the administrator. It has been suggested that an objective approach to skin tone classification using a spectrophotometer should be employed as the "gold standard" to eliminate the shortcomings of the current practice [86], however spectrophotometers are expensive preventing wide-spread adoption, and there is evidence that "skin color evaluation with a spectrophotometer is correlated with visual skin tone assessment" and that "in both objective and subjective measurement methods, human error may be introduced through improper measurement methodology" [123].

There have been cases of racial bias, due to lack of proportional representation in validation studies, appearing in the surrounding disciplines. In the medical field, students have been petitioning to remove the "white skin bias" from medical textbooks as an extensive list of skin conditions, such as meningococcemia, appear different in patients with darker skin tones which are not accounted for in the texts [124], [125]. More generally, there is an increased awareness of cases of algorithmic bias against black individuals. Such cases include a healthcare prediction algorithm, used by more than 200 million patients in USA, that was less likely to refer equally sick black patients than white patients to programs aimed to improve care for patients with complex needs [126]. Additionally, in the optical engineering field, there are anecdotal examples of automatic taps and soap dispensers not working for individuals with darker skin tones [127].

There have been promising developments in motion artifact reduction using multi-wavelengths. Using IR light as a motion reference [116] allows for a more efficient solution as motion sensors such as accelerometers are not required as well as algorithmic approaches using several wavelengths [72]. Research in single-wavelength PPG sensing has explored the use of machine learning and deep learning models, such as Convolutional Neural Networks and Long Short-Term Memory networks, to accurately and robustly remove motion artifacts and estimate heart rate [128]–[130] and blood pressure [131]. This methodology would be well suited to multi-wavelength PPG sensing and should be explored further. Studies, however, lacked the exploration of combined adverse features in their experiments such as skin tone and skin temperature which could identify weaknesses in the proposed methodology.

As research into the field grows, standardization of reporting is of paramount importance in order to produce results that can be replicated and compared. Nelson *et al.* have produced a robust descriptive reporting protocol which if used would

standardize the study design, technological factors, participant characteristics as well as data analysis, data reporting and data transparency [64]. It is also recommended that when using consumer-grade PPG sensing devices or off-the-shelf hardware components, accurate and complete information, such as software and firmware versions, should be given in order to allow the replication of experiments [132].

VII. Conclusion

In this paper, we have presented a comprehensive review on multi-wavelength PPG sensors, encompassing state-of-the-art research work and recommending potential directions for future developments with an emphasis on data collection protocols. In the first two sections theoretical details are given regarding the workings of PPG sensing and the optical properties of skin and blood. Additionally, biological factors that affect PPG sensing, such as skin melanin content and skin temperature, are explored showing conflicting findings highlighting the importance for these topics to be explored in greater detail. Multi-wavelength PPG solutions involve design considerations such as measurement site, contact force and sensor geometry as well as data collection protocols were explored to aid the decision process for future research. Finally, state-of-the-art multi-wavelength motion artifact reduction and physiological monitoring methods were summarized showing promising results highlighting the breadth of applications that multi-wavelength PPG sensing is capable of.

Acknowledgment

D. Ray thanks Nicholas C. S. Ray and Iris Ray for their unconditional support throughout his academic ventures.

References

[1] Y. T. Zhang, "Editorial: Health engineering for urgent challenges in cardiovascular disease," IEEE Reviews in Biomedical Engineering, vol. 13. Institute of Electrical and Electronics Engineers, pp. 3–4, 2020, doi: 10.1109/RBME.2019.2959113.

[2] A. Alzahrani, S. Hu, and V. Azorin-Peris, "A comparative study of physiological monitoring with a wearable opto-electronic patch sensor (OEPS) for motion reduction," Biosensors, vol. 5, no. 2, pp. 288–307, 2015, doi: 10.3390/bios5020288.

[3] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," Int. J. Biosens. Bioelectron., vol. 4, no. 4, 2018, doi: 10.15406/ijbsbe.2018.04.00125.

[4] CCS Insight, "Optimistic outlook for wearables," CSS Insight, 2019. [Online]. Available: www.ccsinsight.com/press/company-news/optimistic-outlook-for-wearables/. Accessed on: Mar. 30, 2021

[5] D. Biswas, N. Simoes-Capela, C. Van Hoof, and N. Van Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: a review," IEEE Sens. J., vol. 19, no. 16, pp. 6560–6570, 2019, doi: 10.1109/JSEN.2019.2914166.

[6] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," Physiol. Meas., vol. 28, no. 3, 2007, doi: 10.1088/0967-3334/28/3/R01.

[7] A. B. Hertzman, "Photoelectric plethysmography of the fingers and toes in man," Proc. Soc. Exp. Biol. Med., vol. 37, no. 3, pp. 529–534, Dec. 1937, doi: 10.3181/00379727-37-9630.

[8] T. Pereira et al., "Photoplethysmography based atrial fibrillation detection: a review," npj Digital Medicine, vol. 3, no. 1. Nature Research, pp. 1–12, 01-Dec-2020, doi: 10.1038/s41746-019-0207-9.

[9] T. Tamura, "Current progress of photoplethysmography and SPO2 for health monitoring," Biomedical Engineering Letters, vol. 9, no. 1. 2019, doi: 10.1007/s13534-019-00097-w.

[10] P. H. Charlton et al., "Breathing rate estimation from the electrocardiogram and photoplethysmogram: A Review," IEEE Rev. Biomed. Eng., vol. 11, pp. 2–20, 2018, doi: 10.1109/RBME.2017.2763681.

[11] Y. K. Qawqzeh, "The analysis of PPG time indices to predict aging and atherosclerosis," in Intelligent Computing Paradigm and Cutting-edge Technologies, Springer, Cham, 2020, pp. 218–225.

[12] G. Wang, M. Atef, and Y. Lian, "Towards a continuous non-invasive cuffless blood pressure monitoring system using PPG: systems and circuits review," IEEE Circuits Syst. Mag., vol. 18, no. 3, pp. 6–26, Jul. 2018, doi: 10.1109/MCAS.2018.2849261.

[13] M. Elgendi et al., "The use of photoplethysmography for assessing hypertension," npj Digit. Med., vol. 2, no. 1, pp. 1–11, Dec. 2019, doi: 10.1038/s41746-019-0136-7.

[14] N. Mangathayaru, B. P. Rani, V. Janaki, L. S. Kotturi, M. Vallabhapurapu, and G. Vikas, "Heart rate variability for predicting coronary heart disease using photoplethysmography," in Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020, 2020, pp. 664–671, doi: 10.1109/I-SMAC49090.2020.9243316.

[15] N. Pinheiro et al., "Can PPG be used for HRV analysis?," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2016, vol. 2016-Octob, pp. 2945–2949, doi: 10.1109/EMBC.2016.7591347.

[16] C. Brüser, C. H. Antink, T. Wartzek, M. Walter, and S. Leonhardt, "Ambient and unobtrusive cardiorespiratory monitoring techniques," IEEE Rev. Biomed. Eng., vol. 8, pp. 30–43, 2015, doi: 10.1109/RBME.2015.2414661.

[17] T. Rukasha, S. I. Woolley, T. Kyriacou, and T. Collins, "Evaluation of wearable electronics for epilepsy: a systematic review," Electron., vol. 9, no. 6, pp. 1–16, Jun. 2020, doi: 10.3390/electronics9060968.

[18] M. K. Uçar, S. Örenç, M. R. Bozkurt, and C. Bilgin, "Evaluation of the relationship between chronic obstructive pulmonary disease and photoplethysmography signal," in 2017 Medical Technologies National Conference, 2017, pp. 1–5, doi: 10.1109/TIPTEKNO.2017.8238032.

[19] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, "Wearable-based affect recognition—a review," Sensors (Switzerland), vol. 19, no. 19. MDPI AG, 2019, doi: 10.3390/s19194079.

[20] P. H. Charlton, P. Celka, B. Farukh, P. Chowienczyk, and J. Alastruey, "Assessing mental stress from the photoplethysmogram: A numerical study," Physiol. Meas., vol. 39, no. 5, May 2018, doi: 10.1088/1361-6579/aabe6a.

[21] M. Shokoueinejad et al., "Sleep apnea: A review of diagnostic sensors, algorithms, and therapies," Physiological Measurement, vol. 38, no. 9. Institute of Physics Publishing, pp. R204–R252, 18-Aug-2017, doi: 10.1088/1361-6579/aa6ec6.

[22] H. Scott, L. Lack, and N. Lovato, "A systematic review of the accuracy of sleep wearable devices for estimating sleep onset," Sleep Medicine Reviews, vol. 49. W.B. Saunders Ltd, p. 101227, 01-Feb-2020, doi: 10.1016/j.smrv.2019.101227.

[23] S. Habbu, M. Dale, and R. Ghongade, "Estimation of blood glucose by non-invasive method using photoplethysmography," Sadhana - Acad. Proc. Eng. Sci., vol. 44, no. 6, pp. 1–14, Jun. 2019, doi: 10.1007/s12046-019-1118-9.

[24] P. Adhikari, I. B. Magaña, and P. D. O'Neal, "Multi-wavelength pulse plethysmography for real-time drug delivery monitoring," in Optical Diagnostics and Sensing XIV: Toward Point-of-Care Diagnostics, 2014, vol. 8951, p. 89510P, doi: 10.1117/12.2040064.

[25] P. Adhikari, W. Eklund, and D. P. O'Neal, "Non-invasive in vivo monitoring of circulating amphotericin b using multi-wavelength photoplethysmography," in Optical Diagnostics and Sensing XV: Toward Point-of-Care Diagnostics, 2015, vol. 9332, p. 93320H, doi: 10.1117/12.2083798.

[26] P. Adhikari, W. Eklund, E. A. Sherer, and D. P. O'Neal, "Assessment of multi-wavelength pulse photometry for non-invasive dose estimation of circulating drugs and nanoparticles," in Optical Diagnostics and Sensing XVI: Toward Point-of-Care Diagnostics, 2016, vol. 9715, p. 97150O, doi: 10.1117/12.2213455.

[27] W. Montagna, A. M. Kligman, and K. S. Carlisle, Atlas of normal human skin. 1992.

[28] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000 nm," J. Phys. D. Appl. Phys., vol. 38, no. 15, pp. 2543–2555, Aug. 2005, doi: 10.1088/0022-3727/38/15/004.

[29] M. Brenner and V. J. Hearing, "The protective role of melanin against UV damage in human skin," Photochemistry and Photobiology, vol. 84,

no. 3. NIH Public Access, pp. 539–549, May 2008, doi: 10.1111/j.1751-1097.2007.00226.x.

[30] G. Zonios, J. Bykowski, and N. Kollias, "Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy," J. Invest. Dermatol., vol. 117, no. 6, pp. 1452–1457, 2001, doi: 10.1046/j.0022-202x.2001.01577.x.

[31] B. Müller et al., "High-resolution tomographic imaging of microvessels," Dev. X-Ray Tomogr. VI, vol. 7078, p. 70780B, Sep. 2008, doi: 10.1117/12.794157.

[32] T. Lister, P. A. Wright, and P. H. Chappell, "Optical properties of human skin," J. Biomed. Opt., vol. 17, no. 9, p. 0909011, Sep. 2012, doi: 10.1117/1.jbo.17.9.090901.

[33] R. R. Anderson and J. A. Parrish, "The optics of human skin," J. Invest. Dermatol., vol. 77, no. 1, pp. 13–19, 1981, doi: 10.1111/1523-1747.ep12479191.

[34] M. Lemay, M. Bertschi, J. Sola, P. Renevey, J. Parak, and I. Korhonen, "Application of optical heart rate monitoring," in Wearable Sensors: Fundamentals, Implementation and Applications, Elsevier Inc., 2014, pp. 105–129.

[35] C. R. Simpson, M. Kohl, M. Essenpreis, and M. Cope, "Near-infrared optical properties of ex vivo human skin and subcutaneous tissues measured using the Monte Carlo inversion technique," Phys. Med. Biol., vol. 43, no. 9, pp. 2465–2478, 1998, doi: 10.1088/0031-9155/43/9/003.

[36] P. Taroni, A. Pifferi, A. Torricelli, D. Comelli, and R. Cubeddu, "In vivo absorption and scattering spectroscopy of biological tissues," Photochem. Photobiol. Sci., vol. 2, no. 2, pp. 124–129, 2003, doi: 10.1039/b209651j.

[37] J. Yao and S. Warren, "A novel algorithm to separate motion artifacts from photoplethysmographic signals obtained with a reflectance pulse oximeter," in Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings, 2004, vol. 26 III, pp. 2153–2156, doi: 10.1109/iembs.2004.1403630.

[38] L. Linberg and P. Öberg, "Photoplethysmography: Part 2 Influence of light source wavelength," Med. Biol. Eng. Comput., no. January, pp. 48–49, 1991, doi: 10.1007/bf02446295.

[39] A. Alzahrani et al., "A multi-channel opto-electronic sensor to accurately monitor heart rate against motion artefact during exercise," Sensors (Switzerland), vol. 15, no. 10, pp. 25681–25702, Oct. 2015, doi: 10.3390/s151025681.

[40] V. Vizbara, A. Solosenko, D. Stankevicius, and V. Marozas, "Comparison of green, blue and infrared light in wrist and forehead photoplethysmography," Biomed. Eng. 2014, pp. 78–81, doi: 10.1109/IEMBS.2008.4649649.

[41] J. Spigulis, L. Gailite, A. Lihachev, and R. Erts, "Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography," in Applied Optics, 2007, vol. 46, no. 10, pp. 1754–1759, doi: 10.1364/AO.46.001754.

[42] S. Han, D. Roh, J. Park, and H. Shin, "Design of multi-wavelength optical sensor module for depth-dependent photoplethysmography," Sensors (Switzerland), vol. 19, no. 24, Dec. 2019, doi: 10.3390/s19245441.

[43] D. Barolet, "Light-emitting diodes (LEDs) in dermatology," Semin. Cutan. Med. Surg., vol. 27, no. 4, pp. 227–238, Dec. 2008, doi: 10.1016/j.sder.2008.08.003.

[44] A. L. Ries, L. M. Prewitt, and J. J. Johnson, "Skin color and ear oximetry," Chest, vol. 96, no. 2, pp. 287–290, 1989, doi: 10.1378/chest.96.2.287.

[45] J. R. Feiner, J. W. Severinghaus, and P. E. Bickler, "Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: The effects of oximeter probe type and gender," Anesth. Analg., vol. 105, no. SUPPL. 6, 2007, doi: 10.1213/01.ane.0000285988.35174.d9.

[46] P. E. Bickler, J. R. Feiner, and J. W. Severinghaus, "Effects of skin pigmentation on pulse oximeter accuracy at low saturation," Anesthesiology, vol. 102, no. 4, pp. 715–719, 2005, doi: 10.1097/00000542-200504000-00004.

[47] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial bias in pulse oximetry measurement," https://doi.org/10.1056/NEJMc2029240, vol. 383, no. 25, pp. 2477–2478, Dec. 2020, doi: 10.1056/NEJMC2029240.

[48] P. A. Bothma et al., "Accuracy of pulse oximetry in pigmented patients," South African Med. J., vol. 86, no. 5 SUPPL., pp. 594–596, 1996.

[49] E. E. Foglia et al., "The effect of skin pigmentation on the accuracy of pulse oximetry in infants with hypoxemia," J. Pediatr., vol. 182, pp. 375-377.e2, Mar. 2017, doi: 10.1016/j.jpeds.2016.11.043.

[50] W. Cui, L. E. Ostrander, and B. Y. Lee, "In vivo reflectance of blood and tissue as a function of light wavelength," IEEE Trans. Biomed. Eng., vol. 37, no. 6, pp. 632–639, 1990, doi: 10.1109/10.55667.

[51] L. Yan, S. Hu, A. Alzahrani, S. Alharbi, and P. Blanos, "A multi-wavelength opto-electronic patch sensor to effectively detect physiological changes against human skin types," Biosensors, vol. 7, no. 2, Jun. 2017, doi: 10.3390/bios7020022.

[52] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," npj Digit. Med., vol. 3, no. 1, p. 18, Dec. 2020, doi: 10.1038/s41746-020-0226-6.

[53] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," J. Clin. Monit. Comput., vol. 27, no. 3, pp. 313–317, Jun. 2013, doi: 10.1007/s10877-013-9436-7.

[54] A. Shcherbina et al., "Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort," J. Pers. Med., vol. 7, no. 2, pp. 1–12, Jun. 2017, doi: 10.3390/jpm7020003.

[55] S. P. Preejith, A. Alex, J. Joseph, and M. Sivaprakasam, "Design, development and clinical validation of a wrist-based optical heart rate monitor," 2016 IEEE Int. Symp. Med. Meas. Appl. MeMeA 2016 - Proc., Aug. 2016, doi: 10.1109/MEMEA.2016.7533786.

[56] I. C. Jeong, H. Yoon, H. Kang, and H. Yeom, "Effects of skin surface temperature on photoplethysmograph," J. Healthc. Eng., vol. 5, no. 4, pp. 429–438, 2014, doi: 10.1260/2040-2295.5.4.429.

[57] K. Budidha and P. A. Kyriacou, "In vivo investigation of ear canal pulse oximetry during hypothermia," J. Clin. Monit. Comput., vol. 32, no. 1, pp. 97–107, Feb. 2018, doi: 10.1007/s10877-017-9975-4.

[58] A. C. Ralston, R. K. Webb, and W. B. Runciman, "Potential errors in pulse oximetry: I. pulse oximeter evaluation," Anaesthesia, vol. 46, no. 3, pp. 202–206, 1991, doi: 10.1111/j.1365-2044.1991.tb09410.x.

[59] Y. Maeda, M. Sekine, T. Tamura, A. Moriya, T. Suzuki, and K. Kameyama, "Comparison of reflected green light and infrared photoplethysmography," in Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare through Technology," 2008, pp. 2270–2272, doi: 10.1109/iembs.2008.4649649.

[60] Y. Maeda, M. Sekine, and T. Tamura, "The advantages of wearable green reflected photoplethysmography," J. Med. Syst., vol. 35, no. 5, pp. 829–834, Oct. 2011, doi: 10.1007/s10916-010-9506-z.

[61] B. Askarian, K. Jung, and J. W. Chong, "Monitoring of heart rate from photoplethysmographic signals using a Samsung Galaxy Note8 in underwater environments," Sensors (Switzerland), vol. 19, no. 13, p. 2846, Jul. 2019, doi: 10.3390/s19132846.

[62] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, and M. Sarlo, "Stressing the accuracy: wrist-worn wearable sensor validation over different conditions," Psychophysiology, vol. 56, no. 11, Nov. 2019, doi: 10.1111/psyp.13441.

[63] J. Fine et al., "Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring," Biosens. 2021, Vol. 11, Page 126, vol. 11, no. 4, p. 126, Apr. 2021, doi: 10.3390/BIOS11040126.

[64] B. W. Nelson, C. A. Low, N. Jacobson, P. Areán, J. Torous, and N. B. Allen, "Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research," npj Digit. Med., vol. 3, no. 1, pp. 1–9, Dec. 2020, doi: 10.1038/s41746-020-0297-4.

[65] S. Gillinov et al., "Variable accuracy of wearable heart rate monitors during aerobic exercise," Med. Sci. Sports Exerc., vol. 49, no. 8, pp. 1697–1703, Aug. 2017, doi: 10.1249/MSS.0000000000001284.

[66] R. Wang et al., "Accuracy of wrist-worn heart rate monitors," JAMA Cardiol., vol. 2, no. 1, pp. 104–106, Jan. 2017, doi: 10.1001/jamacardio.2016.3340.

[67] G. Cosoli, S. Spinsante, and L. Scalise, "Wrist-worn and chest-strap wearable devices: systematic review on accuracy and metrological characteristics," Measurement: Journal of the International Measurement Confederation, vol. 159. Elsevier B.V., p. 107789, Jul. 15, 2020, doi: 10.1016/j.measurement.2020.107789.

[68] L. Asare, E. Kviesis-Kipge, A. Grabovskis, U. Rubins, J. Spigulis, and R. Erts, "Multi-spectral photoplethysmography biosensor," in Optical Sensors 2011; and Photonic Crystal Fibers V, May 2011, vol. 8073, p. 80731Z, doi: 10.1117/12.887176.

[69] P. Blanos, S. Hu, D. Mulvaney, and S. Alharbi, "An applicable approach for extracting human heart rate and oxygen saturation during physical movements using a multi-wavelength illumination optoelectronic sensor system," in Society of Photo-Optical Instrumentation Engineers, Feb. 2018, p. 27, doi: 10.1117/12.2287854.

[70] S. Sen Gupta, S. Hossain, C. A. Haque, and K. D. Kim, "In-Vivo estimation of glucose level using PPG signal," Int. Conf. ICT Converg., vol. 2020-October, pp. 733–736, Oct. 2020, doi: 10.1109/ICTC49870.2020.9289629.

[71] K. M. Warren, J. R. Harvey, K. H. Chon, and Y. Mendelson, "Improving pulse rate measurements during random motion using a wearable

multichannel reflectance photoplethysmograph," Sensors (Switzerland), vol. 16, no. 3, Mar. 2016, doi: 10.3390/s16030342.

[72] J. Lee, M. Kim, H. K. Park, and I. Y. Kim, "Motion artifact reduction in wearable photoplethysmography based on multi-channel sensors with multiple wavelengths," Sensors (Switzerland), vol. 20, no. 5, Mar. 2020, doi: 10.3390/s20051493.

[73] C. C. Chang, C. T. Wu, B. Il Choi, and T. J. Fang, "MW-PPG sensor: an on-chip spectrometer approach," Sensors (Switzerland), vol. 19, no. 17, Sep. 2019, doi: 10.3390/s19173698.

[74] S.-H. Chen, Y.-C. Chuang, and C.-C. Chang, "Development of a portable all-wavelength PPG sensing device for robust adaptive-depth measurement: a spectrometer approach with a hydrostatic measurement example," Sensors 2020, Vol. 20, Page 6556, vol. 20, no. 22, p. 6556, Nov. 2020, doi: 10.3390/S20226556.

[75] M. Leier, K. Pilt, D. Karai, and G. Jervan, "Smart photoplethysmographic sensor for pulse wave registration at different vascular depths," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2015, vol. 2015-Novem, pp. 1849–1852, doi: 10.1109/EMBC.2015.7318741.

[76] Y. Maeda, M. Sekine, and T. Tamura, "Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography," in Journal of Medical Systems, Oct. 2011, vol. 35, no. 5, pp. 969–976, doi: 10.1007/s10916-010-9505-0.

[77] Y. Mendelson and C. Pujary, "Measurement site and photodetector size considerations in optimizing power consumption of a wearable reflectance pulse oximeter," in Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings, 2003, vol. 4, pp. 3016–3019, doi: 10.1109/iembs.2003.1280775.

[78] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors—past and present," Electron. , vol. 3, no. 2, pp. 282–302, 2014, doi: 10.3390/electronics3020282.

[79] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng, "Quantitative comparison of photoplethysmographic waveform characteristics: effect of measurement site," Front. Physiol., vol. 10, no. MAR, p. 198, Mar. 2019, doi: 10.3389/fphys.2019.00198.

[80] J. Liu, B. P. Yan, Y. T. Zhang, X. R. Ding, P. Su, and N. Zhao, "Multi-wavelength photoplethysmography enabling continuous blood pressure measurement with compact wearable electronics," IEEE Trans. Biomed. Eng., vol. 66, no. 6, pp. 1514–1525, Jun. 2019, doi: 10.1109/TBME.2018.2874957.

[81] S. K. Longmore, G. Y. Lui, G. Naik, P. P. Breen, B. Jalaludin, and G. D. Gargiulo, "A comparison of reflective photoplethysmography for detection of heart rate, blood oxygen saturation, and respiration rate at various anatomical locations," Sensors (Switzerland), vol. 19, no. 8, Apr. 2019, doi: 10.3390/s19081874.

[82] S. Lee, H. Shin, and C. Hahm, "Effective PPG sensor placement for reflected red and green light, and infrared wristband-type photoplethysmography," in International Conference on Advanced Communication Technology, ICACT, Mar. 2016, vol. 2016-March, pp. 556–558, doi: 10.1109/ICACT.2016.7423470.

[83] E. Geun, H. Heo, K. C. Nam, and Y. Huh, "Measurement site and applied pressure consideration in wrist photoplethysmography," 23rd Int. Tech. Conf. Circuits Systems Comput. Commun. ITCCSCC 2008, vol. 51, no. 3, pp. 1129–1132, 2008, doi: 10.1080/10635150290069913.

[84] Polar Research and Technology, "Polar precision prime OHR," Polar Res. Technol., 2019, Accessed: Mar. 30, 2021. [Online]. Available: https://www.polar.com/sites/default/files/static/science/white-papers/polar-precision-prime-white-paper.pdf.

[85] T. Rukasha, S. I. Woolley, and T. Collins, "Poster: heart rate performance of a medical-grade data streaming wearable device," in Proceedings - 2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2020, Dec. 2020, pp. 12–13, doi: 10.1145/3384420.3431776.

[86] P. J. Colvonen, P. N. DeYoung, N. O. A. Bosompra, and R. L. Owens, "Limiting racial disparities and bias for wearable devices in health science research," Sleep, vol. 43, no. 10. Sep. 07, 2020, doi: 10.1093/sleep/zsaa159.

[87] U.S. Food and Drug Administration, "510(k) premarket notification," U.S. Food and Drug Administration, 2021. www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K172935 (accessed Apr. 09, 2021).

[88] Fitbit, "Important safety and product information," Fitbit, 2021. www.fitbit.com/in/legal/safety-instructions (accessed Mar. 30, 2021).

[89] Garmin, "Accuracy," Garmin, 2021. www.garmin.com/en-US/legal/atdisclaimer/ (accessed Mar. 30, 2021).

[90] Apple, "Important safety information for Apple Watch - Apple Support," Apple, 2021. support.apple.com/en-gb/guide/watch/apdcf2ff54e9/watchos (accessed Mar. 30, 2021).

[91] Analog Devices, "ADPD188GG data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: www.analog.com/media/en/technical-documentation/data-sheets/adpd188gg.pdf.

[92] Analog Devices, "ADPD144RI data sheet," 2019. Accessed: Apr. 09, 2021. [Online]. Available: www.analog.com/media/en/technical-documentation/data-sheets/ADPD144RI.pdf.

[93] Maxim Integrated, "MAX30101 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: datasheets.maximintegrated.com/en/ds/MAX30101.pdf.

[94] Maxim Integrated, "MAX86150 data sheet," 2018. Accessed: Apr. 09, 2021. [Online]. Available: datasheets.maximintegrated.com/en/ds/MAX86150.pdf.

[95] Maxim Integrated, "MAX86916 data sheet," 2019. Accessed: Apr. 09, 2021. [Online]. Available: datasheets.maximintegrated.com/en/ds/MAX86916.pdf.

[96] OSRAM, "SFH 7072 OSRAM opto semiconductors," 2020. www.osram.com/ecat/BIOFY%C2%AE%20SFH%207072/com/en/class_pim_web_catalog_103489/prd_pim_device_2220016/ (accessed Apr. 09, 2021).

[97] OSRAM, "SFH 7050 OSRAM opto semiconductors," 2020. www.osram.com/ecat/BIOFY%C2%AE%20SFH%207050/com/en/class_pim_web_catalog_103489/prd_pim_device_2220012/ (accessed Apr. 09, 2021).

[98] Y. H. Kao, P. C. P. Chao, and C. L. Wey, "Design and validation of a new ppg module to acquire high-quality physiological signals for high-accuracy biomedical sensing," IEEE J. Sel. Top. Quantum Electron., vol. 25, no. 1, p. 8470924, Feb. 2019, doi: 10.1109/JSTQE.2018.2871604.

[99] H. J. Baek, S. Y. Sim, J. S. Kim, and K. S. Park, "Effect of sensor configurations on indirect-contact photoplethysmogram measurement system," in 2010 5th Cairo International Biomedical Engineering Conference, CIBEC 2010, 2010, pp. 244–246, doi: 10.1109/CIBEC.2010.5716070.

[100] B. Bent and J. P. Dunn, "Optimizing sampling rate of wrist-worn optical sensors for physiologic monitoring," J. Clin. Transl. Sci., pp. 1–8, 2020, doi: 10.1017/cts.2020.526.

[101] Analog Devices, "ADPD4000/ADPD4001 data sheet," 2019. Accessed: Apr. 09, 2021. [Online]. Available: www.analog.com/media/en/technical-documentation/data-sheets/ADPD4000-4001.pdf.

[102] Analog Devices, "ADPD4100/ADPD4101 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: www.analog.com/media/en/technical-documentation/data-sheets/adpd4100-4101.pdf.

[103] Maxim Integrated, "MAX30110 data sheet," 2017. Accessed: Apr. 09, 2021. [Online]. Available: datasheets.maximintegrated.com/en/ds/MAX30110.pdf.

[104] Maxim Integrated, "MAXM86146 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: datasheets.maximintegrated.com/en/ds/MAXM86146.pdf.

[105] Texas Instruments, "AFE4950 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: www.ti.com/lit/ds/symlink/afe4950.pdf?ts=1617894847053.

[106] Texas Instruments, "AFE44S30 data sheet," 2019. Accessed: Apr. 09, 2021. [Online]. Available: www.ti.com/lit/ds/symlink/afe44s30.pdf?ts=1617973955101.

[107] Texas Instruments, "AFE4900 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: www.ti.com/lit/ds/symlink/afe4900.pdf?ts=1617950577027.

[108] Texas Instruments, "AFE4404 data sheet," 2020. Accessed: Apr. 09, 2021. [Online]. Available: www.ti.com/lit/ds/symlink/afe4404.pdf?ts=1617893854742.

[109] X. F. Teng and Y. T. Zhang, "The effect of contacting force on photoplethysmographic signals," Physiol. Meas., vol. 25, no. 5, pp. 1323–1335, Oct. 2004, doi: 10.1088/0967-3334/25/5/020.

[110] K. Matsumura, P. Rolfe, J. Lee, and T. Yamakoshi, "iPhone 4s photoplethysmography: which light color yields the most accurate heart rate and normalized pulse volume using the iPhysioMeter application in the presence of motion artifact?," PLoS One, vol. 9, no. 3, Mar. 2014, doi: 10.1371/journal.pone.0091205.

[111] J. Liu, Y. Li, X. R. Ding, W. X. Dai, and Y. T. Zhang, "Effects of cuff inflation and deflation on pulse transit time measured from ECG and multi-wavelength PPG," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2015, vol. 2015-Novem, pp. 5973–5976, doi: 10.1109/EMBC.2015.7319752.

[112] J. Liu, B. P.-Y. Yan, W.-X. Dai, X.-R. Ding, Y.-T. Zhang, and N. Zhao, "Multi-wavelength photoplethysmography method for skin arterial pulse

extraction," Biomed. Opt. Express, vol. 7, no. 10, p. 4313, Oct. 2016, doi: 10.1364/boe.7.004313.
[113] Z. Geng, F. Tang, Y. Ding, S. Li, and X. Wang, "Noninvasive continuous glucose monitoring using a multisensor-based glucometer and time series analysis," Sci. Reports 2017 71, vol. 7, no. 1, pp. 1–10, Oct. 2017, doi: 10.1038/s41598-017-13018-7.
[114] S. Pasta et al., "A novel multi-wavelength procedure for blood pressure estimation using opto-physiological sensor at peripheral arteries and capillaries," in Society of Photo-Optical Instrumentation Engineers, Feb. 2018, p. 39, doi: 10.1117/12.2287845.
[115] S. Alharbi et al., "Oxygen saturation measurements from green and orange illuminations of multi-wavelength optoelectronic patch sensors," Sensors (Switzerland), vol. 19, no. 1, Jan. 2019, doi: 10.3390/s19010118.
[116] Y. Zhang et al., "Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths," Sensors (Switzerland), vol. 19, no. 3, Feb. 2019, doi: 10.3390/s19030673.
[117] V. P. Rachim and W. Y. Chung, "Wearable-band type visible-near infrared optical biosensor for non-invasive blood glucose monitoring," Sensors Actuators B Chem., vol. 286, pp. 173–180, May 2019, doi: 10.1016/J.SNB.2019.01.121.
[118] J. Liu et al., "PCA-based multi-wavelength photoplethysmography algorithm for cuffless blood pressure measurement on elderly subjects," IEEE J. Biomed. Heal. Informatics, pp. 1–1, Jun. 2020, doi: 10.1109/jbhi.2020.3004032.
[119] C. Y. Wang and K. T. Tang, "Active noise cancellation of motion artifacts in pulse oximetry using isobestic wavelength light source," in Proceedings - IEEE International Symposium on Circuits and Systems, 2011, pp. 1029–1032, doi: 10.1109/ISCAS.2011.5937744.
[120] S. Barbeiro, D. Libânio, R. Castro, M. Dinis-Ribeiro, and P. Pimentel-Nunes, "Narrow-band imaging: clinical application in gastrointestinal endoscopy," GE Port. J. Gastroenterol., vol. 26, no. 1, p. 40, Dec. 2018, doi: 10.1159/000487470.
[121] A. Adegbenro and S. Taylor, "Structural, physiological, functional, and cultural differences in skin of color," in Skin of Color, Springer New York, 2013, pp. 1–19.
[122] American Heart Association, "Cardiovascular disease a costly burden," Am. Hear. Assoc., vol. 91, pp. 399–404, 2017, doi: 1/17DS11775.
[123] B. Bent, O. M. Enache, B. Goldstein, W. Kibbe, and J. P. Dunn, "Reply: matters arising 'Investigating sources of inaccuracy in wearable optical heart rate sensors,'" npj Digital Medicine, vol. 4, no. 1. Nature Research, pp. 1–3, Dec. 01, 2021, doi: 10.1038/s41746-021-00409-4.
[124] C. Agu, "Petition: Medical schools must include BAME representation in clinical teaching," Change.org, 2020. www.change.org/p/gmc-medical-schools-must-include-bame-representation-in-clinical-teaching (accessed Mar. 30, 2021).
[125] M. Mukwende, "Mind the gap: A clinical handbook of signs and symptoms in black and brown skin," blackandbrownskin.co.uk, 2020. www.blackandbrownskin.co.uk/ (accessed Mar. 30, 2021).
[126] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.
[127] S. Fussell, "Why Can't This Soap Dispenser Identify Dark Skin?," GIZMODO, 2017. gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773 (accessed Mar. 30, 2021).
[128] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: large-scale heart rate estimation with convolutional neural networks," Sensor, vol. 19, no. 14, pp. 1–27, 2019, doi: 10.3390/s19143079.
[129] L. G. Rocha et al., "Real-time HR estimation from wrist PPG using binary LSTMs," in BioCAS 2019 - Biomedical Circuits and Systems Conference, Proceedings, Oct. 2019, pp. 1–4, doi: 10.1109/BIOCAS.2019.8918726.
[130] H. Chung, H. Ko, H. Lee, and J. Lee, "Deep learning for heart rate estimation from reflectance photoplethysmography with acceleration power spectrum and acceleration intensity," IEEE Access, vol. 8, pp. 63390–63402, 2020, doi: 10.1109/ACCESS.2020.2981956.
[131] C. El-Hajj and P. A. Kyriacou, "A review of machine learning techniques in photoplethysmography for the non-invasive cuffless measurement of blood pressure," Biomedical Signal Processing and Control, vol. 58. Elsevier Ltd, p. 101870, Apr. 01, 2020, doi: 10.1016/j.bspc.2020.101870.
[132] S. I. Woolley, T. Collins, J. Mitchell, and D. Fredericks, "Investigation of wearable health tracker version updates," BMJ Heal. Care Informatics, vol. 26, no. 1, p. 100083, Oct. 2019, doi: 10.1136/bmjhci-2019-100083.

Daniel Ray is currently pursuing Ph.D. in the Department of Engineering at Manchester Metropolitan University, UK. His research interests include nature-inspired computing, wearable computing, health technologies and the ethics of technology. He has a MSc in Data Science at Manchester Metropolitan University, UK and a BSc in Computing at the University of Liverpool, UK. He is a Student Member of the IEEE and the British Computer Society.

Tim Collins is a senior lecturer at Manchester Metropolitan University, UK, where he is the programme leader for Electrical and Electronic Engineering. He has a PhD in Electronic and Electrical Engineering from the University of Birmingham, UK, and is a Chartered Engineer (IET) and a Senior Member of the IEEE. His research interests are in the utilization of signal processing and image processing techniques, including machine learning, to applications including acoustics, communications, 3D computational geometry, extended reality and health technology.

Sandra I. Woolley leads Software and Systems Engineering Research at Keele University, UK. She has a PhD in Electrical Engineering from Manchester University and is a Senior Member of the IEEE and a Fellow of the British Computer Society. She has long-standing research interests in wearable computing, sensing systems, and health technologies that encompass performance evaluations, patient monitoring applications, and system usability. She teaches computer science masters and undergraduate modules in User Interaction Design, Animation and Multimedia, and Professionalism and she chairs Faculty Research and School Taught Programme Ethics Committees.

Prasad V. S. Ponnapalli is a senior lecturer at Manchester Metropolitan University, UK. He teaches Control Engineering, Industrial Automation and Cyber-Physical Systems at UG and PG level. His research interests are intelligent control, industrial automation and applications of AI. He has a PhD in Electronic and Electrical Engineering from the Queen's University of Belfast, UK, and is a Member of the IEEE and the IET.

A.2 DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography

Ray, D., Collins, T., & Ponnapalli, P. V. S. (2022). DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography. Conference Proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2022, 1651–1654. https://doi.org/10.1109/EMBC48229.2022.9871813

DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography

Daniel Ray, Tim Collins, and Prasad V. S. Ponnapalli

***Abstract*— Wearable Photoplethysmography (PPG) has gained prominence as a low cost, unobtrusive and continuous method for physiological monitoring. The quality of the collected PPG signals is affected by several sources of interference, predominantly due to physical motion. Many methods for estimating heart rate (HR) from PPG signals have been proposed with Deep Neural Networks (DNNs) gaining popularity in recent years. However, the "black-box" and complex nature of DNNs has caused a lack of trust in the predicted values. This paper contributes DeepPulse, an uncertainty-aware DNN method for estimating HR from PPG and accelerometer signals, with aims of increasing trust of the predicted HR values. To the best of the authors' knowledge no PPG HR estimation method has considered aleatoric and epistemic uncertainty metrics. The results show DeepPulse is the most accurate method for DNNs with smaller network sizes. Finally, recommendations are given to reduce epistemic uncertainty, validate uncertainty estimates, improve the accuracy of DeepPulse as well as reduce the model size for resource-constrained edge devices.**

I. INTRODUCTION

Wrist-worn reflectance mode PPG sensing is popular in many wearable devices as it provides a means of low cost, unobtrusive and continuous physiological monitoring [1]. The performance of PPG sensing is affected by several sources of interference including biological characteristics, sensor configuration and placement as well as ambient light [1]. However, the main source of interference is physical motion which distorts the collected PPG signal. The removal of motion artefacts from the signal is a challenge due to overlapping frequency bands and amplitudes much larger than the pulsatile component of the signal [1], [2].

Computational methods for estimating HR from PPG signals consist of four main steps: prepossessing, de-noising, heart rate estimation and heart rate tracking [2]. A common approach used across existing methods for de-noising is to incorporate a motion reference sensor, such as a triaxial accelerometer or gyroscope, in order to capture motion data at the measurement site and compensate for the interference the motion causes [3], [4]. Many conventional signal processing approaches to HR estimation rely on expert-tuned parameters [3] leading to difficulties in generalizing the methods [4], [5]. In order to prevent this, researchers have explored the use of deep neural networks (DNNs) for HR estimations [4]–[9]. Although the performance improvements are significant, DNNs for edge devices have their own challenges including data asymmetry, multi-modality sensing and resource constraints of edge devices [10].

Classical approaches to the fusion of heterogeneous sensing modalities rely on feature engineering to extract independent features from each sensing modality which are then fused together. This approach of extracting different features from individual sensors disregards features that use multiple sensors' data to capture information that neither has in isolation [11]. In many applications, DNNs have been adopted instead due to their ability to learn to extract features during training [11]–[13] showing improved performance in applications such as gait recognition [11], human activity recognition [11]–[13], car tracking [12], dynamic gas mixtures estimations [13] and cuffless blood pressure monitoring [13].

One major drawback to the use of DNNs is a lack of trust in the predicted values due to high complexity and uninterpretability of the generated DNNs, mainly from deep and non-linear structures [13]. In order to increase trust in the predicted values of DNNs researchers have explored ways to represent uncertainty within DNNs [13]–[16]. The two main sources of uncertainty are "aleatoric" and "epistemic". Aleatoric uncertainty describes the irreducible uncertainty in the input data due to an inherent property of the data distribution such as randomness or noise [14]. Epistemic uncertainty describes uncertainty in the model that occurs due to inadequate data which may be reduced by increasing the amount and 'diversity' of the training data [14].

Researchers have explored several methods to incorporate and quantify uncertainty in DNNs such as Monte Carlo Dropout (MCDropout), Variational Inference and Ensemble methods [14], [16]. The uncertainty framework proposed by [16] is advantageous as it requires little modification to existing DNNs [14]. The framework uses MCDropout with an aleatoric uncertainty term to simultaneously estimate aleatoric and epistemic uncertainty, showing promising results in several applications [15], [16]. MCDropout has been theorized to approximate Gaussian processes by activating dropout layers during the prediction phase to provide an ensemble of predictions [16]. The variability of the ensemble predictions distribution quantifies epistemic uncertainty [15], [16]. In order to incorporate aleatoric uncertainty, a second output unit is added to the DNN with a specially-designed loss function such as negative log likelihood (NLL). The two output units of the DNN estimate μ and σ of a distribution, where μ represents the mean value of the distribution and σ represents the standard deviation of the distribution used to

This research received no external funding.

D. Ray, T. Collins and P. V. S. Ponnapalli are with the Department of Engineering, Manchester Metropolitan University, Manchester, UK. (e-mail: Daniel.Ray@stu.mmu.ac.uk, T.Collins@mmu.ac.uk and P.Ponnapalli@mmu.ac.uk)

quantify aleatoric uncertainty [15].

II. METHODOLOGY

A. Datasets

1) IEEE SPC 2015: consists of two datasets that employed different protocols, namely IEEE Train and IEEE Test. Both datasets were collected using a green (515 nm) reflectance mode PPG sensor as well as a single lead chest-worn ECG. IEEE Train collected 12 sessions whilst IEEE Test collected 10 sessions. Both datasets employed laboratory-based protocols with IEEE Train using a treadmill and IEEE Test focusing on arm movements, with each session duration being no longer than 15 minutes [3].

2) PPG-DaLiA: was collected using an Empatica E4 wrist-worn reflectance mode PPG sensor used green (520 nm) and red (660 nm) LEDs and a 3-lead chest-worn ECG. A total of 15 sessions were collected using a naturalistic protocol of various daily activities with each session duration being more than 1.5 hours long [4].

3) BAMI-II: was collected using a wrist-worn reflectance mode green (525 nm) PPG sensor and a medical-grade 3-lead chest-worn ECG Holter monitor. A total of 24 sessions were collected employing a laboratory-based protocol using a treadmill with each session duration being 14 minutes [17].

B. Preprocessing and Learning Strategy

The PPG and accelerometer signals were first subject to a 2^{nd} order Butterworth band-pass filter with cutoff frequencies of 0.5 Hz - 4.5 Hz to remove components of the signals outside the range of cardiac activity. The signals were then re-sampled to 64 Hz and normalized to zero mean and unit variance. Finally, a sliding window was applied to the signals with a window length of 8 seconds and a 2 second slide. To reduce the effects of data asymmetry a leave-one-session-out (LOSO) cross-validation scheme was employed [4], [6] where each session was used as test data exactly once. A more detailed explanation of the implemented LOSO cross-validation scheme can be found in [4].

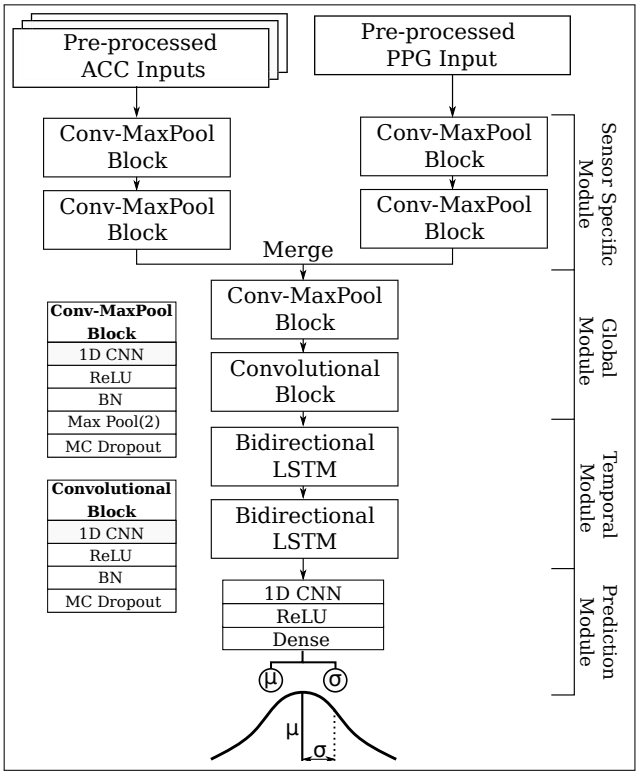


Fig. 1. The Architecture of DeepPulse.

TABLE I
HYPERPARAMETERS OF DEEPPULSE ARCHITECTURE

Sensor-specific Module	
Number of Conv. Filters	64
Global Module	
Number of Conv. Filters	128
Temporal Module	
Number of LSTM Units	32
Network Parameters	
All Conv. Blocks: Convolutional Kernel Size	16
Merge Type	Concatenate Axis = 2
Dropout Rate	0.15
Optimizer	Nadam

C. DeepPulse Architecture and Implementation

DeepPulse contains four main architectural sections: sensor-specific module, global module, temporal module and prediction module (Figure 1). The convolutional blocks in the sensor-specific module extract local interactions within each sensing modality. The sensor-specific features are then merged together and passed through the convolutional blocks in the global module to extract global features. The global features are then used in the temporal module to extract temporal features using bidirectional Long Short-Term Memory (LSTM) layers. The temporal features are passed to the prediction module which contains a convolutional layer to reduce the dimensionality of the features for the fully connected layer. The selected hyper-parameters of the architectural components and network parameters can be found in Table I. A NLL loss function was used to evaluate how the DNN models the data in terms of both accuracy and aleatoric uncertainty. Each convolutional block contains a MCDropout layer used to produce an ensemble of predictions (T=10) for each input to evaluate epistemic uncertainty.

The training phase of DeepPulse was run for 200 epochs with a batch size of 32. During the training phase, early stopping was employed aiming at reducing the risk of overfitting and the learning rate was reduced when the learning had stagnated. DeepPulse was implemented using Tensorflow (Version: 2.7.0) and Tensorflow Probability (Version: 0.14.1). Computation was carried out using 8 Intel Broadwell CPU cores and a NVIDIA Tesla K80 GPU (CUDA Version: 11.2). The implementation of DeepPulse can be found at: `https://github.com/danielray54/DeepPulse`

D. Evaluation Metrics

Mean absolute error (MAE) was employed to assess the accuracy. Predicted values were averaged across all LOSO iteration to obtain a generalized MAE. Additionally, two uncertainty metrics were employed. $u_a(x_i)$ is the aleatoric uncertainty (Equation 1) which is the average of the squared $\sigma_{i,t}$ output unit for an ensemble of predictions, T, for each input window x_i:

$$u_a(x_i) = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_{i,t}^2 \tag{1}$$

u_e is the epistemic uncertainty (Equation 2) which is the variance computed from the predicted mean values $\mu_{i,t}$ from the ensemble of predictions, T, for each input window x_i:

$$u_e(x_i) = \frac{1}{T}\sum_{t=1}^{T}\mu_{i,t}^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\mu_{i,t}\right)^2 \tag{2}$$

III. RESULTS

A. Accuracy & Complexity

The MAE results show that DeepPulse is the second most accurate method of all DNN PPG HR estimation methods for all datasets (Table II). However, when comparing methods with smaller model sizes, under 1 million parameters, (Table III) DeepPulse is the most accurate for all datasets. This is significant as models with large complexities have not accounted for the resource constraints of edge devices [10].

TABLE II
COMPARISON OF MEAN ABSOLUTE ERRORS FOR DNN PPG HR ESTIMATION METHODS

Method	**Datasets**			
	IEEE Train	**IEEE Test**	**PPG-DaLiA**	**BAMI-II**
Deep PPG [4]	4.00 ±5.40	16.51 ±16.10	7.65 ±4.20	N/A
CorNET (LOSO) [6]	4.67 ±3.71	6.61 ±5.35	N/A	N/A
Binary CorNET [6]	6.20 ±4.95	7.31 ±6.14	N/A	N/A
PPGnet [7]	3.36 ±4.10	12.48 ±14.45	N/A	N/A
Chung *et al.* [8]	0.67 ±0.50	0.86 ±0.80	N/A	1.46 ±1.23
MH Conv-LSTM DeepPPG [9]	N/A	N/A	6.28 ±3.53	N/A
DeepPulse	2.76 ±2.95	5.05 ±5.50	2.12 ±3.09	2.38 ±2.57

All values are BPM.

B. Uncertainty

For the IEEE datasets, as the number of input windows per activity decreases the epistemic uncertainty estimates increase (Figure 2(a)). This supports the hypothesis that increasing the size and 'diversity' of the dataset will reduce the epistemic uncertainty. Assuming more intense activity or higher BPM values require more movement from the body thus more noise in the PPG signals then as either BPM values or activity intensity increase so will the aleatoric uncertainty estimates which is shown for the BAMI-II and PPG-DaLiA datasets in Figure 2(b) & 2(c). Finally, Figure 2(d) illustrates that there is little to no relationship between between aleatoric uncertainty and epistemic uncertainty estimates for the BAMI-II dataset.

TABLE III
COMPARISON OF NETWORK COMPLEXITIES FOR DNN PPG HR ESTIMATION METHODS

Method	**Number of Parameters**
Deep PPG [4]	8.5M
CorNET [5]	250K
PPGnet [7]	765K
Chung *et al.* [8]	3.3M
MH Conv-LSTM DeepPPG [9]	680K
DeepPulse	730K

IV. FUTURE WORK

The performance of PPG sensing is affected by several sources of interference and inaccuracies. Some of these sources such as skin tone, skin temperature, age, sex and BMI have not been fully considered in the datasets used. Increasing the size and 'diversity' of the data will be beneficial in improving the accuracy, robustness and generalizability [1] as well as epistemic uncertainty of DNN PPG HR estimation algorithms. Moreover, ensuring that the collected "truth values" are an accurate depiction of the cardiac activity is essential which can achieved by using medically validated chest-worn ECG devices [1].

In order to improve the performance and reduce the model size of DeepPulse, hyperparameter optimization and network architecture search should be carried out [18]. Additionally, weight clustering and model quantization may prove to be effective methods to further reduce the model size.

Finally, further improvement to the accuracy of DeepPulse may be made by introducing a post-processing step that averages predicted values of several input windows when the aleatoric uncertainty is high. To better evaluate aleatoric uncertainty, an accurate signal-to-noise ratio method should be developed to eliminate assumption made based on activity type. Additionally, to validate the epistemic uncertainty estimates, training DeepPulse on subsets of the datasets would provide more insight. Similarly, adding noise to the input windows would enable the validation of the aleatoric uncertainty estimates.

V. CONCLUSION

Wearable Photoplethysmography (PPG) has gained prominence as a method for physiological monitoring but is subject to several sources of interference making the estimation of HR challenging. DNNs have gained popularity in recent years with promising results. However, the "black-box" and complex nature of DNNs has caused a lack of trust in

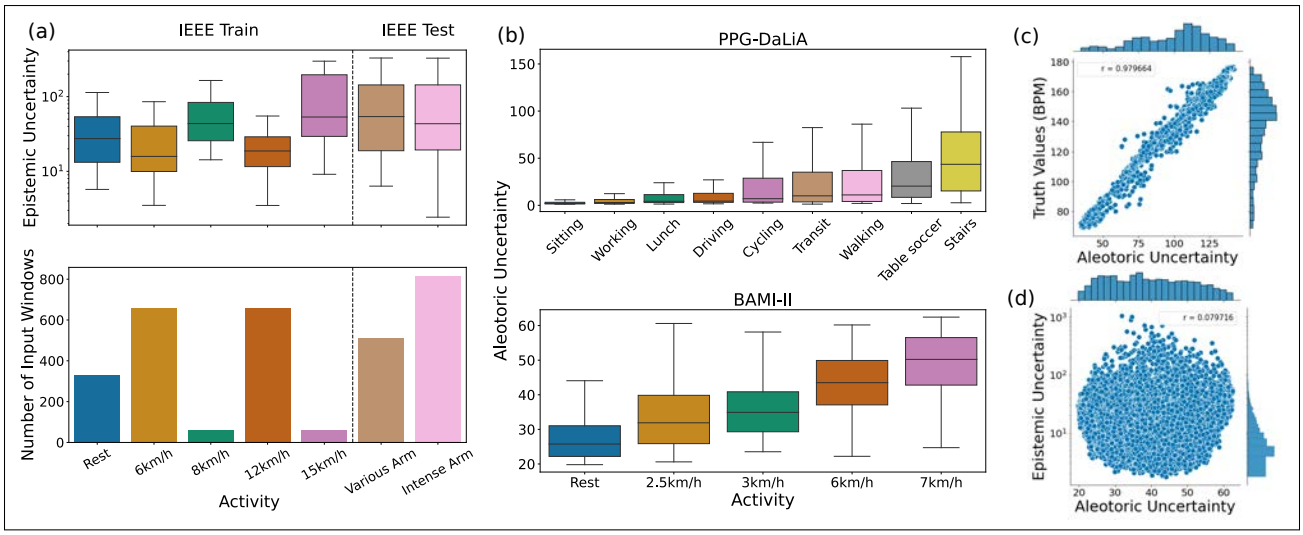


Fig. 2. (a) shows the relationship between epistemic uncertainty and the number of input windows for each activity in both of the IEEE datasets, (b) shows the relationship between aleatoric uncertainty and activity in the BAMI-II and PPG-DaLiA datasets, (c) shows the relationship between aleatoric uncertainty and truth values in BAMI-II dataset and (d) shows the relationship between aleatoric and epistemic uncertainty in the BAMI-II dataset.

the predicted values. This paper contributes DeepPulse, a multimodal uncertainty-aware DNN method for estimating HR from PPG and accelerometer signals. The results show DeepPulse is the most accurate method for DNNs with less than 1 million network parameters. Finally, recommendations have been given to improve the accuracy and reduce the complexity of DeepPulse for resource-constrained edge devices as well as reduce and validate uncertainty estimates.

References

[1] D. Ray, T. Collins, S. Woolley and P. Ponnapalli, "A Review of Wearable Multi-wavelength Photoplethysmography," in IEEE Reviews in Biomedical Engineering, doi: 10.1109/RBME.2021.3121476.

[2] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation," Archives of Computational Methods in Engineering, vol. 1. p. 3, 2021, doi: 10.1007/s11831-021-09597-4.

[3] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," IEEE Trans. Biomed. Eng., vol. 62, no. 2, pp. 522–531, 2015, doi: 10.1109/TBME.2014.2359372.

[4] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks," Sensor, vol. 19, no. 14, pp. 1–27, 2019, doi: 10.3390/s19143079.

[5] D. Biswas et al., "CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment," IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 2, pp. 282–291, 2019, doi: 10.1109/TBCAS.2019.2892297.

[6] L. G. Rocha et al., "Binary CorNET: Accelerator for HR Estimation from Wrist-PPG," IEEE Trans. Biomed. Circuits Syst., vol. 14, no. 4, pp. 715–726, Aug. 2020, doi: 10.1109/TBCAS.2020.3001675.

[7] A. Shyam, V. Ravichandran, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 1899–1902, 2019, doi: 10.1109/EMBC.2019.8856989.

[8] H. Chung, H. Ko, H. Lee, and J. Lee, "Deep Learning for Heart Rate Estimation from Reflectance Photoplethysmography with Acceleration Power Spectrum and Acceleration Intensity," IEEE Access, vol. 8, pp. 63390–63402, 2020, doi: 10.1109/ACCESS.2020.2981956.

[9] M. Wilkosz and A. Szczęsna, "Multi-Headed Conv-LSTM Network for Heart Rate Estimation during Daily Living Activities," Sensors, vol. 21, no. 15, p. 5212, Jul. 2021, doi: 10.3390/s21155212.

[10] X. Zeng, "Mobile sensing through deep learning," in MobiSys 2017 PhD Forum - Proceedings of the 2017 Workshop on MobiSys 2017 Ph.D. Forum, co-located with MobiSys 2017, Jun. 2017, vol. 9908 LNCS, pp. 5–6, doi: 10.1145/3086467.3086476.

[11] V. Radu et al., "Multimodal Deep Learning for Activity and Context Recognition," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 1, no. 4, pp. 1–27, Jan. 2018, doi: 10.1145/3161174.

[12] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing," 26th Int. World Wide Web Conf. WWW 2017, pp. 351–360, Nov. 2016, doi: 10.1145/3038912.3052577.

[13] S. Yao et al., "RDeepSense: Reliable Deep Mobile Computing Models with Uncertainty Estimations," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, Jan. 2018, doi: 10.1145/3161181.

[14] O. Dürr, B. Sick, and E. Murina, Probabilistic Deep Learning: with Python, Keras and Tensorflow Probability. Manning Publications, 2020.

[15] K. Fang, D. Kifer, K. Lawson, and C. Shen, "Evaluating the Potential and Challenges of an Uncertainty Quantification Method for Long Short-Term Memory Models for Soil Moisture Predictions," Water Resour. Res., vol. 56, no. 12, p. e2020WR028095, Dec. 2020, doi: 10.1029/2020WR028095.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in 33rd International Conference on Machine Learning, ICML 2016, Jun. 2016, vol. 3, pp. 1651–1660, Accessed: Jan. 10, 2022. [Online]. Available: https://arxiv.org/abs/1506.02142v6.

[17] H. Lee, H. Chung, and J. Lee, "Motion Artifact Cancellation in Wearable Photoplethysmography Using Gyroscope," IEEE Sens. J., vol. 19, no. 3, pp. 1166–1175, Feb. 2019, doi: 10.1109/JSEN.2018.2879970.

[18] D. Ray, T. Collins, and P. Ponnapalli, "Deep Neural Network Architecture Search for Wearable Heart Rate Estimations," Stud. Health Technol. Inform., vol. 281, pp. 1106–1107, May 2021, doi: 10.3233/SHTI210366.

A.3 Towards Wrist-worn Photoplethysmography Sensing for Medical Applications

Ray, D. (2021). Towards Wrist-worn Photoplethysmography Sensing for Medical Applications. 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), 432–433. https://doi.org/10.1109/ICHI52183.2021.00070

Towards Wrist-worn Photoplethysmography Sensing for Medical Applications

Daniel Ray
Department of Engineering
Manchester Metropolitan University
Manchester, UK
Email: Daniel.Ray@stu.mmu.ac.uk

I. Problem Statement

Photoplethysmography (PPG) sensing takes advantage of hemoglobins' absorbent qualities to visible and infrared (IR) light. Consisting of a light source and a photo-detector, light is emitted into the skin and the intensity of light transmitted into the photo-detector is inversely proportional to the volume of blood in the vascular bed of the measurement site [1], [2].

PPG sensing is widely used in clinical settings for peripheral pulse oximetry measurements (SpO_2). Using the transmittive mode, where the light source and photo-detector are separated by tissue, two wavelengths of light usually red and IR are used to measure SpO_2 on the fingertip. Wrist-worn consumer health monitoring devices have adopted the reflectance mode, where the light source and photo-detector are positioned along side each other, typically using green light to measure heart rate (HR) on the wrist [1], [2].

Selecting the wavelength(s) of light used for PPG sensing is a trade-off and based on the application and measurement site as visible and IR light interact with compounds in human skin and blood to varying degrees. Shorter wavelengths such as blue, green and yellow are highly absorbent to oxyhemoglobin (HbO_2), deoxyhemoglobin (Hb) and melanin and penetrate the skin to a lesser degree than longer wavelengths. Red light is more absorbent to Hb whilst IR is more absorbent to HbO_2 but both produce more complex and sometimes noisy signals [2]. Additionally, exposure to differing temperatures causes varying amounts of blood in the peripheral circulation [3].

The accuracy, robustness & generalizability of wrist-worn PPG sensing is adversely affected by several factors such as: motion, skin-melanin content, skin temperature, sensor geometry and ambient light [2]–[5]. These factors cause the acquired signal to attenuate or contain artifacts making the extraction of accurate physiological measurements challenging [4], [5].

II. A Brief Review of Multi-wavelength PPG Hardware Solutions

Initial multi-wavelength PPG sensing hardware was reliant on fiber optics [6] progressing into patch sensor development [7] due to its low cost and simple form factor. Researchers have also explored ear-worn PPG sensors [8], however wrist-worn devices have a pre-existing cultural acceptance and provide the most convenient and unobtrusive solution. The latest developments in multi-wavelength PPG sensing hardware is an on-chip spectrometer approach based on plasmonic filters [9].

Research suggests that consumer-grade wrist-worn multi-wavelength PPG devices are more accurate at rest than research-grade devices [5], with the most accurate commercial device being the Apple watch producing a Mean Average Error (MAE) of 4.4BPM and 2.7% missingness [5]. Research-grade devices such as Empatica E4 and Biovotion Everion provide data-streamed raw signals, however, the E4 accuracy has been reported to be lower than the consumer-grade devices with Mean Absolute Percentage Errors of 7.2% and 29.2% whilst collecting data on a treadmill [10].

Concerns have arisen due to both consumer and research-grade devices being used in clinical trials, with Fitbit alone having 476 published studies and 449 studies registered on ClinicalTrials.gov [5]. Fitbit, Garmin and Apple all state that their optical heart rate monitors should not be used as medical devices with intent to diagnose, treat, cure or prevent any disease [11]–[13].

III. A Brief Review of PPG Heart Rate Estimation Algorithms

Extracting Heart Rate from PPG signals is challenging in periods of motion due to overlapping frequency bands, with HR frequency range typically being 0-4.5Hz whilst the motion frequency range is within 0-10Hz. Researchers' have employed several approaches with varying results. Statistical approaches such as a correlation-based spectral analysis and harmonic noise dampening produced a MAE of 1.32 ± 1.24BPM [14]. Matrix manipulation techniques such as singular value decomposition have also seen promising results such as in period of intense motion a MAE of 2.92BPM was achieved [15].

Researchers have also explored deep learning methods such as using Convolutional networks [16] as well as a combination of Convolutional and Long short-term memory Networks [17]. In periods of intense activities the Convolutional network method outperformed the classical methods by a MAE of 8.91BPM [16]. However, the generalizability of the methodologies stated above has been called into question as the datasets used to evaluate the performance lacks

proportional representation of skin-melanin content with a skew towards lower skin-melanin content.

IV. Existing & Proposed Work

This research program seeks to investigate, develop and evaluate wrist-worn PPG methodologies that have increased generalizability and robustness to skin characteristics and motion artifacts.

1) **Hardware:** Building on research into sensor placement & arrangement, wavelengths of light selection, light intensity and wrist measurement sites a modular multi-sensor, multi-wavelength wrist-worn PPG device will be constructed and evaluated.
2) **Data Collection:** Current research and datasets lack proportional representation of skin-melanin content. To the best of the authors' knowledge, there is no research exploring the compounded effects of skin-melanin content, skin temperature and motion. The proposed dataset will include both naturalistic and laboratory based collection protocols. A Holter monitor (3-Lead ECG) will be used to collect truth values. Existing commercial multi-wavelength wrist-worn devices will also be used to evaluate the proposed methodologies.
3) **Algorithms:** Adapting and building upon single-wavelength PPG HR estimation algorithms to multi-wavelength PPG data using statistical, signal processing, linear algebra and deep learning learning approaches. Currently, Network Architecture Search algorithms, such as Jin *et al.* [18], are being explore to create elaborate and complex deep learning networks for HR estimations on single-wavelength PPG data which will then be adapted to multi-wavelength PPG data once collected.

V. Expected Contributions & Impact

The expected contributions this project intends to produce upon completion can be separated into three parts: a hardware solution, a diverse dataset in both subjects & activities as well as several single-wavelength and multi-wavelength HR estimation algorithms.

Surrounding disciplines such as medicine and computer science both have cases of racial bias, due to lack of proportional representation in validation studies [19], [20], therefore it is necessary that the influence is made clear. Wrist-worn PPG sensing must work accurately regardless of the individuals' skin-melanin content when used in medical applications.

References

[1] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care", Int. J. Biosens. Bioelectron., vol. 4, no. 4, 2018, doi: 10.15406/ijbsbe.2018.04.00125.

[2] M. Lemay, M. Bertschi, J. Sola, P. Renevey, J. Parak, and I. Korhonen, "Application of Optical Heart Rate Monitoring", in Wearable Sensors: Fundamentals, Implementation and Applications, Elsevier Inc., 2014, pp. 105–129.

[3] Y. Maeda, M. Sekine, T. Tamura, A. Moriya, T. Suzuki, and K. Kameyama, "Comparison of reflected green light and infrared photoplethysmography", in Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare through Technology," 2008, pp. 2270–2272, doi: 10.1109/iembs.2008.4649649.

[4] B. W. Nelson, C. A. Low, N. Jacobson, P. Areán, J. Torous, and N. B. Allen, "Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research", npj Digit. Med., vol. 3, no. 1, pp. 1–9, 2020, doi: 10.1038/s41746-020-0297-4.

[5] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors", npj Digit. Med., vol. 3, no. 1, p. 18, 2020, doi: 10.1038/s41746-020-0226-6.

[6] L. Asare, E. Kviesis-Kipge, A. Grabovskis, U. Rubins, J. Spigulis, and R. Erts, "Multi-spectral photoplethysmography biosensor", in Optical Sensors 2011; and Photonic Crystal Fibers V, 2011, vol. 8073, p. 80731Z, doi: 10.1117/12.887176.

[7] P. Blanos, S. Hu, D. Mulvaney, and S. Alharbi, "An applicable approach for extracting human heart rate and oxygen saturation during physical movements using a multi-wavelength illumination optoelectronic sensor system", in Society of Photo-Optical Instrumentation Engineers, 2018, p. 27, doi: 10.1117/12.2287854.

[8] K. Budidha and P. A. Kyriacou, "In vivo investigation of ear canal pulse oximetry during hypothermia", J. Clin. Monit. Comput., vol. 32, no. 1, pp. 97–107, Feb. 2018, doi: 10.1007/s10877-017-9975-4.

[9] C. C. Chang, C. T. Wu, B. Il Choi, and T. J. Fang, "MW-PPG sensor: An on-chip spectrometer approach", Sensors (Switzerland), vol. 19, no. 17, 2019, doi: 10.3390/s19173698.

[10] T. Rukasha, S. I. Woolley, and T. Collins, "Poster: Heart Rate Performance of a Medical-Grade Data Streaming Wearable Device," in Proceedings - 2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2020, 2020, pp. 12–13, doi: 10.1145/3384420.3431776.

[11] Apple, "Important safety information for Apple Watch - Apple Support," Apple, 2021. [Online]. Available: https://support.apple.com/en-gb/guide/watch/apdcf2ff54e9/watchos. [Accessed: 07-Mar-2021].

[12] Fitbit, "Important Safety and Product Information," Fitbit, 2021. [Online]. Available: https://www.fitbit.com/in/legal/safety-instructions. [Accessed: 07-Mar-2021].

[13] Garmin, "Accuracy," Garmin, 2021. [Online]. Available: https://www.garmin.com/en-US/legal/atdisclaimer/. [Accessed: 07-Mar-2021].

[14] T. Schäck, M. Muma, and A. M. Zoubir, "Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals," in 25th European Signal Processing Conference, EUSIPCO 2017, Oct. 2017, vol. 2017-Janua, pp. 2478–2481, doi: 10.23919/EUSIPCO.2017.8081656.

[15] A. Galli, G. Frigo, C. Narduzzi, and G. Giorgi, "Robust estimation and tracking of heart rate by PPG signal analysis," Jul. 2017, doi: 10.1109/I2MTC.2017.7969715.

[16] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks," Sensor, vol. 19, no. 14, pp. 1–27, 2019, doi: 10.3390/s19143079.

[17] A. Shyam, V. Ravichandran, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram," Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, pp. 1899–1902, 2019, doi: 10.1109/EMBC.2019.8856989.

[18] H. Jin, Q. Song, and X. Hu, "Auto-Keras: An Efficient Neural Architecture Search System," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul. 2019, pp. 1946–1956, doi: 10.1145/3292500.3330648.

[19] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science (80-.)., vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[20] M. Mukwende, "Mind the gap: A clinical handbook of signs and symptoms in black and brown skin," blackandbrownskin.co.uk, 2020. [Online]. Available: https://www.blackandbrownskin.co.uk/. [Accessed: 07-Mar-2021].

A.4 Deep Neural Network Architecture Search for Wearable Heart Rate Estimations

Ray, D., Collins, T., & Ponnapalli, P. (2021). Deep Neural Network Architecture Search for Wearable Heart Rate Estimations. Studies in Health Technology and Informatics, 281, 1106–1107. https://doi.org/10.3233/SHTI210366

Deep Neural Network Architecture Search for Wearable Heart Rate Estimations

Daniel RAY [a,1], Tim COLLINS [a] and Prasad PONNAPALLI [a]
[a] *Manchester Metropolitan University, UK*

Abstract. Extracting accurate heart rate estimations from wrist-worn photoplethysmography (PPG) devices is challenging due to the signal containing artifacts from several sources. Deep Learning approaches have shown very promising results outperforming classical methods with improvements of 21% and 31% on two state-of-the-art datasets. This paper provides an analysis of several data-driven methods for creating deep neural network architectures with hopes of further improvements.

Keywords. wearable, heart rate, photoplethysmography, deep neural networks, network architecture search

1. Introduction

Wrist-worn PPG heart rate monitors provide an unobtrusive and low-cost method for continuous heart rate measurements, widely adopted in both commercial and clinical settings [1,2]. Researchers have identified several factors, such as motion, skin characteristics and ambient light, that cause the acquired signal to contain artifacts making the extraction of accurate heart rate estimations challenging [2]. Deep Neural Networks (DNN) have seen promising results as a method to accurately estimate heart rate from signals that contain artifacts [1]. However, applying a data-driven approach may improve on state-of-the-art deep learning methods by creating elaborate and complex network architectures that would be un-achievable for machine learning engineers to create due to the enormous search space of compounded architectural components.

2. Methods

Several Network Architecture Search strategies will be applied to the current state-of-the-art PPG datasets [1,3,4] to establish an architecture that improves upon the accuracy achieved in [1]. These strategies include Reinforcement Learning [5], NeuroEvolution of Augmenting Typologies [6] and a Bayesian Optimization Network Morphing method [7]. The current standard is to use both PPG and 3-axis accelerometer, as a motion reference, signals which are segmented into overlapping

[1]Corresponding Author: John R. Dalton Building, All Saints Campus, Manchester Metropolitan University, Manchester, UK, M15 6BH; E-mail: daniel.ray@stu.mmu.ac.uk

sliding windows [1]. The 'truth' values are recorded using a chest-worn Electrocardiography device. Each window is then transformed into the frequency domain using a Fourier Transform, then inputted into the network during the training process using Mean Absolute Error (MAE) as comparable metric.

3. Results

Preliminary results show the method of Network Morphing [7] created an architecture that achieved a MAE of 18.2 BPM after 10 iterations. These results would be expected to dramatically improve as more iterations are carried out, allowing for more complex and elaborate architectures to be created.

4. Discussion

Establishing accurate methods for heart rate extraction from wrist-worn PPG devices is becoming increasingly important due to an expanded use in clinical settings [2]. Classical methods using statistical and signal processing techniques have much smaller computational footprints compared to DNN techniques. However, in periods of intense activities the DNN method outperformed the classical methods by a MAE of 8.91 BPM [1].

5. Conclusion

Preliminary results show that network architecture search is a viable method for creating architectures that would be un-achievable to create due to the enormous search space. With more search iterations and analysis of differing strategies, improvements on MAE are likely to be achieved.

References

[1] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks, Sensor. 19 (2019) 1–27. doi:10.3390/s19143079.

[2] B. Bent, B.A. Goldstein, W.A. Kibbe, and J.P. Dunn, Investigating sources of inaccuracy in wearable optical heart rate sensors, Npj Digit. Med. 3 (2020) 18. doi:10.1038/s41746-020-0226-6.

[3] Z. Zhang, IEEE Signal Processing Cup 2015: Heart Rate Monitoring During Physical Exercise Using Wrist-Type Photoplethysmographic (PPG) Signals, (2015). https://sites.google.com/site/researchbyzhang/ieeespcup2015 (accessed June 29, 2020).

[4] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, Introducing WeSAD, a multimodal dataset for wearable stress and affect detection, in: ICMI 2018 - Proc. 2018 Int. Conf. Multimodal Interact., Association for Computing Machinery, Inc, New York, New York, USA, 2018: pp. 400–408. doi:10.1145/3242969.3242985.

[5] B. Zoph, and Q. V. Le, Neural architecture search with reinforcement learning, in: 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc., 2017.

[6] K.O. Stanley, J. Clune, J. Lehman, and R. Miikkulainen, Designing neural networks through neuroevolution, Nat. Mach. Intell. 1 (2019) 24–35. doi:10.1038/s42256-018-0006-z.

[7] H. Jin, Q. Song, and X. Hu, Auto-Keras: An Efficient Neural Architecture Search System, in: Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, New York, NY, USA, 2019: pp. 1946–1956. doi:10.1145/3292500.3330648.

Bibliography

[1] K. Carter and M. Rutherford, "Cardiovascular system – blood vessels and blood," in Building a Medical Terminology Foundation, eCampus, 2020.

[2] J. D. Pollock and A. N. Makaryus, Physiology, Cardiac Cycle. StatPearls Publishing, 2022.

[3] M. D. Cesare et al., "World heart report 2023 confronting the world's Number One killer," 2023.

[4] V. Raleigh, D. Jefferies, and D. Wellings, "Cardiovascular disease in England: Supporting leaders to take actions," The Kings Fund, 2022.

[5] A. Niakouei, M. Tehrani, and L. Fulton, "Health Disparities and Cardiovascular Disease," Healthcare (Basel), vol. 8, no. 1, Mar. 2020, doi: 10.3390/healthcare8010065.

[6] American Heart Association, "Bridging the Gap: CVD and Health Equity," 2015.

[7] D. Ray, T. Collins, S. Woolley, and P. Ponnapalli, "A Review of Wearable Multi-Wavelength Photoplethysmography," IEEE Rev. Biomed. Eng., vol. 16, pp. 136–151, Jan. 2023, doi: 10.1109/RBME.2021.3121476.

[8] J. Lin et al., "Wearable sensors and devices for real-time cardiovascular disease monitoring," Cell Reports Physical Science, vol. 2, no. 8, p. 100541, Aug. 2021, doi: 10.1016/j.xcrp.2021.100541.

[9] A. Zinzuwadia and J. P. Singh, "Wearable devices-addressing bias and inequity," Lancet Digit Health, vol. 4, no. 12, pp. e856–e857, Dec. 2022, doi: 10.1016/S2589-7500(22)00194-7.

[10] J. Achten and A. E. Jeukendrup, "Heart rate monitoring: applications and limitations," Sports Med., vol. 33, no. 7, pp. 517–538, 2003, doi: 10.2165/00007256-200333070-00004.

[11] P. Palatini, "Role of elevated heart rate in the development of cardiovascular disease in hypertension," Hypertension, vol. 58, no. 5, pp. 745–750, Nov. 2011, doi: 10.1161/HYPERTENSIONAHA.111.173104.

[12] C. Perret-Guillaume, L. Joly, and A. Benetos, "Heart rate as a risk factor for cardiovascular disease," Prog. Cardiovasc. Dis., vol. 52, no. 1, pp. 6–10, Jul-Aug 2009, doi: 10.1016/j.pcad.2009.05.003.

[13] F. Custodis, J.-C. Reil, U. Laufs, and M. Böhm, "Heart rate: a global target for cardiovascular disease and therapy along the cardiovascular disease continuum," J. Cardiol., vol. 62, no. 3, pp. 183–187, Sep. 2013, doi: 10.1016/j.jjcc.2013.02.018.

[14] Å. Hjalmarson, "Heart rate: an independent risk factor in cardiovascular disease," Eur. Heart J. Suppl., vol. 9, no. suppl_F, pp. F3–F7, Sep. 2007, doi: 10.1093/eurheartj/sum030.

[15] J. Tian et al., "Association of resting heart rate and its change with incident cardiovascular events in the middle-aged and older Chinese," Sci. Rep., vol. 9, no. 1, p. 6556, Apr. 2019, doi: 10.1038/s41598-019-43045-5.

[16] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," Int J Biosens Bioelectron, vol. 4, no. 4, pp. 195–202, Aug. 2018, doi: 10.15406/ijbsbe.2018.04.00125.

[17] K. Khunti, "Accurate interpretation of the 12-lead ECG electrode placement: A systematic review," Health Educ. J., vol. 73, no. 5, pp. 610–623, Sep. 2014, doi: 10.1177/0017896912472328.

[18] S. Woolley, "Wearables and Connected Health Futures," ITNOW, vol. 65, no. 1, pp. 24–25, Feb. 2023, doi: 10.1093/combul/bwad012.

[19] F. Scardulla et al., "Photoplethysmograhic sensors, potential and limitations: Is it time for regulation? A comprehensive review," Measurement, vol. 218, p. 113150, Aug. 2023, doi: 10.1016/j.measurement.2023.113150.

[20] "ClinicalTrials.gov Database," 2023. Available: `https://clinicaltrials.gov/search?intr=Fitbit&viewType=Table`. [Accessed: Oct. 25, 2023]

[21] A.-N. Heizmann, C. Chapelle, S. Laporte, F. Roche, D. Hupin, and C. Le Hello, "Impact of wearable device-based interventions with feedback for increasing daily walking activity and physical capacities in cardiovascular patients: a systematic review and meta-analysis of randomised controlled trials," BMJ Open, vol. 13, no. 7, p. e069966, Jul. 2023, doi: 10.1136/bmjopen-2022-069966.

[22] "How Wearables Can Help Improve Cardiovascular Health." Available: `https://www.radcliffecardiology.com/news/how-wearables-can-help-improve-cardiovascular-health`. [Accessed: Oct. 25, 2023]

[23] J. Fine et al., "Sources of Inaccuracy in Photoplethysmography for Continuous Cardiovascular Monitoring," Biosensors, vol. 11, no. 4, p. 126, Apr. 2021, doi: 10.3390/bios11040126.

[24] R. Al-Halawani, P. H. Charlton, M. Qassem, and P. A. Kyriacou, "A review of the effect of skin pigmentation on pulse oximeter accuracy," Physiol. Meas., vol. 44, no. 5, Jun. 2023, doi: 10.1088/1361-6579/acd51a.

[25] F. Y. Sinaki et al., "Ethnic disparities in publicly-available pulse oximetry databases," Commun. Med., vol. 2, p. 59, May 2022, doi: 10.1038/s43856-022-00121-8.

[26] D. Ray, T. Collins, and P. V. S. Ponnapalli, "DeepPulse: An Uncertainty-aware Deep Neural Network for Heart Rate Estimations from Wrist-worn Photoplethysmography," Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 2022, pp. 1651–1654, Jul. 2022, doi: 10.1109/EMBC48229.2022.9871813.

[27] V. Bieri, P. Streli, B. U. Demirel, and C. Holz, "BeliefPPG: Uncertainty-aware Heart Rate Estimation from PPG signals via Belief Propagation," arXiv [cs.LG], Jun. 13, 2023.

[28] P. A. Kyriacou and J. Allen, Photoplethysmography: Technology, Signal Analysis and Applications. Elsevier Science, 2021.

[29] T. Pereira et al., "Photoplethysmography based atrial fibrillation detection: a review," Nature Research, 12 2020, pp. 1–12. doi: 10.1038/s41746-019-0207-9.

[30] E. Sazonov and M. R. Neuman, Wearable sensors: Fundamentals, implementation and applications. 2014, pp. 1–634. doi: 10.1016/C2013-0-06896-X.

[31] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," Physiol. Meas., vol. 28, no. 3, pp. R1–39, Mar. 2007, doi: 10.1088/0967-3334/28/3/R01.

[32] A. B. Hertzman, "Photoelectric plethysmography of the fingers and toes in man," Exp. Biol. Med. , vol. 37, no. 3, pp. 529–534, Dec. 1937, doi: 10.3181/00379727-37-9630.

[33] W. Montagna, A. M. Kligman, and K. S. Carlisle, Atlas of Normal Human Skin. Springer-Verlag, 1992. doi: 10.1007/978-1-4613-9202-6.

[34] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000 nm," J. Phys. D Appl. Phys., vol. 38, no. 15, pp. 2543–2555, Aug. 2005, doi: 10.1088/0022-3727/38/15/004.

[35] M. Brenner and V. J. Hearing, "The protective role of melanin against UV damage in human skin," Photochem. Photobiol., vol. 84, no. 3, pp. 539–549, May-Jun 2008, doi: 10.1111/j.1751-1097.2007.00226.x.

[36] G. Zonios, J. Bykowski, and N. Kollias, "Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy," J. Invest. Dermatol., vol. 117, no. 6, pp. 1452–1457, Dec. 2001, doi: 10.1046/j.0022-202x.2001.01577.x.

[37] B. Müller et al., "High-resolution tomographic imaging of microvessels," Developments in X-Ray Tomography VI, vol. 7078, p. 70780B, 9 2008, doi: 10.1117/12.794157.

[38] T. Lister, P. A. Wright, and P. H. Chappell, "Optical properties of human skin," J. Biomed. Opt., vol. 17, no. 9, pp. 90901–90901, Sep. 2012, doi: 10.1117/1.JBO.17.9.090901.

[39] R. R. Anderson and J. A. Parrish, "The optics of human skin," J. Invest. Dermatol., vol. 77, no. 1, pp. 13–19, Jul. 1981, doi: 10.1111/1523-1747.ep12479191.

[40] C. R. Simpson, M. Kohl, M. Essenpreis, and M. Cope, "Near-infrared optical properties of ex vivo human skin and subcutaneous tissues measured using the Monte Carlo inversion technique," Phys. Med. Biol., vol. 43, no. 9, pp. 2465–2478, Sep. 1998, doi: 10.1088/0031-9155/43/9/003.

[41] P. Taroni, A. Pifferi, A. Torricelli, D. Comelli, and R. Cubeddu, "In vivo absorption and scattering spectroscopy of biological tissues," Photochem. Photobiol. Sci., vol. 2, no. 2, pp. 124–129, Feb. 2003, doi: 10.1039/b209651j.

[42] J. Yao and S. Warren, "A novel algorithm to separate motion artifacts from photoplethysmographic signals obtained with a reflectance pulse oximeter," in The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Sep. 2004, pp. 2153–2156. doi: 10.1109/IEMBS.2004.1403630.

[43] L. G. Lindberg and P. A. Oberg, "Photoplethysmography. Part 2. Influence of light source wavelength," Med. Biol. Eng. Comput., vol. 29, no. 1, pp. 48–54, Jan. 1991, doi: 10.1007/BF02446295.

[44] A. Alzahrani et al., "A multi-channel opto-electronic sensor to accurately monitor heart rate against motion artefact during exercise," Sensors, vol. 15, no. 10, pp. 25681–25702, Oct. 2015, doi: 10.3390/s151025681.

[45] J. Spigulis, L. Gailite, A. Lihachev, and R. Erts, "Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography," 2007, pp. 1754–1759. doi: 10.1364/AO.46.001754.

[46] S. Han, D. Roh, J. Park, and H. Shin, "Design of Multi-Wavelength Optical Sensor Module for Depth-Dependent Photoplethysmography," Sensors, vol. 19, no. 24, p. 5441, Dec. 2019, doi: 10.3390/s19245441.

[47] V. Vizbara, "Comparison of Green, Blue and Infrared Light in Wrist and Forehead Photoplethysmography," Biomed. Eng. , vol. 17, no. 1, pp. 78–81, Nov. 2013,

[48] D. Barolet, "Light-emitting diodes (LEDs) in dermatology," Semin. Cutan. Med. Surg., vol. 27, no. 4, pp. 227–238, Dec. 2008, doi: 10.1016/j.sder.2008.08.003.

[49] A. Shcherbina et al., "Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort," J Pers Med, vol. 7, no. 2, pp. 1–12, May 2017, doi: 10.3390/jpm7020003.

[50] P. Mohapatra, S. P. Preejith, and M. Sivaprakasam, "A novel sensor for wrist based optical heart rate monitor," in 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), May 2017, pp. 1–6. doi: 10.1109/I2MTC.2017.7969842.

[51] S. P. Preejith, A. Alex, J. Joseph, and M. Sivaprakasam, "Design, development and clinical validation of a wrist-based optical heart rate monitor," in 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Institute of Electrical and Electronics Engineers Inc., May 2016, pp. 1–6. doi: 10.1109/MeMeA.2016.7533786.

[52] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," J. Clin. Monit. Comput., vol. 27, no. 3, pp. 313–317, Jun. 2013, doi: 10.1007/s10877-013-9436-7.

[53] A. L. Ries, L. M. Prewitt, and J. J. Johnson, "Skin color and ear oximetry," Chest, vol. 96, no. 2, pp. 287–290, Aug. 1989, doi: 10.1378/chest.96.2.287.

[54] P. E. Bickler, J. R. Feiner, and J. W. Severinghaus, "Effects of skin pigmentation on pulse oximeter accuracy at low saturation," Anesthesiology, vol. 102, no. 4, pp. 715–719, Apr. 2005, doi: 10.1097/00000542-200504000-00004.

[55] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial Bias in Pulse Oximetry Measurement," N. Engl. J. Med., vol. 383, no. 25, pp. 2477–2478, Dec. 2020, doi: 10.1056/NEJMc2029240.

[56] J. R. Feiner, J. W. Severinghaus, and P. E. Bickler, "Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender," Anesth. Analg., vol. 105, no. 6 Suppl, pp. S18–S23, Dec. 2007, doi: 10.1213/01.ane.0000285988.35174.d9.

[57] P. A. Bothma et al., "Accuracy of pulse oximetry in pigmented patients," S. Afr. Med. J., vol. 86, no. 5 Suppl, pp. 594–596, May 1996,

[58] E. E. Foglia et al., "The Effect of Skin Pigmentation on the Accuracy of Pulse Oximetry in Infants with Hypoxemia," J. Pediatr., vol. 182, pp. 375–377.e2, Mar. 2017, doi: 10.1016/j.jpeds.2016.11.043.

[59] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," NPJ Digit Med, vol. 3, p. 18, Feb. 2020, doi: 10.1038/s41746-020-0226-6.

[60] L. Yan, S. Hu, A. Alzahrani, S. Alharbi, and P. Blanos, "A Multi-Wavelength Opto-Electronic Patch Sensor to Effectively Detect Physiological Changes against Human Skin Types," Biosensors, vol. 7, no. 2, p. 22, Jun. 2017, doi: 10.3390/bios7020022.

[61] I. C. Jeong, H. Yoon, H. Kang, and H. Yeom, "Effects of skin surface temperature on photoplethysmograph," J. Healthc. Eng., vol. 5, no. 4, pp. 429–438, 2014, doi: 10.1260/2040-2295.5.4.429.

[62] K. Budidha and P. A. Kyriacou, "In vivo investigation of ear canal pulse oximetry during hypothermia," J. Clin. Monit. Comput., vol. 32, no. 1, pp. 97–107, Feb. 2018, doi: 10.1007/s10877-017-9975-4.

[63] A. C. Ralston, R. K. Webb, and W. B. Runciman, "Potential errors in pulse oximetry. I. Pulse oximeter evaluation," Anaesthesia, vol. 46, no. 3, pp. 202–206, Mar. 1991, doi: 10.1111/j.1365-2044.1991.tb09410.x.

[64] B. Askarian, K. Jung, and J. W. Chong, "Monitoring of Heart Rate from Photoplethysmographic Signals Using a Samsung Galaxy Note8 in Underwater Environments," Sensors, vol. 19, no. 13, Jun. 2019, doi: 10.3390/s19132846.

[65] Y. Maeda, M. Sekine, T. Tamura, A. Moriya, T. Suzuki, and K. Kameyama, "Comparison of reflected green light and infrared photoplethysmography," in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Aug. 2008, pp. 2270–2272. doi: 10.1109/IEMBS.2008.4649649.

[66] Y. Maeda, M. Sekine, and T. Tamura, "The advantages of wearable green reflected photoplethysmography," J. Med. Syst., vol. 35, no. 5, pp. 829–834, Oct. 2011, doi: 10.1007/s10916-010-9506-z.

[67] P. H. Charlton, P. A. Kyriaco, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey, "Wearable photoplethysmography for cardiovascular monitoring," Proc. IEEE Inst. Electr. Electron. Eng., vol. 110, no. 3, pp. 355–381, Mar. 2022, doi: 10.1109/JPROC.2022.3149785.

[68] L. Asare, E. Kviesis-Kipge, A. Grabovskis, U. Rubins, J. Spigulis, and R. Erts, "Multi-spectral photoplethysmography biosensor," in Optical Sensors 2011; and Photonic Crystal Fibers V, SPIE, May 2011, pp. 374–379. doi: 10.1117/12.887176.

[69] S. Alharbi, S. Hu, D. Mulvaney, and P. Blanos, "An applicable approach for extracting human heart rate and oxygen saturation during physical movements using a multi-wavelength illumination optoelectronic sensor system," in Design and Quality for Biomedical Technologies XI, SPIE, Feb. 2018, pp. 85–99. doi: 10.1117/12.2287854.

[70] S. S. Gupta, S. Hossain, C. A. Haque, and K.-D. Kim, "In-Vivo Estimation of Glucose Level Using PPG Signal," in 2020 International Conference on Information and Communication Technology Convergence (ICTC), IEEE Computer Society, Oct. 2020, pp. 733–736. doi: 10.1109/ICTC49870.2020.9289629.

[71] K. M. Warren, J. R. Harvey, K. H. Chon, and Y. Mendelson, "Improving Pulse Rate Measurements during Random Motion Using a Wearable Multichannel Reflectance Photoplethysmograph," Sensors, vol. 16, no. 3, Mar. 2016, doi: 10.3390/s16030342.

[72] J. Lee, M. Kim, H.-K. Park, and I. Y. Kim, "Motion Artifact Reduction in Wearable Photoplethysmography Based on Multi-Channel Sensors with Multiple Wavelengths," Sensors, vol. 20, no. 5, Mar. 2020, doi: 10.3390/s20051493.

[73] C.-C. Chang, C.-T. Wu, B. I. Choi, and T.-J. Fang, "MW-PPG Sensor: An on-Chip Spectrometer Approach," Sensors, vol. 19, no. 17, p. 3698, Aug. 2019, doi: 10.3390/s19173698.

[74] S.-H. Chen, Y.-C. Chuang, and C.-C. Chang, "Development of a Portable All-Wavelength PPG Sensing Device for Robust Adaptive-Depth Measurement: A Spectrometer Approach with a Hydrostatic Measurement Example," Sensors, vol. 20, no. 22, p. 6556, Nov. 2020, doi: 10.3390/s20226556.

[75] Y. Maeda, M. Sekine, and T. Tamura, "Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography," J Med Syst, 10 2011, pp. 969–976. doi: 10.1007/s10916-010-9505-0.

[76] Y. Mendelson and C. Pujary, "Measurement site and photodetector size considerations in optimizing power consumption of a wearable reflectance pulse oximeter,"

in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), IEEE, Sep. 2003, pp. 3016–3019 Vol.4. doi: 10.1109/IEMBS.2003.1280775.

[77] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable Photoplethysmographic Sensors—Past and Present," Electronics, vol. 3, no. 2, pp. 282–302, Apr. 2014, doi: 10.3390/electronics3020282.

[78] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng, "Quantitative Comparison of Photoplethysmographic Waveform Characteristics: Effect of Measurement Site," Front. Physiol., vol. 10, p. 198, Mar. 2019, doi: 10.3389/fphys.2019.00198.

[79] J. Liu, B. P. Yan, Y.-T. Zhang, X.-R. Ding, P. Su, and N. Zhao, "Multi-Wavelength Photoplethysmography Enabling Continuous Blood Pressure Measurement With Compact Wearable Electronics," IEEE Trans. Biomed. Eng., vol. 66, no. 6, pp. 1514–1525, Jun. 2019, doi: 10.1109/TBME.2018.2874957.

[80] S. K. Longmore, G. Y. Lui, G. Naik, P. P. Breen, B. Jalaludin, and G. D. Gargiulo, "A Comparison of Reflective Photoplethysmography for Detection of Heart Rate, Blood Oxygen Saturation, and Respiration Rate at Various Anatomical Locations," Sensors, vol. 19, no. 8, p. 1874, Apr. 2019, doi: 10.3390/s19081874.

[81] S. Lee, H. Shin, and C. Hahm, "Effective PPG sensor placement for reflected red and green light, and infrared wristband-type photoplethysmography," in 2016 18th International Conference on Advanced Communication Technology (ICACT), Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 556–558. doi: 10.1109/ICACT.2016.7423470.

[82] E. G. Kim, H. Heo, K. C. Nam, and Y. Huh, "Measurement Site and Applied Pressure Consideration in Wrist Photoplethysmography," IEICE Proceedings Series, vol. 39, no. P1–32, Jul. 2008,

[83] P. Research and Technology, "Polar Precision Prime OHR," Polar Research and Technology, 2019, Available: `https://www.polar.com/sites/default/files/static/science/white-papers/polar-precision-prime-white-paper.pdf`

[84] T. Rukasha, S. I. Woolley, and T. Collins, "Poster: Heart Rate Performance of a Medical-Grade Data Streaming Wearable Device," in 2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), IEEE Press, Jan. 2021, pp. 12–13. doi: 10.1145/3384420.3431776.

[85] Y.-H. Kao, P. C.-P. Chao, and C.-L. Wey, "Design and validation of a new PPG module to acquire high-quality physiological signals for high-accuracy biomedical

sensing," IEEE J. Sel. Top. Quantum Electron., vol. 25, no. 1, pp. 1–10, Jan. 2019, doi: 10.1109/jstqe.2018.2871604.

[86] H. J. Baek, S. Y. Sim, J. S. Kim, and K. S. Park, "Effect of sensor configurations on indirect-contact photoplethysmogram measurement system," in 2010 5th Cairo International Biomedical Engineering Conference, IEEE, Dec. 2010, pp. 244–246. doi: 10.1109/cibec.2010.5716070.

[87] B. Bent and J. P. Dunn, "Optimizing sampling rate of wrist-worn optical sensors for physiologic monitoring," J Clin Transl Sci, vol. 5, no. 1, p. e34, Aug. 2020, doi: 10.1017/cts.2020.526.

[88] X. F. Teng and Y. T. Zhang, "The effect of contacting force on photoplethysmographic signals," Physiol. Meas., vol. 25, no. 5, pp. 1323–1335, Oct. 2004, doi: 10.1088/0967-3334/25/5/020.

[89] Y. Zhang et al., "Motion Artifact Reduction for Wrist-Worn Photoplethysmograph Sensors Based on Different Wavelengths," Sensors, vol. 19, no. 3, Feb. 2019, doi: 10.3390/s19030673.

[90] K. Matsumura, P. Rolfe, J. Lee, and T. Yamakoshi, "iPhone 4s photoplethysmography: which light color yields the most accurate heart rate and normalized pulse volume using the iPhysioMeter Application in the presence of motion artifact?," PLoS One, vol. 9, no. 3, p. e91205, Mar. 2014, doi: 10.1371/journal.pone.0091205.

[91] C.-Y. Wang and K.-T. Tang, "Active noise cancellation of motion artifacts in pulse oximetry using isobestic wavelength light source," in 2011 IEEE International Symposium of Circuits and Systems (ISCAS), May 2011, pp. 1029–1032. doi: 10.1109/ISCAS.2011.5937744.

[92] S. Ismail, U. Akram, and I. Siddiqi, "Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review," EURASIP J. Adv. Signal Process., vol. 2021, no. 1, pp. 1–27, Jan. 2021, doi: 10.1186/s13634-020-00714-2.

[93] V. Periyasamy, M. Pramanik, and P. K. Ghosh, "Review on Heart-Rate Estimation from Photoplethysmography and Accelerometer Signals During Physical Exercise," J. Indian Inst. Sci., vol. 97, no. 3, pp. 313–324, Sep. 2017, doi: 10.1007/s41745-017-0037-1.

[94] Pankaj, A. Kumar, R. Komaragiri, and M. Kumar, "A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation," Arch. Comput. Methods Eng., vol. 1, p. 3, 6 2021, doi: 10.1007/s11831-021-09597-4.

[95] D. Biswas, N. Simões-Capela, C. Van Hoof, and N. Van Helleputte, "Heart Rate Estimation From Wrist-Worn Photoplethysmography: A Review," IEEE Sens. J., vol. 19, no. 16, pp. 6560–6570, Aug. 2019, doi: 10.1109/JSEN.2019.2914166.

[96] T. Tamura, "Current progress of photoplethysmography and SPO2 for health monitoring," Association for Computing Machinery, Inc, 10 2019, pp. 400–408. doi: 10.1007/s13534-019-00097-w.

[97] P. H. Charlton et al., "Breathing Rate Estimation From the Electrocardiogram and Photoplethysmogram: A Review," IEEE Rev. Biomed. Eng., vol. 11, pp. 2–20, 2018, doi: 10.1109/RBME.2017.2763681.

[98] G. Wang, M. Atef, and Y. Lian, "Towards a Continuous Non-Invasive Cuffless Blood Pressure Monitoring System Using PPG: Systems and Circuits Review," IEEE Circuits and Systems Magazine, vol. 18, no. 3, pp. 6–26, thirdquarter 2018, doi: 10.1109/MCAS.2018.2849261.

[99] E. Lam, S. Aratia, J. Wang, and J. Tung, "Measuring Heart Rate Variability in Free-Living Conditions Using Consumer-Grade Photoplethysmography: Validation Study," JMIR Biomedical Engineering, vol. 5, no. 1, p. e17355, Nov. 2020, doi: 10.2196/17355.

[100] M. Elgendi et al., "The use of photoplethysmography for assessing hypertension," NPJ Digit Med, vol. 2, p. 60, Jun. 2019, doi: 10.1038/s41746-019-0136-7.

[101] Y. K. Qawqzeh, "The Analysis of PPG Time Indices to Predict Aging and Atherosclerosis," in Intelligent Computing Paradigm and Cutting-edge Technologies, Springer International Publishing, 2020, pp. 218–225. doi: 10.1007/978-3-030-38501-9_22.

[102] N. Mangathayaru, B. P. Rani, V. Janaki, L. S. Kotturi, M. Vallabhapurapu, and G. Vikas, "Heart Rate Variability for Predicting Coronary Heart Disease using Photoplethysmography," in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 664–671. doi: 10.1109/I-SMAC49090.2020.9243316.

[103] T. Rukasha, S. I Woolley, T. Kyriacou, and T. Collins, "Evaluation of Wearable Electronics for Epilepsy: A Systematic Review," Electronics, vol. 9, no. 6, p. 968, Jun. 2020, doi: 10.3390/electronics9060968.

[104] M. K. Uçar, S. Örenç, M. R. Bozkurt, and C. Bilgin, "Evaluation of the relationship between Chronic Obstructive Pulmonary Disease and photoplethysmography signal," in 2017 Medical Technologies National Congress (TIPTEKNO),

Institute of Electrical and Electronics Engineers Inc., Oct. 2017, pp. 1–4. doi: 10.1109/TIPTEKNO.2017.8238032.

[105] P. H. Charlton, P. Celka, B. Farukh, P. Chowienczyk, and J. Alastruey, "Assessing mental stress from the photoplethysmogram: a numerical study," Physiol. Meas., vol. 39, no. 5, p. 054001, May 2018, doi: 10.1088/1361-6579/aabe6a.

[106] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, "Wearable-Based Affect Recognition-A Review," Sensors, vol. 19, no. 19, pp. 1–42, Sep. 2019, doi: 10.3390/s19194079.

[107] H. Scott, L. Lack, and N. Lovato, "A systematic review of the accuracy of sleep wearable devices for estimating sleep onset," Sleep Med. Rev., vol. 49, p. 101227, Feb. 2020, doi: 10.1016/j.smrv.2019.101227.

[108] M. Shokoueinejad et al., "Sleep apnea: a review of diagnostic sensors, algorithms, and therapies," Physiol. Meas., vol. 38, no. 9, pp. R204–R252, Aug. 2017, doi: 10.1088/1361-6579/aa6ec6.

[109] E. Susana and K. Ramli, "Review of Non-Invasive Blood Glucose Level Estimation based on Photoplethysmography and Artificial Intelligent Technology," pp. 158–163, Oct. 2021, doi: 10.1109/QIR54354.2021.9716164.

[110] P. Adhikari, I. B. Magaña, and P. D. O'Neal, "Multi-wavelength pulse plethysmography for real-time drug delivery monitoring," in Optical Diagnostics and Sensing XIV: Toward Point-of-Care Diagnostics, SPIE, Feb. 2014, pp. 148–152. doi: 10.1117/12.2040064.

[111] P. Adhikari, W. Eklund, and P. D. O'Neal, "Non-invasive in vivo monitoring of circulating amphotericin b using multi-wavelength photoplethysmography," in Optical Diagnostics and Sensing XV: Toward Point-of-Care Diagnostics, SPIE, Mar. 2015, pp. 62–67. doi: 10.1117/12.2083798.

[112] P. Adhikari, W. Eklund, E. A. Sherer, and D. Patrick O'Neal, "Assessment of multi-wavelength pulse photometry for non-invasive dose estimation of circulating drugs and nanoparticles," in Optical Diagnostics and Sensing XVI: Toward Point-of-Care Diagnostics, SPIE, Mar. 2016, pp. 110–116. doi: 10.1117/12.2213455.

[113] P. H. Charlton et al., "The 2023 wearable photoplethysmography roadmap," Physiol. Meas., Jul. 2023, doi: 10.1088/1361-6579/acead2.

[114] S. Alharbi et al., "Oxygen Saturation Measurements from Green and Orange Illuminations of Multi-Wavelength Optoelectronic Patch Sensors," Sensors, vol. 19, no. 1, Dec. 2018, doi: 10.3390/s19010118.

[115] F. Scardulla et al., "A novel multi-wavelength procedure for blood pressure estimation using opto-physiological sensor at peripheral arteries and capillaries," in Design and Quality for Biomedical Technologies XI, SPIE, Feb. 2018, pp. 135–145. doi: 10.1117/12.2287845.

[116] J. Liu, B. P.-Y. Yan, W.-X. Dai, X.-R. Ding, Y.-T. Zhang, and N. Zhao, "Multi-wavelength photoplethysmography method for skin arterial pulse extraction," Biomed. Opt. Express, vol. 7, no. 10, pp. 4313–4326, Oct. 2016, doi: 10.1364/BOE.7.004313.

[117] Z. Geng, F. Tang, Y. Ding, S. Li, and X. Wang, "Noninvasive Continuous Glucose Monitoring Using a Multisensor-Based Glucometer and Time Series Analysis," Sci. Rep., vol. 7, no. 1, p. 12650, Oct. 2017, doi: 10.1038/s41598-017-13018-7.

[118] V. P. Rachim and W.-Y. Chung, "Wearable-band type visible-near infrared optical biosensor for non-invasive blood glucose monitoring," Sens. Actuators B Chem., vol. 286, pp. 173–180, May 2019, doi: 10.1016/j.snb.2019.01.121.

[119] "UK Biobank." Available: `https://www.ukbiobank.ac.uk/`. [Accessed: Sep. 08, 2023]

[120] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, in ICMI '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 400–408. doi: 10.1145/3242969.3242985.

[121] V. Markova, "Database for Cognitive Load Affect and Stress recognition." IEEE Dataport, Jan. 31, 2020. doi: 10.21227/ybsw-yr53. Available: `https://ieee-dataport.org/open-access/database-cognitive-load-affect-and-stress-recognition`. [Accessed: Sep. 08, 2023]

[122] M. A. F. Pimentel et al., "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters," IEEE Trans. Biomed. Eng., vol. 64, no. 8, pp. 1914–1923, Aug. 2017, doi: 10.1109/TBME.2016.2613124.

[123] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," IEEE Trans. Biomed. Eng., vol. 60, no. 7, pp. 1946–1953, Jul. 2013, doi: 10.1109/TBME.2013.2246160.

[124] P. H. Charlton et al., "Detecting beats in the photoplethysmogram: benchmarking open-source algorithms," Physiol. Meas., vol. 43, no. 8, Aug. 2022, doi: 10.1088/1361-6579/ac826d.

[125] B. W. Nelson, C. A. Low, N. Jacobson, P. Areán, J. Torous, and N. B. Allen, "Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research," NPJ Digit Med, vol. 3, p. 90, Jun. 2020, doi: 10.1038/s41746-020-0297-4.

[126] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," IEEE Trans. Biomed. Eng., vol. 62, no. 2, pp. 522–531, Feb. 2015, doi: 10.1109/TBME.2014.2359372.

[127] H. Chung, H. Ko, H. Lee, and J. Lee, "Deep Learning for Heart Rate Estimation From Reflectance Photoplethysmography With Acceleration Power Spectrum and Acceleration Intensity," IEEE Access, vol. 8, pp. 63390–63402, 2020, doi: 10.1109/ACCESS.2020.2981956.

[128] L. Alkhoury, J. Choi, V. D. Chandran, G. B. De Carvalho, S. Pal, and M. Kam, "Dual Wavelength Photoplethysmography Framework for Heart Rate Calculation," Sensors, vol. 22, no. 24, Dec. 2022, doi: 10.3390/s22249955.

[129] D. Jarchi and A. J. Casson, "Description of a Database Containing Wrist PPG Signals Recorded during Physical Exercise with Both Accelerometer and Gyroscope Measures of Motion," Brown Univ. Dig. Addict. Theory Appl., vol. 2, no. 1, p. 1, Dec. 2016, doi: 10.3390/data2010001.

[130] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks," Sensors, vol. 19, no. 14, pp. 1–27, Jul. 2019, doi: 10.3390/s19143079.

[131] P. H. Charlton, K. Pilt, and P. A. Kyriacou, "Establishing best practices in photoplethysmography signal acquisition and processing," Physiol. Meas., vol. 43, no. 5, p. 050301, May 2022, doi: 10.1088/1361-6579/ac6cc4.

[132] Z. Zhang, Z. Pi, and B. Liu, "IEEE PPG Signal Processing Cup 2015," TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. Jun. 21, 2020. doi: 10.5281/zenodo.3902710. Available: `https://zenodo.org/record/3902710`

[133] "IEEE Signal Processing Cup 2015," IEEE Signal Processing Society, Dec. 30, 2020. Available: `https://signalprocessingsociety.org/community-involvement/ieee-signal-processing-cup-2015`. [Accessed: Sep. 06, 2023]

[134] A. Casson and J. Delaram, "Wrist PPG During Exercise," Description of a Database Containing Wrist PPG Signals Recorded during Physical Exercise with Both Accelerometer and Gyroscope Measures of Motion. Oct. 20, 2017. doi: 10.13026/C2PQ1X.

Available: https://physionet.org/content/wrist/1.0.0/. [Accessed: Jun. 08, 2023]

[135] H. Lee, H. Chung, and J. Lee, "BAMI 1 and 2," Deep Learning for Heart Rate Estimation from Reflectance Photoplethysmography with Acceleration Power Spectrum and Acceleration Intensity. May 11, 2021. Available: https://github.com/HeewonChung92/CNN_LSTM_HeartRateEstimation. [Accessed: Jun. 08, 2023]

[136] L. Alkhoury, DWL Dataset. Github. Available: https://github.com/ludvikalkhoury/DWL-Method. [Accessed: Sep. 08, 2023]

[137] C. Orphanidou, "Quality Assessment for the Photoplethysmogram (PPG)," in Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations, C. Orphanidou, Ed., Cham: Springer International Publishing, 2018, pp. 41–63. doi: 10.1007/978-3-319-68415-4_3.

[138] M. Elgendi, "Optimal Signal Quality Index for Photoplethysmogram Signals," Bioengineering (Basel), vol. 3, no. 4, Sep. 2016, doi: 10.3390/bioengineering3040021.

[139] M. Rinkevičius, P. H. Charlton, R. Bailón, and V. Marozas, "Influence of Photoplethysmogram Signal Quality on Pulse Arrival Time during Polysomnography," Sensors, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042220.

[140] S. Moscato, S. L. Giudice, G. Massaro, and L. Chiari, "Wrist Photoplethysmography Signal Quality Assessment for Reliable Heart Rate Estimate and Morphological Analysis," Sensors, vol. 22, no. 15, Aug. 2022, doi: 10.3390/s22155831.

[141] N. Pradhan, S. Rajan, and A. Adler, "Evaluation of the signal quality of wrist-based photoplethysmography," Physiological Measurement, vol. 40, 2019, doi: 10.1088/1361-6579/ab225a.

[142] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring," IEEE J Biomed Health Inform, vol. 19, no. 3, pp. 832–838, May 2015, doi: 10.1109/JBHI.2014.2338351.

[143] T. Pereira et al., "A Supervised Approach to Robust Photoplethysmography Quality Assessment," IEEE J Biomed Health Inform, vol. 24, no. 3, pp. 649–657, Mar. 2020, doi: 10.1109/JBHI.2019.2909065.

[144] W.-K. Beh, Y.-C. Yang, Y.-C. Lo, Y.-C. Lee, and A.-Y. Wu, "Machine-aided PPG Signal Quality Assessment (SQA) for Multi-mode Physiological Signal Monitoring," ACM Trans. Comput. Healthcare, vol. 4, no. 2, pp. 1–20, Apr. 2023, doi: 10.1145/3587256.

[145] ANSI, "Physical Activity Monitoring for Heart Rate - Real World Analysis," ANSI/CTA-2065.1, 2018. Available: `https://shop.cta.tech/products/physical-activity-monitoring-for-heart-rate`. [Accessed: Sep. 07, 2023]

[146] ANSI/AAMI, "Cardiac Monitors, Heart Rate Meters, And Alarms," EC13-2002. Available: `https://webstore.ansi.org/standards/aami/ansiaamiec132002`. [Accessed: Sep. 07, 2023]

[147] S. M. Bishop and A. Ercole, "Multi-Scale Peak and Trough Detection Optimised for Periodic and Quasi-Periodic Neuroscience Data," in Intracranial Pressure & Neuromonitoring XVI, Springer International Publishing, 2018, pp. 189–195. doi: 10.1007/978-3-319-65798-1_39.

[148] A. N. Vest et al., "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," Physiol. Meas., vol. 39, no. 10, p. 105004, Oct. 2018, doi: 10.1088/1361-6579/aae021.

[149] M. Aboy, J. McNames, T. Thong, D. Tsunami, M. S. Ellenby, and B. Goldstein, "An automatic beat detection algorithm for pressure signals," IEEE Trans. Biomed. Eng., vol. 52, no. 10, pp. 1662–1670, Oct. 2005, doi: 10.1109/TBME.2005.855725.

[150] F. Scholkmann, J. Boss, and M. Wolf, "An Efficient Algorithm for Automatic Peak Detection in Noisy Periodic and Quasi-Periodic Signals," Algorithms, vol. 5, no. 4, pp. 588–603, Nov. 2012, doi: 10.3390/a5040588.

[151] H. S. Shin, C. Lee, and M. Lee, "Adaptive threshold method for the peak detection of photoplethysmographic waveform," Comput. Biol. Med., vol. 39, no. 12, pp. 1145–1152, Dec. 2009, doi: 10.1016/j.compbiomed.2009.10.006.

[152] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," PLoS One, vol. 8, no. 10, p. e76585, Oct. 2013, doi: 10.1371/journal.pone.0076585.

[153] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "HeartPy: A novel heart rate algorithm for the analysis of noisy signals," Transp. Res. Part F Traffic Psychol. Behav., vol. 66, pp. 368–378, Oct. 2019, doi: 10.1016/j.trf.2019.09.015.

[154] W. Karlen, J. M. Ansermino, and G. Dumont, "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications," in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 2012, pp. 3131–3134. doi: 10.1109/EMBC.2012.6346628.

[155] E. J. Argüello Prada and R. D. Serna Maldonado, "A novel and low-complexity peak detection algorithm for heart rate estimation from low-amplitude photoplethysmographic (PPG) signals," J. Med. Eng. Technol., vol. 42, no. 8, pp. 569–577, Nov. 2018, doi: 10.1080/03091902.2019.1572237.

[156] B. N. Li, M. C. Dong, and M. I. Vai, "On an automatic delineator for arterial blood pressure waveforms," Biomed. Signal Process. Control, vol. 5, no. 1, pp. 76–81, Jan. 2010, doi: 10.1016/j.bspc.2009.06.002.

[157] J. Lázaro, E. Gil, J. M. Vergara, and P. Laguna, "Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children," IEEE J Biomed Health Inform, vol. 18, no. 1, pp. 240–246, Jan. 2014, doi: 10.1109/JBHI.2013.2267096.

[158] C. Pettit and P. Aston, "Photoplethysmogram Beat Detection Using Symmetric Projection Attractor Reconstruction," [in preparation], Available: `https://ppg-beats.readthedocs.io/en/latest/functions/spar_beat_detector/`

[159] S. Vadrevu and M. S. Manikandan, "A Robust Pulse Onset and Peak Detection Method for Automated PPG Signal Analysis System," IEEE Trans. Instrum. Meas., vol. 68, no. 3, pp. 807–817, Mar. 2019, doi: 10.1109/TIM.2018.2857878.

[160] N. J. Conn and D. A. Borkholder, "Wavelet based photoplethysmogram foot delineation for heart rate variability applications," in 2013 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Dec. 2013, pp. 1–5. doi: 10.1109/SPMB.2013.6736782.

[161] S. M. A. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon, "A Novel Time-Varying Spectral Filtering Algorithm for Reconstruction of Motion Artifact Corrupted Heart Rate Signals During Intense Physical Activities Using a Wearable Photoplethysmogram Sensor," Sensors, vol. 16, no. 1, p. 10, Dec. 2015, doi: 10.3390/s16010010.

[162] T. Schäck, M. Muma, and A. M. Zoubir, "Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals," in 2017 25th European Signal Processing Conference (EUSIPCO), Institute of Electrical and Electronics Engineers Inc., Aug. 2017, pp. 2478–2481. doi: 10.23919/EUSIPCO.2017.8081656.

[163] F. Chollet, Deep Learning with Python, Second Edition. New York, NY: Simon and Schuster, 2021.

[164] D. Biswas et al., "CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment," IEEE Trans.

Biomed. Circuits Syst., vol. 13, no. 2, pp. 282–291, Apr. 2019, doi: 10.1109/TBCAS.2019.2892297.

[165] A. Shyam, V. Ravichandran, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "PPGnet: Deep Network for Device Independent Heart Rate Estimation from Photoplethysmogram," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jul. 2019, pp. 1899–1902. doi: 10.1109/EMBC.2019.8856989.

[166] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "PP-Net: A Deep Learning Framework for PPG-Based Blood Pressure and Heart Rate Estimation," IEEE Sens. J., vol. 20, no. 17, pp. 10000–10011, Sep. 2020, doi: 10.1109/JSEN.2020.2990864.

[167] L. G. Rocha et al., "Binary CorNET: Accelerator for HR Estimation From Wrist-PPG," IEEE Trans. Biomed. Circuits Syst., vol. 14, no. 4, pp. 715–726, Aug. 2020, doi: 10.1109/TBCAS.2020.3001675.

[168] S. B. Song, J. W. Nam, and J. H. Kim, "NAS-PPG: PPG-Based Heart Rate Estimation Using Neural Architecture Search," IEEE Sens. J., vol. 21, no. 13, pp. 14941–14949, Jul. 2021, doi: 10.1109/JSEN.2021.3073047.

[169] S. Ismail, I. Siddiqi, and U. Akram, "Heart rate estimation in PPG signals using Convolutional-Recurrent Regressor," Comput. Biol. Med., vol. 145, p. 105470, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105470.

[170] A. Burrello et al., "Q-PPG: Energy-Efficient PPG-Based Heart Rate Monitoring on Wearable Devices," IEEE Trans. Biomed. Circuits Syst., vol. 15, no. 6, pp. 1196–1209, Dec. 2021, doi: 10.1109/TBCAS.2021.3122017.

[171] M. Risso et al., "Robust and Energy-Efficient PPG-Based Heart-Rate Monitoring," in 2021 IEEE International Symposium on Circuits and Systems (ISCAS), May 2021, pp. 1–5. doi: 10.1109/ISCAS51556.2021.9401282.

[172] A. Burrello et al., "Embedding Temporal Convolutional Networks for Energy-efficient PPG-based Heart Rate Monitoring," ACM Trans. Comput. Healthcare, vol. 3, no. 2, pp. 1–25, Mar. 2022, doi: 10.1145/3487910.

[173] X. Chang, G. Li, G. Xing, K. Zhu, and L. Tu, "DeepHeart: A Deep Learning Approach for Accurate Heart Rate Estimation from PPG Signals," ACM Trans. Sen. Netw., vol. 17, no. 2, pp. 1–18, Jan. 2021, doi: 10.1145/3441626.

[174] P. Sarkar and A. Etemad, "CardioGAN: Attentive Generative Adversarial Network with Dual Discriminators for Synthesis of ECG from PPG," arXiv [cs.LG], Sep. 30, 2020.

[175] P. Kasnesis, L. Toumanidis, A. Burrello, C. Chatzigeorgiou, and C. Z. Patrikakis, "Multi-Head Cross-Attentional PPG and Motion Signal Fusion for Heart Rate Estimation," arXiv [eess.SP], Oct. 14, 2022.

[176] D. Ray, T. Collins, and P. Ponnapalli, "Deep Neural Network Architecture Search for Wearable Heart Rate Estimations," Stud. Health Technol. Inform., vol. 281, pp. 1106–1107, May 2021, doi: 10.3233/SHTI210366.

[177] M. Wilkosz and A. Szczęsna, "Multi-Headed Conv-LSTM Network for Heart Rate Estimation during Daily Living Activities," Sensors, vol. 21, no. 15, Jul. 2021, doi: 10.3390/s21155212.

[178] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based Framework for Multivariate Time Series Representation Learning," arXiv [cs.LG], Oct. 06, 2020.

[179] K. Kazemi, J. Laitala, I. Azimi, P. Liljeberg, and A. M. Rahmani, "Robust PPG Peak Detection Using Dilated Convolutional Neural Networks," Sensors, vol. 22, no. 16, Aug. 2022, doi: 10.3390/s22166054.

[180] B. Ngoc-Thang, T. M. Tien Nguyen, T. T. Truong, B. L.-H. Nguyen, and T. T. Nguyen, "A dynamic reconfigurable wearable device to acquire high quality PPG signal and robust heart rate estimate based on deep learning algorithm for smart healthcare system," Biosensors and Bioelectronics: X, vol. 12, p. 100223, Dec. 2022, doi: 10.1016/j.biosx.2022.100223.

[181] P. Mehrgardt, M. Khushi, S. Poon, and A. Withana, "Deep Learning Fused Wearable Pressure and PPG Data for Accurate Heart Rate Monitoring," IEEE Sens. J., vol. 21, no. 23, pp. 27106–27115, Dec. 2021, doi: 10.1109/JSEN.2021.3123243.

[182] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," arXiv [cs.LG], Aug. 23, 2019.

[183] J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Machine Bias," May 23, 2016. Available: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`. [Accessed: Sep. 18, 2023]

[184] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[185] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger, "Uncertainty Toolbox: an Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification," arXiv [cs.LG], Sep. 2021,

[186] D. Ray, "Towards Wrist-worn Photoplethysmography Sensing for Medical Applications," in 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), Aug. 2021, pp. 432–433. doi: 10.1109/ICHI52183.2021.00070.

[187] "QardioCore User Manual," Qardio, 2020. Available: `https://support.qardio.com/hc/en-us/articles/115004855669-EN-User-Manual`. [Accessed: Oct. 02, 2023]

[188] "Glossary: Sex," European Institute for Gender Equality. Available: `https://eige.europa.eu/publications-resources/thesaurus/terms/1048`. [Accessed: Sep. 25, 2023]

[189] T. B. Fitzpatrick, "The Validity and Practicality of Sun-Reactive Skin Types I Through VI," Arch. Dermatol., vol. 124, no. 6, pp. 869–871, Jun. 1988, doi: 10.1001/archderm.1988.01670060015008.

[190] A. S. Jackson, S. N. Blair, M. T. Mahar, L. T. Wier, R. M. Ross, and J. E. Stuteville, "Prediction of functional aerobic capacity without exercise testing," Med. Sci. Sports Exerc., vol. 22, no. 6, pp. 863–870, Dec. 1990, doi: 10.1249/00005768-199012000-00021.

[191] P. J. Colvonen, "Response To: Investigating sources of inaccuracy in wearable optical heart rate sensors," Nature Publishing Group, 2 2021, pp. 1–2. doi: 10.1038/s41746-021-00408-5.

[192] C. Barr, "Comparison of Accuracy and Diagnostic Validity of a Novel Non-Invasive Electrocardiographic Monitoring Device with a Standard 3 Lead Holter Monitor and an ECG Patch over a 24 hours Period," Journal of Cardiovascular Diseases & Diagnosis, 2019.

[193] A. Cardona, T. T. Harfi, H. N. Nagaraja, M. Tong, and U. Paliani, "Enhanced wearability experience and accuracy of a novel FDA-validated noninvasive ambulatory ECG device compared with a conventional Holter monitor: The qardio-ECG study," J. Am. Coll. Cardiol., vol. 79, no. 9, p. 2053, Mar. 2022, doi: 10.1016/s0735-1097(22)03044-3.

[194] "EVAL-HCRWATCH4Z User Guide," Analog Devices, 2021. Available: `https://www.analog.com/media/en/technical-documentation/user-guides/eval-hcrwatch4z-ug-1221.pdf`

[195] "ADXL362 Datasheet," Analog Devices, 2023. Available: `https://www.analog.com/media/en/technical-documentation/data-sheets/adxl362.pdf`. [Accessed: Oct. 05, 2023]

[196] "NTCG104EF104FTDSX Datasheet," TDK Corporation, 2023. Available: https://product.tdk.com/system/files/dam/doc/product/sensor/ntc/chip-ntc-thermistor/data_sheet/datasheet_ntcg104ef104ftdsx.pdf

[197] "ADPD4100/4101 Datasheet," Analog Devices, 2020. Available: https://www.analog.com/media/en/technical-documentation/data-sheets/adpd4100-4101.pdf

[198] "VEMD8080 Datasheet," Vishay, 2023. Available: https://www.vishay.com/docs/84565/vemd8080.pdf. [Accessed: Oct. 05, 2023]

[199] "IR Compact Line Datasheet," Luxeon, 2018. Available: https://lumileds.com/wp-content/uploads/files/DS190.pdf. [Accessed: Oct. 05, 2023]

[200] "Z Color Line Datsheet," Luxeon, 2021. Available: https://lumileds.com/wp-content/uploads/files/DS105-LUXEON-Z-Color-Line-datasheet.pdf

[201] "nRF52840 BLE Dongle Product Specification," Nordic Semiconductor , 2021. Available: https://infocenter.nordicsemi.com/pdf/nRF52840_PS_v1.7.pdf

[202] J. Jorge et al., "Technical Note: Solving the problem of inter-device time delays," University of Oxford, 2019.

[203] R. Xiao, C. Ding, and X. Hu, "Time Synchronization of Multimodal Physiological Signals through Alignment of Common Signal Types and Its Technical Considerations in Digital Health," J. Imaging Sci. Technol., vol. 8, no. 5, Apr. 2022, doi: 10.3390/jimaging8050120.

[204] Y. Jiang et al., "EventDTW: An Improved Dynamic Time Warping Algorithm for Aligning Biomedical Signals of Nonuniform Sampling Frequencies," Sensors, vol. 20, no. 9, May 2020, doi: 10.3390/s20092700.

[205] F. C. Bennis, C. van Pul, J. J. L. van den Bogaart, P. Andriessen, B. W. Kramer, and T. Delhaas, "Artifacts in pulse transit time measurements using standard patient monitoring equipment," PLoS One, vol. 14, no. 6, p. e0218784, Jun. 2019, doi: 10.1371/journal.pone.0218784.

[206] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," IEEE Trans. Biomed. Eng., vol. 33, no. 12, pp. 1157–1165, Dec. 1986, doi: 10.1109/tbme.1986.325695.

[207] P. Hamilton, "Open source ECG analysis," in Computers in Cardiology, IEEE, 2003, pp. 101–104. doi: 10.1109/cic.2002.1166717.

[208] I. I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold," Biomed. Eng. Online, vol. 3, no. 1, p. 28, Aug. 2004, doi: 10.1186/1475-925X-3-28.

[209] M. Elgendi, M. Jonkman, and F. DeBoer, "Frequency bands effects on qrs detection," in Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing, SciTePress - Science and and Technology Publications, 2010, pp. 428–431. doi: 10.5220/0002742704280431.

[210] V. Kalidas and L. Tamil, "Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Institute of Electrical and Electronics Engineers Inc., Oct. 2017, pp. 457–461. doi: 10.1109/BIBE.2017.00-12.

[211] B. Porr, L. Howell, I. Stournaras, and Y. Nir, Popular ECG R peak detectors written in python. 2023. doi: 10.5281/zenodo.7652725. Available: `https://zenodo.org/record/7652725`.

[212] U. K. Okoji, S. C. Taylor, and J. B. Lipoff, "Equity in skin typing: why it is time to replace the Fitzpatrick scale," Br. J. Dermatol., vol. 185, no. 1, pp. 198–199, Jul. 2021, doi: 10.1111/bjd.19932.

[213] P. J. Colvonen, P. N. DeYoung, N.-O. A. Bosompra, and R. L. Owens, "Limiting racial disparities and bias for wearable devices in health science research," Sleep, vol. 43, no. 10, pp. 1–16, Oct. 2020, doi: 10.1093/sleep/zsaa159.

[214] B. Bent, O. M. Enache, B. Goldstein, W. Kibbe, and J. P. Dunn, "Reply: Matters Arising 'Investigating sources of inaccuracy in wearable optical heart rate sensors,'" NPJ Digit Med, vol. 4, no. 1, p. 39, Feb. 2021, doi: 10.1038/s41746-021-00409-4.

[215] F. Q. Nuttall, "Body Mass Index: Obesity, BMI, and Health: A Critical Review," Nutr. Today, vol. 50, no. 3, pp. 117–128, May 2015, doi: 10.1097/NT.0000000000000092.

[216] NHS, "Body Mass Index," nhs.uk, 2023. Available: `https://www.nhs.uk/live-well/healthy-weight/bmi-calculator/`. [Accessed: Oct. 05, 2023]

[217] D. Shookster, B. Lindsey, N. Cortes, and J. R. Martin, "Accuracy of Commonly Used Age-Predicted Maximal Heart Rate Equations," Int. J. Exerc. Sci., vol. 13, no. 7, pp. 1242–1250, Sep. 2020.

[218] T. T. Um et al., "Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring using Convolutional Neural Networks," arXiv [cs.CV], Jun. 2017.

[219] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," Brief. Bioinform., vol. 23, no. 2, Mar. 2022, doi: 10.1093/bib/bbab569.

[220] V. Radu et al., "Multimodal Deep Learning for Activity and Context Recognition," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, pp. 1–27, Jan. 2018, doi: 10.1145/3161174.

[221] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing," in Proceedings of the 26th International Conference on World Wide Web, in WWW '17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 351–360. doi: 10.1145/3038912.3052577.

[222] S. Yao et al., "RDeepSense: Reliable Deep Mobile Computing Models with Uncertainty Estimations," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 1, no. 4, pp. 1–26, Jan. 2018, doi: 10.1145/3161181.

[223] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv [cs.LG], Feb. 11, 2015.

[224] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15, no. 56, pp. 1929–1958, 2014,

[225] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," J. R. Stat. Soc. Series B Stat. Methodol., vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[226] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," arXiv [cs.LG], Mar. 12, 2020.

[227] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, and L. Invernizzi, "Keras-Tuner." 2019. Available: `https://keras.io/keras_tuner/`. [Accessed: Nov. 21, 2023]

[228] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[229] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv [cs.LG], Sep. 15, 2016.

[230] J. Gawlikowski et al., "A Survey of Uncertainty in Deep Neural Networks," arXiv [cs.LG], Jul. 07, 2021.

[231] K. Fang, D. Kifer, K. Lawson, and C. Shen, "Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions," Water Resour. Res., vol. 56, no. 12, p. e2020WR028095, Dec. 2020, doi: 10.1029/2020wr028095.

[232] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Inf. Fusion, vol. 76, pp. 243–297, Dec. 2021, doi: 10.1016/j.inffus.2021.05.008.

[233] N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 6417–6428.

[234] S. Mohamed and B. Lakshminarayanan, "Learning in Implicit Generative Models," arXiv [stat.ML], Oct. 11, 2016.

[235] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach," vol. 80, pp. 4075–4084, 10–15 Jul 2018.

[236] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," arXiv [cs.CV], Mar. 15, 2017.

[237] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in Proceedings of The 33rd International Conference on Machine Learning, M. F. Balcan and K. Q. Weinberger, Eds., in Proceedings of Machine Learning Research, vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059.

[238] Y. Gal, J. Hron, and A. Kendall, "Concrete Dropout," arXiv [stat.ML], May 2017.

[239] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," arXiv [stat.ML], Dec. 05, 2016.

[240] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv [cs.SD], Sep. 12, 2016.

[241] C. Szegedy et al., "Going Deeper with Convolutions," arXiv [cs.CV], Sep. 17, 2014.

[242] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based Framework for Multivariate Time Series Representation Learning," arXiv [cs.LG], Oct. 06, 2020.

[243] H. J. Davies, J. Monsen, and D. P. Mandic, "Interpretable pre-trained transformers for heart time-series data," arXiv [cs.LG], Jul. 30, 2024.

[244] H. Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," arXiv [cs.LG], Dec. 14, 2020.

[245] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv [cs.CV], Dec. 10, 2015.

[246] E. K. Naeini, F. Sarhaddi, I. Azimi, P. Liljeberg, N. Dutt, and A. M. Rahmani, "A Deep Learning–based PPG Quality Assessment Approach for Heart Rate and Heart Rate Variability," ACM Trans. Comput. Healthcare, vol. 4, no. 4, pp. 1–22, Nov. 2023, doi: 10.1145/3616019.

[247] Z. Meng, X. Yang, X. Liu, D. Wang, and X. Han, "Non-invasive blood pressure estimation combining deep neural networks with pre-training and partial fine-tuning," Physiol. Meas., vol. 43, no. 11, Nov. 2022, doi: 10.1088/1361-6579/ac9d7f.

[248] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep Evidential Regression," arXiv [cs.LG], Oct. 07, 2019.

[249] N. Durasov, N. Dorndorf, H. Le, and P. Fua, "ZigZag: Universal Sampling-free Uncertainty Estimation Through Two-Step Inference," Transactions on Machine Learning Research, Jan. 16, 2024.