UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Remote Heart Rate Estimation Leveraging Eulerian Video Magnification

Gonçalo Martins Rodrigues

Mestrado em Engenharia Biomédica e Biofísica

Dissertação orientada por:
Nuno M. Garcia

2024

”All models are wrong, but some are useful”

George E. P. Box

Ackowledgements

I would like to express my deepest gratitude to everyone who supported me throughout this journey over the past five years. First, to my family, especially my parents and sister, who have been my greatest pillars of support, not only in this final chapter but at every step along the way. Your unwavering belief in me was essential, and I could not have completed this chapter without your unconditional love and encouragement. It is because of you that I am who I am, and you will always be my strength. Thank you from the bottom of my heart. A special note of thanks to you, Andreia, for being an incredible source of support and strength in these past years.

I am profoundly grateful to my closest friends, who witnessed this journey firsthand. You stood by me throughout these years, and I know how challenging that can be. Being able to call you my friends is truly a blessing, and I am immensely thankful for your presence in this chapter of my life. To those friends who have drifted away over time, know that you, too, left your mark and influenced the path I followed.

My sincere thanks go to all the professors who were fundamental to my academic journey, particularly those at IBEB, who nurtured my passion for knowledge. A special thanks to Professor Nuno M. Garcia for his invaluable support and constant availability throughout this project. His expertise and guidance made completing this work possible.

Lastly, I thank myself for my resilience and commitment to excellence, for maintaining an insatiable curiosity and a drive to learn. This journey has been as much about personal growth as academic achievement, and I am proud of how far I have come.

Abstract

Introduction: Measuring physiological signals like Heart Rate, Respiration Rate, and Blood Pressure is essential for health assessment but often requires physical contact, which is a limitation emphasized by situations such as the SARS-CoV-2 pandemic. This constraint has spurred the development of remote monitoring solutions, with Remote Photoplethysmography (rPPG) being a prominent method. Via captured video, rPPG detects subtle color changes in the skin due to blood flow, though it is sensitive to noise like motion and lighting variations. To enhance robustness, the study investigates the use of Eulerian Video Maggnification (EVM), which amplifies color changes for better explainability and signal processing.

Objectives: The study aimed to develop a robust remote Heart Rate estimation model using EVM and deep learning. It also aimed to create a public dataset of facial videos, supporting further research in this field.

Methods: The dataset included videos with Electrocardiogram as ground truth and additional subject/environmental details. The EVM technique followed Wu's original methodology, emphasizing color magnification to highlight subtle skin changes [1]. The deep learning model, incorporates 1D CNNs and LSTM layers for effective temporal pattern recognition. Performance was assessed using MAE, MAPE, and RMSE metrics.

Results: EVM showed potential but suffered from performance drops under noise like lighting changes, particularly at higher heart rates. The forehead region yielded better results than cheeks. The proposed model achieved a significant performance improvement over baselines, with an MAE of 4.57 ± 0.87 bpm and overall better performance when trained on the forehead region. Cross-region training improved results but with variable success.

Conclusion: EVM and rPPG remain sensitive to noise and require either controlled conditions or adaptations for practical use. The deep learning model demonstrated improvements over baseline methods but was affected data variability, highlighting the need for further refinement.

Keywords — Remote Photoplethysmography (rPPG), Eulerian Video Magnification (EVM), Heart Rate Estimation, Deep Learning.

Resumo

Introdução: A medição dos sinais fisiológicos como o Batimento Cardíaco, o Ritmo Respiratório e a Pressão arterial estão altamente correlacionados com a saúde de cada indivíduo e a sua medição tornou-se crucial no ambiente clínico. Estas medições apresentam uma grande limitação: a necessidade de contacto físico para conseguir aferi-los de forma correta. Esta condição não pode ser assegurada sempre, e a pandemia do SARS-CoV-2 só veio enfatizar isso, sendo esta uma das grandes limitações da telemedicina. Esta limitação combinada com a necessidade crescente de monitorizar estes parâmetros, tanto no ambiente hospitalar como fora, estimularam a pesquisa de soluções alternativas remotas.

Uma das técnicas mais pesquisadas é a fotopletismografia remota. Esta técnica permite realizar a medição destes parâmetros de modo remoto através de captura de vídeo. Esta analisa a ligeira mudança de cor na superfície da pele gerada pelo fluxo de sangue nos capilares. Embora invisível a olho nu, esta mudança pode ser rastreada através de técnicas de processamento de vídeo e sinal, e utilizada para medir estes parâmetros. Apesar dos avanços recentes nesta área, esta técnica continua altamente experimental no espaço médico devido às suas várias limitações, nomeadamente a sensibilidade ao ruído provocado por fatores como o movimento e iluminação. A pesquisa nesta área foca-se em reduzir os efeitos do ruído, adicionando camadas para mitigar o impacto, mas com sucessos limitados.

Uma das abordagens reportadas na literatura que pode possivelmente ter um efeito positivo é a Magnificação Euleriana de Vídeo (EVM). Esta técnica visual realça as ligeiras mudanças de cor nos vídeos e, como tal, ajuda a interpretar os resultados obtidos pela técnica de fotopletismografia. Sugere-se ainda que a incorporação desta técnica no algoritmo de fotopletismografia remota pode ser vantajosa devido ao processamento de sinal realizado, adicionando robustez.

De modo a analisar melhor as limitações da fotopletismografia remota e as típicas características da técnica, particularmente quando é aplicada em conjunto com EVM, foi feita uma revisão extensiva da literatura de acordo com o modelo PRISMA. Nesta foram pesquisados 238 artigos, dos quais 32 foram incluídos na revisão. Nesta revisão foram listadas as principais características dos 32 artigos com o objetivo de perceber que fatores mais influenciam os resultados.

Objetivos: O principal objetivo deste trabalho é investigar e desenvolver um modelo de estimação remota de sinais fisiológicos a partir de captura de vídeo mais robusto, aproveitando as vantagens da técnica de EVM e do uso de Aprendizagem Profunda. A Aprendizagem Profunda, um tipo de Aprendizagem Automática, tem registado um crescimento significativo nos últimos anos. O objetivo é otimizar um modelo deste tipo para melhorar os resultados, mantendo baixas as necessidades computacionais, um equilíbrio difícil de alcançar.

O segundo objetivo é criar um conjunto de vídeos faciais que permita o estudo nesta área, com a intenção de o disponibilizar publicamente. A motivação por detrás deste segundo objetivo é contornar a

escassez de recursos disponíveis online. Ao criar um conjunto de dados público e ao partilhar o código utilizado, pretende-se aumentar a transparência do estudo realizado e expandir os recursos disponíveis na área.

Métodos: Para iniciar o desenvolvimento do modelo, foi necessário criar o conjunto de dados. Este conjunto foi previamente aprovado pela comissão de ética de ciências (CEC) e todo o procedimento seguiu as diretrizes impostas, pois o objetivo era posteriormente publicá-lo. Não houve critérios restritivos para a seleção dos sujeitos, procurando-se maximizar a variabilidade e, assim, a generalização dos dados. Foram impostas condições naturais e menos controladas de gravação, recorrendo à iluminação natural durante a recolha, o que adiciona complexidade aos dados e permite avaliar a robustez dos modelos testados. Além dos vídeos, foi gravado o sinal de electrocardiograma (ECG) para servir como referência real. Também foram anotadas características dos sujeitos, como idade, sexo, altura, entre outras, e fatores ambientais, como temperatura, humidade e iluminação. O equipamento de medição, o procedimento e a configuração do espaço físico foram padronizados e mantidos constantes para todos os sujeitos. Os sinais e vídeos não foram armazenados no estado bruto; passaram por processamento de sinal/imagem, incluindo a divisão em intervalos de 20 segundos (a janela de estimação desejada) e a eliminação de intervalos com artefactos significativos. Os vídeos no conjunto de dados correspondem apenas às regiões de interesse: testa e bochechas; garantindo a anonimidade dos sujeitos por questões éticas. Ainda mais, foi preparado, para cada sujeito, um ficheiro com o batimento cardíaco médio por intervalo. O conjunto de dados está disponível publicamente mediante pedido aos autores, com um ficheiro adicional detalhando a sua organização.

A metodologia restante foi dividida em duas partes, correspondendo às técnicas de magnificação de vídeo e de estimação do batimento cardíaco. A magnificação de vídeo foi implementada segundo o artigo original [1], que propõe duas abordagens de EVM: magnificação de cor e de movimento. Conforme o objetivo imposto e com a literatura, foi usada a magnificação de cor. De forma simplificada, o modelo pode ser descrito em quatro etapas principais: decomposição espacial, filtragem temporal, amplificação e reconstrução do vídeo. Como não existe um parâmetro objetivo de avaliação, a eficácia da técnica foi avaliada de forma algo subjetiva e, neste caso, foi avaliada pela sua capacidade de revelar mudanças de cor de forma fiel à realidade.

Para a estimativa do batimento cardíaco, foi usada uma abordagem de Aprendizagem Profunda, em vez das técnicas tradicionais de processamento de sinal. A combinação desta técnica com a magnificação visa melhorar a robustez do modelo. Antes da reconstrução do vídeo, a série temporal amplificada é guardada e posteriormente usada pelo modelo para prever o batimento cardíaco. A arquitetura do modelo inclui convoluções 1D e uma camada de memória de curto longo prazo (do inglês ”Long Short Term Memory”), camadas especializadas em capturar padrões temporais, permitindo ao modelo prever o batimento cardíaco ao identificar dependências temporais. O modelo foi treinado com 60% dos dados e dos restantes, 20% foram usados para validação e otimização do mesmo, e 20% para uma avaliação final da sua generalização em dados não analisados até então. A avaliação do modelo foi feita usando métricas de erro: Erro Absoluto Médio (MAE), Erro Percentual Absoluto Médio (MAPE) e Raiz do Erro Quadrático Médio (RMSE). Foram também desenvolvidos dois modelos de referência, com os quais se comparou o modelo proposto para medir o seu desempenho.

Resultados: Os resultados foram analisados separadamente para a magnificação de vídeo e a estimação do batimento cardíaco. Inicialmente, validou-se o modelo de magnificação, verificando-se a reprodução fiel dos resultados do artigo original. Posteriormente, avaliou-se o modelo no conjunto de dados

criado. Embora tenha funcionado corretamente, o modelo sofreu quedas significativas de desempenho devido ao ruído presente nos dados recolhidos. Esse ruído foi causado por fatores que comprometeram a refletância da pele, entre estes a luminosidade durante a gravação e o tom da pele de cada indivíduo. Além disso, amostras com batimentos cardíacos mais elevados mostraram-se mais suscetíveis à interação com o ruído. Foi também realizada uma análise do impacto de cada parâmetro da técnica nos resultados, determinando-se que a amplitude de frequências amplificadas tem o maior potencial de redução de ruído, embora com algumas nuances. Os demais parâmetros afetaram apenas a qualidade visual do resultado, com pouco impacto no ruído. Foram ainda testadas as diferentes regiões de interesse, onde os resultados mostraram variabilidade dependendo da região utilizada, com a testa demonstrando claramente um desempenho superior às outras. Este provém da diferente quantidade de ruído presente em cada uma delas e não de uma vantagem inerente à área em específico.

A seguir, avaliou-se o desempenho da estimação. No caso da testa, o modelo proposto obteve um desempenho muito superior aos modelos de referência. Os resultados foram um MAE de $4,57 \pm 0,87$, um MAPE de $5,93 \pm 1,21$ e um RMSE de $6,34 \pm 1,07$, representando uma melhoria geral de cerca de 30% em relação ao melhor modelo base testado. Estes resultados também apresentaram grande variabilidade entre as regiões de interesse, sendo que as restantes regiões não conseguiram superar todos os modelos base devido à dificuldade do modelo em aprender a variância dos dados nesses casos. Esta foi mais uma vez devida à presença de ruído em quantidades diferentes em cada uma das regiões estudadas. Foi ainda testada uma técnica de treino inter-regional, onde o modelo foi treinado na região da testa e testado nas outras regiões, o que permitiu a melhoria dos resultados. No entanto, apenas a bochecha esquerda conseguiu superar todos os métodos base com esta abordagem devido à quantidade de ruído superior na bochecha direita.

Conclusão: Embora a aplicação da magnificação tenha funcionado corretamente, o seu desempenho não é robusto perante grandes quantidades de ruído e é altamente influenciado por fatores como a intensidade da iluminação. Os resultados obtidos refletem a dificuldade dos dados utilizados para esta técnica, que ainda requer condições controladas ou o desenvolvimento de novas abordagens para se adaptar ao ruído. Relativamente à tarefa de estimação, o modelo superou os modelos de referência, mas de forma inconsistente, possivelmente devido a um viés gerado pela otimização. No geral, os resultados não foram excelentes, mas foram elucidativos, destacando as principais limitações destas técnicas.

Palavras-chave — Fotopletismografia Remota (rPPG), Magnificação Euleriana de Vídeo (EVM), Estimação do Batimento Cardíaco, Aprendizagem Profunda.

Contents

List of Figures

List of Tables

Acronyms

ACMD Adaptive Chirp Mode Decomposition

AI Artificial Intelligence

BP Blood Pressure

bpm beats per minute

BSS Blind Source Separation

BVP Blood Volume Pulse

CCA Canonical Component Analysis

CEC-FCUL Ethics Committee of the Faculty of Sciences at the University of Lisbon

CEEMDAN Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

CNN Convolutional Neural Network

DWT Discrete Wavelet Transform

ECG Electrocardiogram

EEMD Ensemble Empirical Mode Decomposition

EVM Eulerian Video Magnification

FFT Fast Fourier Transform

fps frames per second

HR Heart Rate

HRV Heart Rate Variability

ICA Independent Component Analysis

IR Infrared

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

NIR Near-Infrared

OpenCV Open Computer Vision Library

PCA Principle Component Analysis

PPG Photoplethysmography

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

ROI Region of Interest

rPPG Remote Photoplethysmography

RR Respiration Rate

SD Standard Deviation

SNR Signal-to-Noise Ratio

ST Map Spatio-Temporal Map

Chapter 1

Introduction

1.1 Motivation

The rapid advancement of Artificial Intelligence (AI) has led to the development of various highly impactful applications in the healthcare field such as image analysis, medical device automation, and patient monitoring [2]. AI has always fascinated me, particularly due to its ability to solve problems with complex internal patterns that the human mind cannot easily comprehend by itself [3]. My interest in this field emerged spontaneously and has since been nurtured by numerous courses during my academic journey. I believe that this technology will significantly change our perception of problem solving and will enable us (or perhaps it has already begun) to take a new step towards a new era of innovation.

Although the methods used are influenced by my personal preferences, the context that led me to address this specific topic has a different origin. The motivation for the topic approached in this dissertation was fueled by the pandemic faced in 2020. This outbreak had a huge impact on several factors in society, redefining some of the basic beliefs of modern healthcare and leading to the search for solutions to newly emerging problems [4]. The pandemic highlighted various weaknesses in our healthcare systems [5], while also exposing limitations in some of the standard methods that are used in healthcare daily. A fundamental aspect of patient assessment, measuring the vital signs [6] such as Heart Rate (HR), Respiration Rate (RR), Blood Pressure (BP), etc., presents a significant limitation: conventional equipment requires direct contact with the patient's skin. For patients quarantined at home, this posed challenges due to the constraints of telemedicine [7] and lack of techniques to monitor their vital parameters. Meanwhile, for hospitalized patients, the need for close contact increased the risk of infection for healthcare professionals [8], adding to the burden on hospitals. This posed a major challenge during the pandemic, emphasizing that new approaches must be found, like making such type of equipment more accessible to the population, which can be provided through devices such as smartwatches [9], or either trough adapting the existing methods to overcome this difficulty, which can be done by skewing the research towards non-contact monitoring of the vital parameters. The search for different approaches is not new and situations like the pandemic have merely brought the problem into sharper focus.

In this dissertation, an approach to estimate HR remotely using Eulerian Video Magnification (EVM) and Remote Photoplethysmography (rPPG), combined with a deep learning estimation procedure, was developed. This research aimed to draw attention to the topic and advance the existing literature and research in this area. It addresses a topic with gaps in the current literature and has the potential to impact

remote health monitoring, particularly in resource-constrained environments. Additionally, a dataset was developed as part of this research, which is intended to be made publicly available to address the scarcity of data for research on this topic. Further research in this area could open up new possibilities, ranging from more effective methods for measuring vital signs to customizable cybersecurity applications based on vital parameters

Ultimately, this dissertation represents a significant step toward my long-term goal of contributing to the development of AI-driven healthcare solutions that are both innovative and impactful.

1.2 Objectives

The primary objective of this master's dissertation is to investigate and develop a more robust non-contact method for vital sign monitoring, particularly HR, from video recordings using the Eulerian Video Magnification technique. Other significant contributions of this dissertation include:

1. Systematically reviewing the literature to analyze the current state of the art in the application of techniques utilizing EVM;
2. Optimizing the model by integrating EVM with efficient machine learning techniques;
3. Developing a custom dataset of facial videos featuring various subjects and their corresponding heart rates;
4. Implementing a fully automated preprocessing pipeline for video data to prepare it for the estimation model;
5. Evaluating the efficiency and effectiveness of the proposed model, particularly in natural lightning conditions;
6. Discuss future work and research directions.

1.3 Structure of the Document

This dissertation is divided into five distinct chapters. This chapter introduces the dissertation's topic, its motivation, and outlines the structure of the document. Chapter 2 provides a brief contextualization of the theoretical themes related to the methods employed throughout the dissertation, as well as, a systematic literature review based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology. Chapter 3 details the methodologies used, describing each component of the work individually as well as the integration of all the steps. Chapter 4 presents the results obtained during the research and their discussion. Finally, in Chapter 5, a final assessment is made, considering the challenges and limitations of the work, as well as future prospects.

Chapter 2

State of the Art

2.1 Introduction

The relationship between physiological signals, such as Heart Rate (HR), Respiration Rate (RR), Blood Pressure (BP), etc., and an individual's health has been well established since the previous century [10]. Analyzing these signals and their variations can provide critical insights into a patient's condition. Among these signals, HR stands out as a key indicator, capable of monitoring both cardiac and non-cardiac diseases and even serving as a predictor of mortality [11]. Given its significance, it is essential to monitor this vital sign, particularly in routine medical settings. Typically, HR is measured using one of two methods: Electrocardiogram (ECG) [12] or Photoplethysmography (PPG) [13]. Although these methods are based on distinct concepts and focus on entirely different physical properties, both yield highly reliable results. However, a major limitation of these techniques is their dependency on physical contact, yielding problems especially with babies, patients with burnt or sensitive skin, cognitive impairments, those undergoing certain medical procedures or severe illnesses [14], and even quarantined people as was seen in the recent SARS-CoV-2 pandemic. The constraints imposed by this requirement, combined with the growing need for continuous monitoring of vital signs both within and outside hospital environments, have spurred research into alternative, non-contact methods. This trend is mirrored in the monitoring of other physiological signals such as blood pressure and respiratory rate.

The extraction of these physiological signals in a non-contact manner, particularly through video capture, is known as Remote Photoplethysmography (rPPG). The human visual system exhibits limited sensitivity to minor temporal variations, even though various signals with significant information exist at this level. For instance, during a heartbeat, subtle and often imperceptible changes occur, such as slight alterations in facial skin color, momentarily causing a shift in facial hue. Similarly, some imperceptible small movements can be induced by respiration. rPPG is a method capable of analysing this imperceptible changes and, by tracking them, reconstruct the underlying biomedical signal. Through signal and image processing it is then possible to determine the HR remotely. The initial evidence for the use of this technique to determine HR dates back to 2008 [15]. Even though it is a seemingly affordable technology, rPPG remains largely experimental, with limited application in medical settings due to significant environmental constraints. A primary challenge is the susceptibility of the extracted signal to noise, which arises from sensitivity to lighting conditions and motion artifacts [14]. Although recent advancements in fields such as computer vision, signal processing, and image processing have mitigated some of these limitations, rPPG's adoption in clinical environments remains limited. Some examples are the refinement

of methods such as Region of Interest (ROI) tracking and the development of Eulerian Video Magnification (EVM), which allowed for significant advancements. These were reported to enable the overcoming of challenges related to subject movement, lighting conditions and robustness overall [15, 16].

Eulerian Video Magnification, proposed by Wu et al. in 2012, is a Visual Computing method used to amplify slight temporal variations in videos [1]. By amplifying imperceptible physical phenomena, such as slight changes in color or small movements both caused by the heartbeat, it enables their real-time visualization. Although EVM is not strictly necessary for measuring physiological variables, it is suggested to serve as a tool that can enhance the robustness of rPPG algorithms. This type of visualization aids in explainability, as they often complement each other, allowing for a better understanding of how the algorithm obtains its results. While EVM is primarily a visualization technique rather than an estimation method, it is closely associated with rPPG, and both are often integrated into the same processing pipeline. Furthermore, EVM may be advantageous for estimating physiological parameters that rely on spatial information, leveraging its ability to preserve spatial details during video reconstruction and to amplify motion. Examples include the estimation of respiratory rate (through motion magnification) [17, 18] or the diagnosis of health issues affecting blood circulation patterns (using spatial information) like BP [19]. However, despite EVM asnwering to some of rPPG's limitations, it also has its trade-offs. Due to the amplification of a frequency spectra, certain types of noise might also be magnified, worsening the same issue it tries to solve.

EVM and rPPG methods have been gaining popularity in recent years. The significant need to develop cost-effective and non-invasive techniques for collecting certain biometric characteristics has led to an increasing demand for remote practices. Due to their substantial recent development, they still face some obstacles that complicate their widespread use. This review aims to better understand these techniques and their respective challenges. Although there are other reviews addressing similar themes, in the searched databases, no review has exclusively focused on EVM methods for facial videos.

This chapter undertakes a comprehensive and systematic review of EVM methods employed in the extraction of biometric features from facial videos over recent years. The objective is to provide an in-depth analysis that not only serves as a foundation for subsequent research endeavors in this domain but also facilitates the comprehension of the present landscape within the field. This systematic literature review was conducted following the updated guidelines of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [20, 21].

The main contributions of this chapter are as follows:

- It summarizes the existing studies regarding the estimation of biometric features that employ EVM. The data for analysis is drawn from five distinct databases;
- It presents the predominant advantages and disadvantages, addressing parameters such as databases, estimation algorithms, and other processing techniques within the examined studies;
- It meticulously analyses the performance by incorporating reported metrics. This analysis establishes a benchmark for comparing the performance of newly developed algorithms.

The remainder of this chapter is organized as follows: this paragraph concludes the Introduction where the motivation and the contributions of this review were presented; the Methods section describes in detail the dynamics of the search, selection of articles and their synthesis as well as defining research questions; next, in the Results section, the most important characteristics of each article have been sum-

marized to facilitate understanding of the state of the art; the Discussion revisits the research questions; and finally, in the Conclusion, some limitations and future prospects of this review are highlighted.

2.2 Methods

In this section, a detailed description is provided on how this review was carried out. With the overarching research interests in view, this section starts by presenting the research questions, upon which the entire research and article eligibility criteria were based. Subsequently, the methods of search, selection, information retrieval, and synthesis will be analyzed thoroughly, in this order.

2.2.1 Research Questions

The search for articles focused on addressing the following research questions:

- (RQ1) What biometric features can be estimated through facial videos?
- (RQ2) How do the different methods affect the model's performance?
- (RQ3) What are the factors that can induce error in the estimation of such biometric features?

While the primary focus of this search was on the estimation of HR and its variability as a well-known biometric feature, it remained crucial to explore what other features could be assessed using this technique, hence the first research question (RQ1).

Both (RQ2) and (RQ3) focus on the most commonly employed models for estimation and their respective performances. At the same time, these research questions aim to understand the typical limitations of such models.

2.2.2 Eligibility Criteria

The searched studies had to meet specific criteria for inclusion in this review. Regarding the timeline, the analyzed studies were exclusively those from 2016 up to January 2024. Apart from 2016 being a common date barrier observed in other reviews, up until then, the few studies in this field were mainly proof-of-concept. Only after this date did emerge a surge in the exploration of more sophisticated methodologies based on EVM, since it was introduced in 2012. Furthermore, the primary objective was to explore the application of this technique in a broader context. Consequently, studies were omitted if they involved subjects with specific medical conditions or if the primary aim was to diagnose a particular ailment. Likewise, research involving exclusively individuals under the age of 18 was excluded for the same rationale. All other considered criteria [22] are summarized in Table 2.1.

Although not explicitly stated in the criteria presented in Table 2.1, the included studies are expected to demonstrate an application of EVM. Additionally, studies conducted on subjects other than humans (such as animals or objects) do not align with the research topic and will therefore be excluded.

Regarding the reported outcomes, it is anticipated that the included studies provide some form of performance metric. The majority of studies reported one or more of the following: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), or Root Mean Squared Error (RMSE). It was considered mandatory to report such metrics for the estimated feature, since it establishes a comparative baseline. When sufficient data were available, it was possible to infer the values of some of these metrics. As a result, certain articles that would initially have been excluded were considered in this review. This

Table 2.1: Summary of the Eligibility Criteria

Eligibility Criteria	Inclusion	Exclusion
Date	From 2016 onwards	All others
Exposure of interest	No condition/diagnose	Any condition or diagnose
Geographic Location	All	None
Language	English	Any other language
Participants	18 years and above	Below 18 years
Setting	All	None
Reported Outcomes	Performance Metrics: MAE, MAPE, RMSE	All others
Study Design	Region of Interest: Face	All others
Type of Publication	Journal and Conference Proceedings	All others

enabled a direct comparison between more studies, but nonetheless, it was still not possible to compare many of them due to reporting different outcomes.

2.2.3 Search Strategy

The databases and research engines employed to conduct the study searches, guided by reviews with similar themes and according to their main domains, were Scopus[1], Web of Science[2], PubMed[3], Xplore[4] and Science Direct[5]. The database search was completed in 31^{st} of January 2024.

The first two were selected for their multidisciplinary content. PubMed was included due to the nature of the research topic, given its extensive collection of biomedical articles. Similarly, Xplore was incorporated for its significant biomedical articles related to engineering topics. Lastly, ScienceDirect was chosen for its abundance of journal articles.

The search was conducted using a straightforward keyword approach that was executed consistently across all databases. The comprehensive line-by-line search strategy is detailed in Table 2.2. Initially, the search was conducted without any date criteria limitations. However, date criteria were subsequently added to the search for the reasons mentioned earlier. In the subsequent stage, only articles within the considered date range underwent screening.

Table 2.2: Search Strategy

	Search Strategy
1	(”Heart Rate” OR ”Biometry” OR ”Biometrics”)
2	(”Eulerian Video Magnification” OR ”Video Magnification”)
3	1 AND 2
4	(”Heart Rate Biometry”)
5	3 OR 4
6	Limit 5 to English language

[1]Elsevier | Scopus - Available at https://www.scopus.com/
[2]Clarivate | Web of Science - Available at https://www.webofscience.com/wos/woscc/basic-search
[3]NIH | PubMed - Available at https://pubmed.ncbi.nlm.nih.gov
[4]IEEE | Xplore Digital Library - Available at https://ieeexplore.ieee.org/Xplore/home.jsp
[5]Elsevier | Science Direct - Available at https://www.sciencedirect.com

It is noteworthy that no explicit search was conducted for parameters other than HR, this being the main focus. Although other parameters emerged in the search and were part of Heart Rate estimation studies, this aspect might present a limitation to the search strategy regarding (RQ1).

2.2.4 Study Selection

The objective of this review was to identify any application of EVM used for extracting biometric features from facial videos. The eligibility criteria, as previously mentioned, were broadly defined to enhance the sensitivity of the search and focus on pertinent articles. The study selection process was carried out by a single investigator individually, potentially introducing a limitation to the review by increasing the risk of overlooking relevant studies. To help navigate this situation, supervisor Nuno Garcia was consulted to avoid doubts during the screening process.

All identified studies in the databases underwent screening of titles and abstracts for eligibility. Only those aligning with the scope of this review and meeting the inclusion criteria progressed to the next phase. This step also allowed for the elimination of the retrieved duplicate articles. In the subsequent stage, their full texts were analysed for a more in-depth analysis of eligibility, in accordance with the criteria outlined in Table 2.1. This process was checked twice to avoid inconsistencies. Those that remained were utilized to conduct this review.

2.2.5 Data Collection

Relevant data from each article was extracted through a comprehensive reading of the full text by one investigator individually. For each article, the aim was to gather information on:

1. Study characteristics: title, citations, publication year, type of publication;
2. Population characteristics: age, gender, ethnicity;
3. Study design: estimated physiological parameters, addressed image artifacts, camera settings, estimation methodology, etc;
4. Performance: reported error performance metrics and the respective reference devices.

The performance metrics, even when not reported, could be inferred if there was sufficient available information to do so clearly. This was done using the formulas in Equations 2.1, 2.2, and 2.3. It is crucial that this process can be carried out clearly. Any case where data is not presented clearly was discarded due to the uncertainty of inferring the correct values of the metrics.

The following equations present the formulas used for MAE, MAPE, and RMSE, respectively:

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.3}$$

where y_i represents the actual value of the ith sample for a given parameter, $\hat{y}_i$ represents the predicted value for the ith sample, and n represents the number of samples in a given set.

2.2.6 Synthesis Methods

The data presentation in the results section mainly was built upon tables. Bar charts were also prepared to organize the collected information, and furthermore, a box plot was created to organize the performance metric data. The charts were generated using Microsoft Excel [23].

The synthesis was subdivided into four distinct sections, each with a focus on different aspects of the retrieved information. Firstly, the Study Characteristics which focus on general information extracted from the articles, such as the publication year, type of publication, and number of citations. Followed by details about the Population Characteristics utilized in each study. Then, the Study Design, encompassing its procedures, objectives, and imposed conditions. And lastly, the Performance Metrics are scrutinized, representing the outcomes presented in each paper.

2.3 Results

The search, as described in the previous section, yielded 280 articles from various databases. After removing duplicate articles (123) and others, mainly reviews (7) or articles published before the year 2016 (4), a total of 146 studies remained for further analysis. Subsequently, the titles and abstracts of the articles were screened, resulting in the exclusion of an additional 76 articles. The most common reason for their removal was their irrelevance to the research topic or being out of scope. Additionally, some articles, despite involving video magnification, did not employ an EVM-related approach, were aimed at diagnosing or treating a specific condition, focused on a particular subject group such as those in fetal monitoring, or were not human-related, etc. It is also worth noting that any application of EVM focusing on regions of interest other than the face was considered out of scope since the research primarily focused on this type of applications. Two articles were also excluded due to the unavailability of the full text [24] (no information could be retrieved from the second one). Following a comprehensive reading of the full text, 36 more articles were eliminated based on the eligibility criteria. Consequently, 32 articles were included in the review. The process can be observed in Figure 2.1. The PRISMA flow diagram was developed using an online open-access tool [25].

Some studies were included in the review despite potential concerns regarding their eligibility according to the defined criteria. The articles, along with the rationale for their inclusion, are as follows:

Articles [18, 26–28] were included despite involving subjects under the age of 18. Although this may appear to contradict the eligibility criteria, since the focus of these articles is not exclusively on this age group, their contributions can be generalized without causing concern;

The articles [28–32], for which performance metrics were not reported but had their data clearly presented, had their values inferred through equations 2.1, 2.2, 2.3. Another paper [33] did not report the results in their entirety; instead, it shared an ”accuracy” metric. As the process of calculating this metric was presented, it was still possible to infer the MAPE;

Authors of [34] solely reported MAPE, presenting it in bar charts without providing the exact values. Despite the inability to infer the precise values, it was retained due to its valuable contributions to the topic;

Several studies also presented applications for ROIs other than the face but were kept in the review due to their contributions for the facial applications: Neck [18, 35–37], Chest [17], Hand/Wrist [18, 30, 38] and others [18];

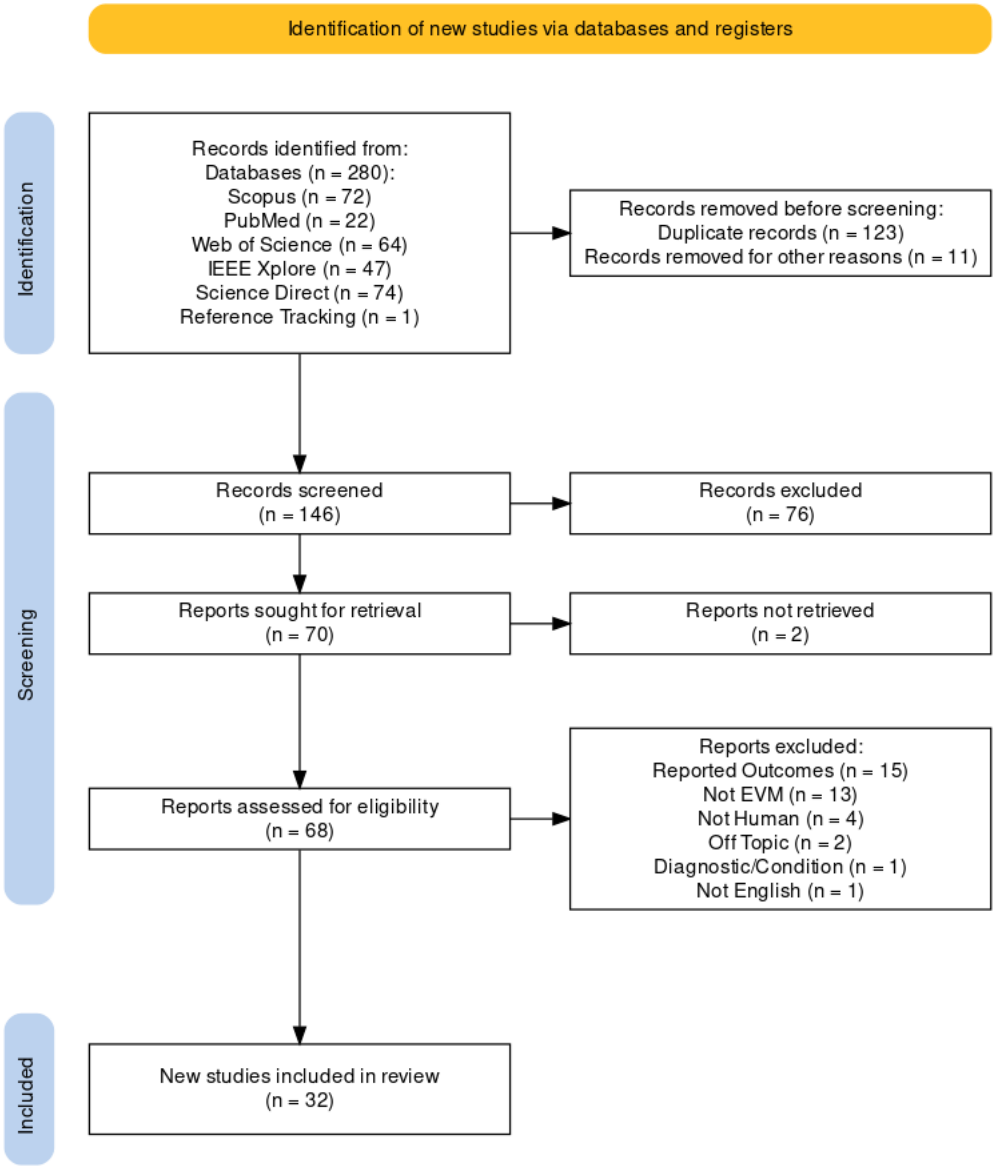


Figure 2.1: PRISMA Flow Diagram

Additionally, article [39] was removed due to presenting performance metrics unclearly, thereby hindering comprehension.

2.3.1 Study Characteristics

Statistical data regarding the characteristics of the analyzed studies have been organized. Out of the 32 studies included in the review, 19 (59.4%) of them are papers published in Conference Proceedings, while the remaining 13 (40.6%) are articles published in various journals (Table 2.3). Figure 2.2 represents a bar chart of the quantity of studies included in the review per year, categorizing them into Conference Proceedings and Journals. Similarly, the bar chart in Figure 2.3 displays the number of studies per publication citation, dividing them in the same manner.

Regarding the publication year, as observed in Figure 2.2, among the retrieved articles, the most recent ones are predominantly conference proceedings. Concerning the number of citations, journal articles had a higher number of citations, with the maximum observed value being a total of 106 citations for

Table 2.3: Publication Type

Type of Publication	Number of Studies	Studies
Journal Article	13	[17–19, 26, 27, 32, 35, 37, 40–44]
Conference Proceedings Paper	19	[28–31, 33, 34, 36, 38, 45–55]

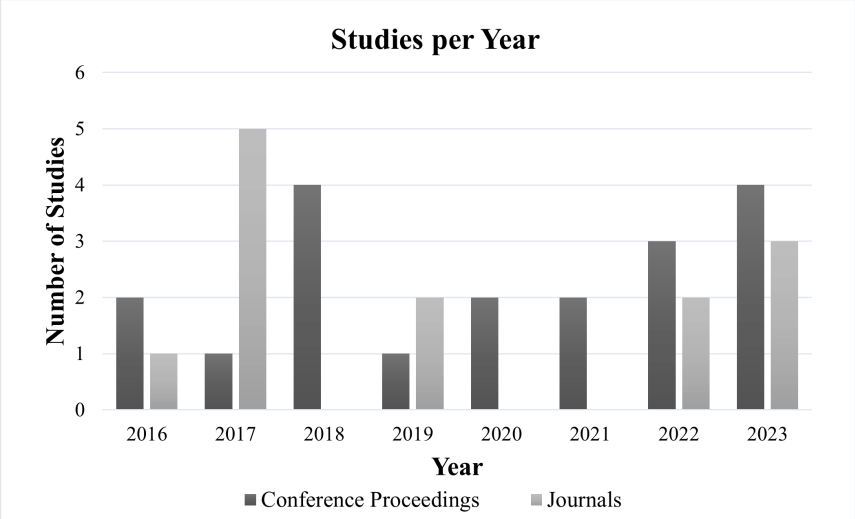


Figure 2.2: Bar Chart of Studies per Year

article [44], whereas the conference proceedings with the highest number of citations had 26 [31]. The rounded average citation count per journal article (28 citations) exceeded this value; whereas conference proceedings averaged 5 citations per paper.

2.3.2 Population Characteristics

Age and Gender

Among the included studies, only 18 (56.3%) reported the age of the participants. Among these, 5 reported the mean age of the individuals included. The minimum reported mean age was 22.5 years [36, 37], and the maximum was 44.5 years [29, 48], with the remaining study falling within this range (35.6 years, [17]). Furthermore, 15 reported the age range of the participants. By excluding the ages reported from studies [18, 26–28], which reported ages under 18 years, the overall age range of subjects included in the remaining studies was from 18 to 85 years old. Observing exclusively the previously mentioned articles, their age range was from 2 to 50 years old. Thus, the overall range of the subjects in the included studies was from 2 to 85 year old.

Regarding the gender of participants, it was reported by only 16 (50%) of the articles, while the other half did not share relevant information on this matter. The vast majority presented a sample predominantly male (12) or entirely male (1), with only 3 articles presenting a balance of both genders in their sample [36, 37, 40]. Of all the articles, 2 (6.25%) indicated having a sample exclusively of the same gender, with one specifying that they were all male [32], with no information provided for the other [45].

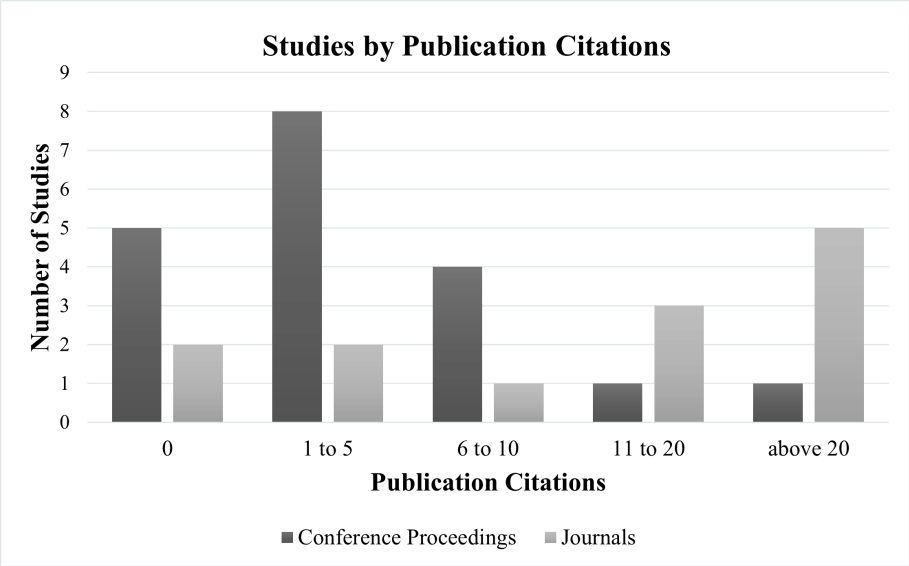


Figure 2.3: Bar Chart of Studies by Publication Citations

Ethnicity and Skin Tone

Numerous studies in the past have shown the significance of considering ethnicity or skin color due to darker skin tones posing more challenges for estimation tasks compared to white skin. Therefore, the information reported from each study on this matter was also analyzed. Out of the total articles, 12 (37.5%) reported presenting various skin tones and/or ethnicities and showed concern about displaying a diverse range of these characteristics due to the potential impact on the results. Only one study reported having subjects all from the same ethnicity (Chinese) [52]. The remaining studies did not share any information in this regard.

Number of Participants and Databases

The number of participants in each study was further analyzed and organized into a bar chart (Figure 2.4). Out of the total included studies, 25 (78.1%) provided information on the number of subjects included, where approximately half of the studies (13) were conducted with a cohort of 16 or more subjects, while the remainder utilized a cohort ranging from 1 to 15 subjects. The study with the highest number of participants had 48 subjects [40], and the minimum number was 1 [49], yielding a rounded average of 18 subjects per study. The mode was 40 with four studies [17, 41, 44, 45].

The databases used by each study were also analyzed. Out of the 32 articles, 5 (15.6%) did not report any information on this aspect, while 27 (84.4%) shared information that allowed understanding whether a public or existing database was used, or if they developed their own database. Among the 27 that shared information, 24 created their own dataset, and only 3 used existing databases. The databases utilized by these studies were COHFACE [56] in study [17], MR-NIRP-INDOOR [57] and UBFC-rPPG [58] in study [46], and MMSE-HR [59] in study [44].

2.3.3 Study Design

In this section, the main characteristics of the study design reported in each article are analyzed. It is worth noting that since the data reported here mostly consist of processing techniques or other non-exclusive characteristics, the same study may report two or more different aspects. Therefore, the total

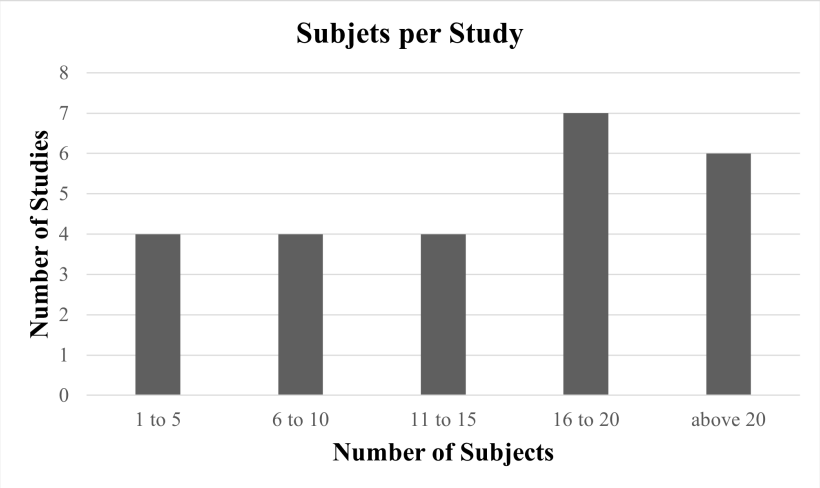


Figure 2.4: Bar Chart of Number of subjects by Study

number of studies presented in the graphs and other shared information may exceed 32.

Biometric Parameters

The included studies aimed to estimate five biometric parameters: Heart Rate (HR), Respiration Rate (RR), Blood Pressure (BP), Heart Rate Variability (HRV), and Temperature. Out of the 32 screened studies, 31 (96.9%) estimated HR, 7 (21.9%) estimated RR [17, 18, 26, 27, 31, 43, 51], 2 (6.25%) estimated BP [19, 35], and only 1 (3.13%) investigated HRV [42] or Temperature [31], as can be seen in Figure 2.5. Apart from the HR, the only other parameter to be estimated exclusively in a study was the BP [19]. All others were estimated along with the HR.

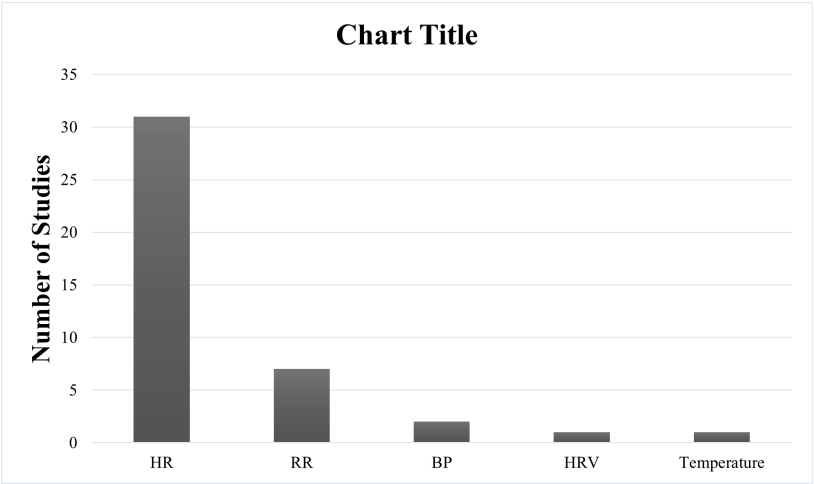


Figure 2.5: Bar Chart of Studies per Estimated Biometric Parameter

Magnification Type

While the majority of studies strictly employed color magnification (27, 84.4%), only 1 (3.13%) exclusively tested motion magnification [52]. Another 4 (12.1%) studies simultaneously tested both approaches [17, 18, 27, 43]. The vast majority of estimations for almost all parameters were thus carried out with color magnification, with motion magnification being primarily utilized only for RR estimation.

Furthermore, the estimation of Temperature was not performed along with EVM, thus, losing relevance for this review [31].

Study Conditions

Through the analysis of the included studies, two main causes of image artifacts were identified: lighting issues and motion issues. The studies were thus divided into 4 categories based on the addressed artifacts to understand the types of conditions imposed during their execution. Firstly, those that presented completely controlled conditions to test their algorithm where participants had to avoid movement and were exposed to constant lighting. These comprised 20 out of 32 articles (62.5%). Secondly, those that, although not studying the effect of artifacts, did not impose conditions regarding subject movement and/or did not control lighting conditions, posing greater challenges to the algorithm. This was carried out by 5 articles (15.6%). The other two types consist of studying either the effect of motion or lighting, producing image artifacts by pushing these conditions to the extreme to test their algorithms. Those reporting to study motion artifacts were 5 (12.5%) and image artifacts were 9 (28.1%). The information is further summarized in Table 2.4.

Table 2.4: Addressed Artifacts

Artifact	Number of Studies	Studies
None - Controlled Conditions	20	[17–19, 27, 29–32, 34–38, 40, 42, 45, 46, 49, 52, 53]
None - Natural Conditions	5	[28, 44, 51, 54, 55]
Motion	5	[26, 27, 41, 43, 50]
Illumination	9	[26, 33, 34, 36, 37, 41, 43, 47, 48]

Other conditions were tested throughout the studies. The most relevant, along with their respective articles, were as follows: 3 (9.38%) studied the impact of different Regions of Interest [32, 37, 49], 2 (6.25%) tested different recording distances [43, 53], and 5 (15.6%) tried different angles/poses during recording [33, 38, 49, 51, 53].

Regions of Interest

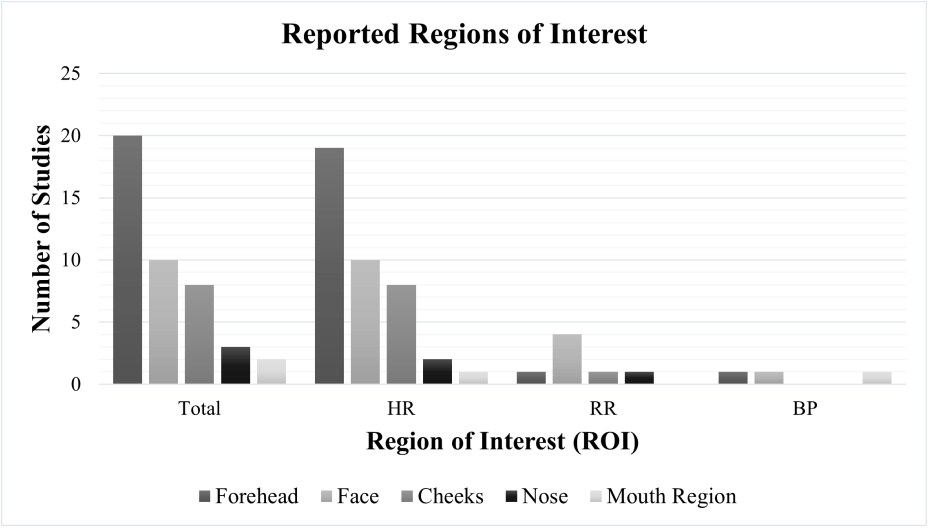


Figure 2.6: Bar Chart of Studies per ROI

The facial ROIs used by the studies were as marked in Figure 2.6. Several studies utilized more than one of the listed facial regions to determine their respective parameter (12, 37.5%), while the remaining limited their investigation to just one of them (20, 62.5%). Overall, the most utilized region was the forehead (20), followed by the entire face (10), cheeks (8), nose (3), and mouth region (2). The trend was similar for HR estimation; however, for RR, the entire face was the preferred region of interest. It is not possible to determine a trend in blood pressure estimation due to the limited number of studies in this parameter. The general information for all parameters was also summarized in Table 2.5.

Table 2.5: Reported Regions of Interest

ROIs	Number of Studies	Studies
Forehead	20	[17–19, 28–32, 34, 36, 37, 40–43, 47–49, 52, 55]
Face	10	[18, 26, 27, 32, 33, 35, 38, 45, 51, 54]
Cheeks	8	[17, 34, 36, 37, 42–44, 50]
Nose	3	[31, 44, 53]
Mouth Region	2	[19, 46]

To track/extract the regions of interest, various different algorithms were employed. Out of the 32 articles, only 26 (81.3%) reported any information in this regard. Two of these reported not using any algorithm, instead asking participants to control their movement as best as possible. The remaining 24 reported using one or more algorithms for this purpose, but only 22 shared the employed algorithms. Among these, 11 used the Viola & Jones algorithm [60] or a similar Haar Cascade-like approach for face detection, 6 used the Kanade-Lucas-Tomasi algorithm [61] for face tracking, 2 used Mediapipe Face Mesh [62] for feature extraction, and 2 reported using a self-developed algorithm for skin segmentation (Table 2.6). Additionally, 10 others reported using different algorithms.

Table 2.6: ROI Tracking/Extraction

Algorithm	Number of Studies	Studies
Viola & Jones/Haar Cascade	11	[18, 31, 34, 36, 37, 40, 42, 45, 51, 52, 55]
Kanade-Lucas-Tomasi	6	[18, 31, 36, 37, 47, 51]
Mediapipe Face Mesh	2	[17, 46]
Skin Segmentation	2	[51, 54]
Other	10	[26, 27, 32, 40, 41, 43, 44, 46, 53, 54]
None	2	[49, 50]

Camera Settings

While performing non-contact estimation, a camera is necessary for video acquisition of the ROI. Therefore, camera specifications such as video resolution, frame rate, shooting distance, etc; can impact the accuracy of parameter estimation.

The first characteristic addressed was the spectrum/color space used during image acquisition (color spaces used during image processing were included in this part of the review). Out of the total articles, 26 studies reported or allowed inference about the spectra and/or color spaces used throughout the procedure. As expected, the most used spectrum was the visible light spectrum, which provides more information about individuals' pulses. Within this spectrum, the most utilized color space was RGB, by

21 out of the 26 studies. Other color spaces used included HSV, YIQ, and YCbCr. Out of the 26 studies, 3 captured images in the Near-Infrared (NIR) or Infrared (IR) spectrum. There was also a study that captured depth images using a multimodal camera [33]. This information is organized in Table 2.7.

Table 2.7: Reported Color Space/Spectrum

Color Space/Spectrum	Number of Studies	Studies
RGB	21	[17, 18, 26–29, 31, 33, 36–38, 40–42, 44, 46–49, 51, 53]
YIQ	5	[34, 36, 37, 43, 50]
HSV	3	[17, 31, 38]
NIR/IR	3	[31–33]
YCbCr	2	[27, 52]
Depth	1	[33]

Subsequently, the camera resolution and frames per second (fps) of the recording were analyzed. In terms of image resolution, the most frequently used resolution was 1920×1080, in 8 out of the 24 articles that provided this information, followed by resolutions of 1080×720 and 1280×720, each with 4 articles. The shared resolutions ranged from 80×60 [31] (for capture in the IR/NIR spectrum, which typically utilizes reduced resolutions) to 3840×2160 [51], varying greatly in shape and size. The more frequent image resolutions are relatively high when compared to the shared spectrum, which aligns with expectations due to their ability to convey more information.

Regarding the fps during image capture, only 24 articles reported any information. The most common frame rate used was 30fps, with 14 articles. Another 7 reported frame rates lower than 30fps. Lastly, 5 studies reported frame rates equal to or greater than 60fps. Only one article tested the impact of different frame rates (between 30 and 60fps) [34]. While maximizing these two parameters can provide a significant amount of information, they may also increase algorithm processing time, which could be an undesirable effect. The studies that reported information about their recording frame rates are organized in Table 2.8.

Table 2.8: Reported Frame Rates

Frame Rate	Number of Studies	Studies
Lower than 30	7	[17, 32, 41, 44, 50, 53, 55]
30	14	[18, 28, 29, 34, 36, 37, 40, 42, 47–49, 51, 52, 54]
60 or higher	5	[18, 19, 26, 34, 43]

The reported recording time was analyzed as it can also impact the results. Performing prediction within a shorter recording interval may have a negative outcome. Out of the 32 studies, only 24 (75%) reported the recording times used. Among these, only two tested the effect of different recording times: [53] tested various values between 6 and 60 seconds; and [34] tested 15 and 30 seconds of recording. Additionally, two studies used different recording times, although they did not consider their effect [18, 44]. The articles were subdivided into different ranges, as shown in Table 2.9, except for [53], which was compatible with various subdivisions. The shortest recording time analyzed was 6 seconds [53], and the longest recording time conducted was 5 minutes [32].

Lastly, the recording distance was analyzed. A total of 21 (65.6%) articles reported the distance from the subject to the camera during recording. Among these, only 1 study actively tested the influence

Table 2.9: Reported Recording Duration

Recording Duration	Number of Studies	Studies
30s or less	6	[18, 27, 31, 33, 34, 40]
>30s to 60s	9	[17, 19, 26, 36, 37, 44, 47, 54, 55]
More than 60s	8	[29, 32, 41, 42, 48–51]

of different recording distances [53]. Additionally, 2 studies reported various distances, but not aiming to test their effect [18, 33]. The remaining 18 were subdivided into different ranges, as shown in Table 2.10. The vast majority of studies used distances less than 1 meter. One article stood out due to testing distances greater than 50 meters [43]. The smallest distance reported was 20 centimeters [33].

Table 2.10: Reported Recording Distance

Recording Distance	Number of Studies	Studies
1m or less	10	[19, 28, 31, 36, 37, 40, 41, 45, 47, 52]
>1m to 2m	2	[32, 51]
More than 2m	6	[26, 27, 29, 43, 48, 49]

Signal Extraction

The first step in executing the prediction of a given parameter, after magnification is performed, is to extract the biomedical signal through which the prediction would be made. Out of the 32 articles, 3 (9.38%) did not share information on how the signal extraction was conducted [19, 45, 55], and another 2 (6.25%) did not need to perform extraction as they used a method solely based on a Convolutional Neural Network (CNN) [44, 46]. The remaining 27 (84.4%) shared their approaches for conducting this extraction. The most frequently extracted signal is the Blood Volume Pulse (BVP) given its importance for the determination of HR and BP.

The most utilized method, in 19 out of the 27 articles, involved separating a color channel (or vertical motion in the case of motion magnification) from the pixels of the region of interest and using the intensity of this channel as the biomedical signal. Separation was predominantly performed on the green channel (11) for the RGB color space, followed by the Y channel (4) for the YIQ color space. Detailed information has been organized in Table 2.11.

The second most utilized method, by 8 out of the 27 articles, is Blind Source Separation (BSS), where the given signal is assumed to be a mixed one and through separation it is divided into the different source signals. This was achieved through methods based on component analysis such as Principle Component Analysis (PCA) [63], Independent Component Analysis (ICA) [64], and Canonical Component Analysis (CCA) [65], or through methods based on intrinsic mode functions such as Ensemble Empirical Mode Decomposition (EEMD) [66], Adaptive Chirp Mode Decomposition (ACMD) [67], and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) [68]. The most common employed one was ICA, reported by 4 different articles. Further information is provided in Table 2.12.

Out of the 27 articles, 6 of them also used an approach of the average intensity of the pixels in the region of interest, without employing any of the algorithms mentioned here to conduct the separation [31–33, 50, 52, 54].

Table 2.11: Reported Channel Separation

Channel	Number of Studies	Studies
R	3	[17, 33, 38]
G	11	[26–28, 30, 38, 41, 42, 47–49, 51]
B	1	[38]
H	2	[31, 38]
S	1	[38]
V	2	[17, 38]
Y	4	[34, 36, 37, 43]
I	1	[34]
Vertical Component	1	[27]
Not Specified	1	[29]

Table 2.12: Reported Blind Source Separation Methods

BSS Method	Number of Studies	Studies
PCA	1	[35]
ICA	4	[18, 40, 43, 53]
CCA	2	[26, 27]
EEMD	1	[18]
ACMD	1	[41]
CEEMDAN	2	[26, 27]

Estimation

The next step, following the extraction of the biomedical signal, is to perform the estimation of the biometric parameter in question. This was employed by the studies as described in Table 2.13. The most common reported procedures involve using either the Fast Fourier Transform (FFT) to determine the predominant frequency in the extracted signal or by simply employing a peak detection algorithm to count the relative maxima in the extracted signal. All studies provided information on this aspect except one [35]. Additionally, only one study tested two different approaches to carry out the estimation of a single parameter, the FFT and Zero Crossing [32]. There was also a study that stood out for using a mathematical model to determine BP through the BVP [19].

Table 2.13: Reported Estimation Methods

Estimation Methods	Number of Studies	Studies
FFT	16	[29, 31–33, 35–37, 40, 42, 47–53]
Peak Detection Algorithm	12	[17, 18, 26–28, 30, 34, 38, 41, 43, 54, 55]
CNN	3	[35, 44, 46]
Zero Crossing	1	[32]
Mathematical Model	1	[19]
Not Specified	1	[45]

Reference Devices

Only the devices used for ground truth in articles that developed their own dataset were considered, as the reference devices used for publicly available datasets can be easily accessed. Excluding the 3 articles that used such datasets, 28 articles reported the device used to compare their results. Only 1 study did not share any information on this aspect [55]. The vast majority of studies reported different devices, with no consensus in this regard. To facilitate this analysis, these devices were generalized by device type. The most common type of device for determining HR was the pulse oximeter, which was reported by 15 studies, followed by Electrocardiogram (ECG) measurement systems with 5 articles. For RR, the most common device used was respiration belts (3 articles), and for BP, blood pressure monitors (2 articles). Other mentioned reference devices to measure the biometric parameters include Holter devices, HR sensor bands, amongst others.

2.3.4 Performance Metrics

The performance analysis of the studies primarily focused on three different metrics: MAE, MAPE, and RMSE. At least one of these was reported or calculated based on information available in the articles. Other metrics were reported by some of the articles, such as Pearson correlation score, standard deviation, Bland-Altman analysis, among others. However, due to the reduced frequency with which these were reported, they were not considered as they hinder comparison among various studies.

The most frequently reported metric overall was MAPE, which was reported by 16 out of 32 articles (50%) and calculated in 6 of them (18.8%), totaling 22 articles. The second most reported metric was MAE, also reported by 16 out of 32 articles (50%), but could only be inferred in another 5 (15.6%), totaling 21 articles. Lastly, RMSE was reported by 14 articles (43.8%) and could also be calculated in 5 articles (15.6%) for a total of 19 papers. The articles that reported each metric are organized in Table 2.14.

Table 2.14: Reported Performance Metrics

Error Metric	Number of Studies	Studies
MAE	21	[17–19, 28–32, 35, 41–44, 46, 47, 50–55]
MAPE	22	[17, 19, 26, 28–38, 41, 44–46, 48–50, 52]
RMSE	19	[17, 18, 26–32, 35, 40, 41, 43, 44, 46, 50, 51, 53, 55]

The remaining analysis in this section will be conducted separately for each biometric parameter: HR, RR, and BP. The other parameters were discarded as they were identified in only one study each, which does not allow for a performance comparison. It is important to mention that, in the case of articles that reported values, the metrics extracted from each represent the performance of the most successful approach within each article. That is, the retrieved values are a sample of the best performances of each article. The details of the reported metrics for all the parameters are available in Appendix A (Tables A.1 and A.2 for RR and HR, respectively).

Heart Rate Studies

There were 31 studies that attempted to estimate HR, of which 20 (64.5%) reported MAE, 20 (64.5%) reported MAPE, and 18 (58.1%) reported RMSE. The Mean and Standard Deviation (SD) of the reported metrics have been organized in Table 2.15. The distribution of these metrics has also been

organized in a box-whisker plot (Figure 2.7) for a more detailed analysis. One study was not included in this analysis due to the unavailability of the exact MAPE value (between 2% and 3%), which was the only reported metric [34].

Table 2.15: Performance Metrics Statistics (HR)

Error Metric	Number of Studies	Mean ± SD
MAE	20	2.84 ± 2.64 (bpm)
MAPE	20	5.90 ± 3.92 (%)
RMSE	18	4.40 ± 3.82 (bpm)

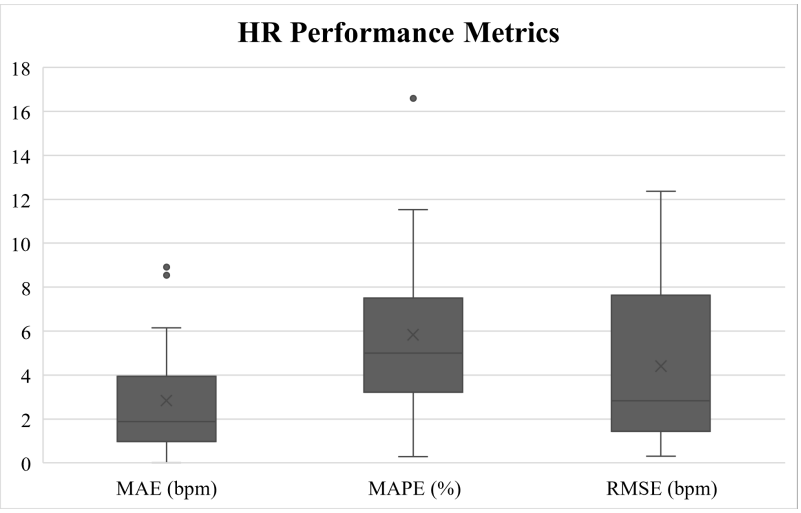


Figure 2.7: Box-Whisker Plot of HR Performance Metrics

Among the 20 studies that reported MAE, 5 (1^{st} Quartile) managed to achieve a value below 1 bpm [18, 35, 42, 50, 53]. The minimum reported value was 0.006 bpm [42]. Analyzing the box-whisker plot, it is possible to identify two outliers regarding MAE. The first outlier tested an extreme lighting situation, where the subject was illuminated with a light source of about 60 lux [47]. The second outlier is an article that used a simple linear EVM algorithm [28]. Additionally, the measurements in the second outlier were taken under natural and ambient lighting conditions, which in certain cases may not be sufficient to obtain good results.

Among the 20 studies that reported MAPE, 6 (1^{st} Quartile) managed to achieve a value below 3% [17, 26, 34, 35, 37, 45]. The minimum reported value was 0.029% [26]. Analyzing the box-whisker plot, one outlier regarding MAPE can be identified. This article overall tested very low illumination levels, leading to a lower performance of the algorithm [33].

Lastly, among the 18 studies that reported RMSE, 4 (22.2%) managed to achieve a value below 1 bpm [18, 26, 27, 53]. The minimum reported value was 0.31 [26]. This metric did not present any outliers.

Respiration Rate Studies

Regarding Respiration Rate, there were only 7 articles trying to approach the estimation of said parameter. Of the 7 included articles, 5 (71.4%) reported MAE, 3 (42.9%) reported MAPE, and all 7 (100%) reported the RMSE. The Mean and SD of the reported metrics are presented in Table 2.16. A

box-whisker plot was not developed for the RR parameter due to the scarcity of results resulting from the limited number of articles.

Table 2.16: Performance Metrics Statistics (RR)

Error Metric	Number of Studies	Mean ± SD
MAE	5	1.31 ± 0.33 (bpm)
MAPE	3	2.03 ± 2.63 (%)
RMSE	7	1.22 ± 0.68 (bpm)

From the 5 studies that reported MAE, 2 (40%) achieved values below 1 bpm [31, 43], but all of the included studies managed to develop approaches under with an MAE value below 2 bpm. The minimum reported amount was 0.91 [31].

Only 3 articles reported MAPE. Of those, 2 accomplished very low values: 0.158% [17] and 0.18% [26]. Comparatively, the last value was much higher than the previous ones, 5.74% [31].

Lastly, among the 7 studies that reported RMSE, 2 (28.6%) managed to achieve a value below 1 bpm [26, 27]. The minimum reported value was 0.31 [26]. Only one value reported surpassed an RMSE of 2 bpm [17].

Blood Pressure Studies

Only two studies investigating the estimation of BP were included, rendering it impractical to conduct a comparative analysis. The outcomes reported from each study, for systolic and diastolic BP, are delineated in Tables 2.17 and 2.18, respectively. Notably, one of these studies demonstrated a favorable performance, suggesting potential applicability of the EVM approach to methods intended for predicting this parameter. However, further research is imperative to facilitate a more comprehensive analysis.

Table 2.17: Reported Performance Metrics (systolic BP)

Study	MAE (mmHg)	MAPE (%)	RMSE (mmHg)
[35]	0.1	3.73	4.11
[19]	9.4	8.2	-

Table 2.18: Reported Performance Metrics (diastolic BP)

Study	MAE (mmHg)	MAPE (%)	RMSE (mmHg)
[35]	0.4	4.95	3.75
[19]	8	9.8	-

2.4 Discussion

In this section, the proposed research questions will be addressed one by one, aiming to comprehend the state of the art. By addressing these questions, the key points to understand the state of the art will be covered. They will be answered based on the analysis of the data collected throughout the preceding section.

2.4.1 (RQ1) What biometric features can be estimated through facial videos?

One of the main objectives of this review was to understand which biometric data could be extracted from magnified videos. Five biometric parameters to be estimated were identified: Heart Rate (HR), Respiration Rate (RR), Blood Pressure (BP), Heart Rate Variability (HRV), and Temperature. Considering the properties of this magnification technique, it would be expected that features based on color change or motion augmentation could be extracted. This is the case with RR, which can be easily measured with the subject's chest movement, for example; or with HR, which causes a change in the subject's skin color, this change being magnified by the technique used, or by magnifying the slight movement caused by the heartbeat on smaller blood vessels. HR was addressed more frequently than the others, solidifying the possibility of estimating this parameter. Even though the retrieval of numerous articles attempting to estimate this parameter may result from some bias due to the keywords used for the search, its estimation is vastly researched in the available literature given the HR monitoring importance. As expected, this parameter could be estimated with excellent results by several articles. RR was the second most addressed parameter, also estimated with good results. Contrary to expectations, this parameter could also be determined through color magnification by some articles [18, 26, 51], not just through motion magnification. These articles used a methodology very similar to that of HR. The estimation of the remaining parameters may not be as obvious. Blood pressure can be determined through mathematical models based on the extracted BVP (the same signal extracted to estimate the HR, [19]) or by using black box models like CNNs [35]. Although its estimation is evidently possible, few studies have been found on this subject, making it difficult to study this parameter with the same depth as the previous ones. More research in this area is needed. HRV was assessed by only one study [42]. Its estimation is analogous to that of HR, although it was hardly addressed in the included studies. Based on that, it is expected that its estimation is equally accessible but further research is needed to draw any conclusions. Lastly, Temperature estimation was also addressed by one study [31]. Its estimation was not performed with video magnification and was based solely on infrared spectrum image analysis. Although not highly relevant to this review, this aspect is noteworthy as an advantage of this imaging modality, which allows for the analysis of an additional parameter at the expense of some image resolution.

2.4.2 (RQ2) How do the different methods affect the model's performance?

In this question, only HR and RR will be addressed due to the scarcity of articles on the other parameters. Each part of the methodology will be analyzed separately, assuming that they are independent from each other and contribute individually to the results. This assumption may introduce some bias into the analysis, but it allows for the deconstruction of the procedure used by each article to make the analysis more accessible.

Heart Rate

The most commonly used magnification method for this parameter was definitely color magnification (31 out of 32). This type of approach is well established in the literature, as only 2 studies attempted to estimate HR through motion amplification [18, 52]. The use of motion magnification is widespread in methods that utilize different regions of interest, such as the wrist or neck. However, a disadvantage of using this type of magnification on the face is the presence of movements with greater amplitude, which may hinder the extraction of this type of information. Furthermore, employing this type of magnification did not show any trend of improvement in the results. Two studies tested the impact of using two different

types of magnification, finding no significant differences between them [26, 43].

The most commonly used ROI was the forehead (19 out of 32 articles), but there did not seem to be a trend of improved results compared to some of the other regions used, since the cheeks and nose produced similar results. All of these are highly vascularized areas and are expected to produce a good response to color magnification. The most reported reason for using the forehead was its flat and larger area, and unlike other regions, it is not subject to as much noise caused by movement. The use of the entire face was also proposed. Although it allows for more information retrieval, it is also subject to more noise due to including regions without relevant information. An example is the mouth region, which was used in a study with less success [46]. Certain studies that used the face as the ROI also took advantage of a skin segmentation algorithm that avoids regions without interest for the problem [51, 54]. One study attempted to test different ROIs and combinations of them (forehead, cheeks, and neck) and concluded that using all simultaneously produced the best results [37]. The regions tested were limited, and the study focused heavily on the neck region, making it difficult to draw conclusions from its results. One other tested between the forehead, upper half of the face and the whole facial region, resulting in similar metrics between the three even though the whole face region had slightly better results. There was also a study that attempted to test the impact of the size of the ROIs used; however, among the sizes tested, no significant difference was identified [36]. Another study also tested different sizes of ROIs and concluded that there is a threshold size that provides better results [49]; that is, regions smaller than 64×64 pixels showed worse performance due to the scarcity of information that could be acquired through such a small region. Larger regions did not show any significant improvement.

The majority of articles acknowledged the need to mitigate noise caused by subject movement, with a trend towards using algorithms to track the ROI. The most commonly used algorithm was Viola & Jones due to its low computational requirements and processing speed, enabling fast tracking of the ROI. However, this feature, like others, represents a trade-off; while more effective methods exist, they are generally heavier and slower in regards to computation, which could be disadvantageous depending on the application. There did not appear to be any specific trend regarding results and proposed solutions. Rather, the most important aspect was addressing the challenges posed by motion noise. The only two articles that reported not using these types of algorithms needed to find ways to counteract movement noise. One minimized movement as much as possible by controlling test conditions [49], while the other developed a magnification approach combined with Discrete Wavelet Transform (DWT) and PCA to reduce noise, with the latter yielding excellent results in terms of MAE [50]. Scenarios where the issue of motion noise was taken to the extreme were also tested by certain articles. Most of them developed robust responses to this type of noise, based on more complex systems with better tracking of the ROI and respective features [26, 27, 41, 43, 50], and/or on stronger preprocessing techniques and BSS methods that can eliminate noise more efficiently [26, 27, 41, 50]. Some of these are among the most successful in terms of the results retrieved [26, 27, 50].

To generate the BVP signal used to estimate parameters such as HR, HRV, and BP, the preferred method was to use the intensity of one of the color channels, particularly the green channel of the RGB color space. One article studied which color channels yielded the best results, effectively concluding that the green channel of said color space was the most successful [38]. The use of this approach (selecting one channel) did not seem to impact the results significantly, as its application was evenly distributed among all articles, both those that achieved relatively better performances and those with poorer results. In contrast, the use of BSS methods (alone or in conjunction with color channel separation) appeared to

skew towards better results, with 5 out of 8 articles reporting the use of these methods among those that achieved relatively better outcomes (within the 1^{st} quartile) [18, 26, 27, 35, 53]. This type of approach was the second most reported, with ICA being the most commonly used technique within BSS. One article investigated the impact of different BSS methods on the results, among which ACMD yielded the best outcomes. Additionally, it proposed a signal quality assessment method to create a more robust methodology [41]. Another reported approach of BVP extraction, although less frequently, was simply using the average pixel intensity of the region of interest using a linear combination of all channels, which, like using the intensity of one of the color channels, did not seem to impact the results significantly, yielding outcomes in line with the average performance. Two studies did not need to perform the signal extraction as they employed CNN-based approaches, where the input corresponds to images obtained after magnification [44, 46]. Although they did not exhibit standout performances, it is a rapidly growing and promising technology that deserves further research.

In addition to signal extraction, possibly the most important factor is the estimation of the HR value. This was mainly carried out in one of two ways: frequency spectrum analysis, determining the predominant frequency, or employing a peak counting algorithm on the extracted signal to estimate the HR value. In this regard, there also did not seem to be a trend in the results between the two presented approaches.

One of the problems with the estimation observed by some of the articles was the need for prior knowledge of the HR frequency to determine which frequency range to magnify, in order to reveal the color changes related to it without magnifying a lot of noise. Two articles attempted to overcome this requirement. The first one tested magnification and estimation across various frequency intervals ranging from 35 to 110 bpm, assuming as real the HR frequency obtained in the highest number of intervals [33]. Another proposed approach was to perform an initial estimation with a broader amplitude covering the most common spectrum of Heart Rates and use this estimation to perform a more specific magnification with a narrower filter based on a confidence interval to produce an appropriate response [54].

Respiration Rate

Regarding the type of magnification used, there is greater variability for this parameter than for HR. Despite this, the article that achieved the best results used color magnification [26]. Furthermore, two studies tested the impact of both approaches, concluding that color magnification slightly improved the results [27, 43].

The entire face was undoubtedly the preferred region for determining this parameter. It is difficult to identify a trend in the results due to the lack of articles on this biometric parameter. Furthermore, no study directly tested the impact of different ROIs on the estimation of this parameter. Regarding the tracking of these regions, the most important aspects have already been addressed in the section on HR, particularly in articles [26, 27, 43]. This issue is analogous for both biometric parameters, therefore what was mentioned can be generalized onto the estimation of the RR.

The most commonly used signal extraction for motion magnification was performed by analyzing the vertical component of motion. One study used the distance between the nose and the shoulder line to determine this component [17]. On the other hand, in the case of color magnification, it was mostly carried out similarly to HR estimation, namely using BSS methods, which were used by 4 out of 7 articles that attempted to estimate this parameter (3 of which were the most successful) [18, 26, 27, 43]. Regarding

the estimation, for this parameter it was mostly performed by leveraging peak counting algorithms for the biomedical signal (5 out of 7 articles).

The limited number of studies on RR in this review complicates the analysis of which approach might be best for its estimation. Further research is needed on this parameter to draw conclusions about the best approach for its overall framework.

2.4.3 (RQ3) What are the factors that can induce error in the estimation?

The factors most addressed throughout the articles were lighting and motion artifacts. Motion magnification introduces unwanted noise and generally complicates the estimation process. On the other hand, insufficient lighting in the images makes it challenging to extract information as the information becomes less abundant, resulting in a low signal-to-noise ratio after magnification. To address motion-related issues, as mentioned in the previous research question, methods for tracking the ROI and robust data preprocessing approaches were employed, leveraging BSS techniques to remove noise. The latter was also beneficial for addressing lighting issues and noise removal in general. The lighting problem was also addressed by utilizing other spectra/color spaces. In particular, the infrared spectrum resolves lighting issues, although it may be affected by humidity and temperature and have lower resolution. However, studies [31–33], which employed infrared, did not yield significant results, possibly due to the lower resolution of the images and frame rates. In addition, the use of this type of camera reduces accessibility. Another approach involved using different color spaces such as YIQ, HSV, and YCbCr, all of which include a component related to color brightness (Y for YIQ and YCbCr, and V for HSV), aiding in recovering more information from the image. However, study [17] determined that despite being slightly superior, there was no significant difference between HSV and RGB. Additionally, study [31] employed HSV but yielded relatively poor results. A different approach was attempted by a study to counteract the effect of lower illumination by leveraging the low-light image enhancement function of the camera response model, albeit without achieving satisfactory results relative to this approach [47].

Nine studies pushed the boundaries of lighting conditions to extremes and attempted to devise more robust solutions to this problem, with varying degrees of success. Five out of the nine studies that tested extreme lighting situations utilized the YIQ color space to mitigate this effect [34, 36, 37, 43, 50] with more success, of which only 2 used it solely during preprocessing before estimation [36, 37]. Additionally, study [34] concluded that when performing channel separation to extract the BVP, the Y channel outperformed isolating the I channel.

The reality remains that these approaches continue to be tested in highly controlled situations, and many other factors of this nature introducing unpredictability or, more aptly put, naturalness in recording, may ultimately influence the results. Other mentioned features that were tested across the included studies were, for example, different Regions of Interest, various angles and poses, and differing distances. For these, no new solutions distinct from those already mentioned were identified, and many studies merely focused on assessing their impact. In the case of distance, minor variations did not seem to have a significant impact. One study tested the effects of several distances ranging from 50 to 250 centimeters, finding no major differences in the results for the various distances tested [53]. Additionally, there was a study that examined distances exceeding 50 meters but achieved above-average results by applying the various aforementioned solutions [43]. Another mentioned factor was the presence of obstructions such as glasses, hair, etc., which can obscure the region of interest. The solution found by two studies was to perform skin segmentation [51, 54].

Another important consideration pertains to the characteristics of the camera and the recording itself, including resolution, fps, and recording time, which essentially define the amount of information conveyed in the utilized images. The primary issue with these three characteristics is that while a greater quantity of information allows for better prediction of biometric parameters, it also exponentially increases the processing time required to achieve such predictions, presenting a trade-off that may be perceived as a disadvantage.

Regarding camera resolution, the trend has been to utilize devices with higher resolutions. Nowadays, devices with higher resolution are more accessible, enabling the use of images with more information. Despite this, several studies have found success with devices of lower resolution, indicating that resolution is not necessarily a barrier to obtaining good results, particularly because most studies end up calculating the average pixel intensity in the region of interest. However, it cannot be definitively stated that this does not influence the results, as no identified study has examined this impact, and it is expected that a minimum resolution is required to achieve satisfactory results, given that low resolutions of infrared images have not been successful [31–33]. As for frames per second, it did not appear to be a limiting factor, as several studies found success with frame rates both below 30 fps and up to 60 fps. Furthermore, contrary to expectations, one study concluded that 30 fps provided a better response than 60 fps [34]. The most limiting characteristic was undoubtedly the recording time, which can indeed be a limiting factor. The same study [34] tested recording intervals between 15 and 30 seconds and favored the use of the longer interval. Another study tested intervals ranging from 6 to 60 seconds, concluding that less than 12 seconds resulted in significantly higher errors, and that the error stabilized around 30 seconds [53]. One final study attempted to predict instantaneous HR using 4-second recording windows and achieved optimistic MAE values but with high inconsistency considering the higher values of SD, RMSE, MAPE.

It is also worth noting that each study used its own ground truth device with few using the same. Since each one has its own limitations and associated errors, this factor may limit the comparative analysis of the articles if it is done solely based in their reported results.

2.5 Conclusion

This review aimed to consolidate knowledge on approaches for extracting biometric characteristics from videos leveraging the EVM methodology. It extensively addressed typical framework features and aimed to discern essential points to address and those of lesser importance for the obtained results. Despite researching various parameters, only HR could be thoroughly analyzed due to the scarcity of articles on other biometric data. Further research is needed in this area, which is undergoing significant and constant growth, promising the development of new applications that could enhance people's daily lives, such as vital parameter monitoring.

Several limitations exist with the technique used for this review, as enumerated throughout the text. It is noteworthy to mention some, such as the search using keywords heavily focused on HR and the search for facial EVM applications, which, representing specific criteria, may have excluded other important articles and limited findings. Also, the reviewed process is that of a rPPG, although with the scope of including EVM into the pipeline, which may overlook different approaches of estimation not covered in this review. Furthermore, the discussion was conducted from a perspective where the various phases of the procedure were considered independent, potentially introducing bias. Missing information

in certain features of some articles may have also contributed to this aspect.

Chapter 3

Methodology

This chapter depicts a comprehensive description of the applied tools and methods to achieve the objectives of this study. The primary contributions of this chapter are as follows:

- To detail the process of dataset formation, including a thorough description of the data;
- To explain the procedures for data preprocessing and its preparation for the techniques employed;
- To introduce the techniques responsible for magnification and estimation tasks, along with the rationale for their application.

All the code implementation of the algorithms described in this chapter is available online in this **repository**.

3.1 Dataset

As part of this project, a dataset was produced due to the lack of publicly available datasets online. This shortage potentially arises from the fact that the information contained in such datasets (videos of subjects' faces) typically requires approval from an ethics committee, due to its sensitive nature. The dataset was created with the intention of being made publicly available and, for this purpose, was previously approved itself by the Ethics Committee of the Faculty of Sciences at the University of Lisbon (CEC-FCUL) [69]. The procedures used all aligned with the guidelines proposed by the ethics committee.

3.1.1 Subject Selection

The selection of subjects was made with certain conditions in mind. First, the aim was to achieve a balance between male and female participants to improve the generalizability of the results obtained from this dataset. Second, an objective of this dataset was to include individuals from various ethnic backgrounds and with different skin tones. To achieve this, the Fitzpatrick scale was used [70], which assesses skin tone based on visual observation and the subject's susceptibility to sunburn, a factor that was individually discussed with each participant.

A limitation of this scale is that the assessment of skin tone can be somewhat subjective and may vary depending on the evaluator [70]. To mitigate this, skin tone assessments were conducted by two researchers. Another issue that arose was that, since the study was passively shared through email and

fliers, without intensive recruitment efforts in this regard, ethnic diversity was limited, resulting in a narrower distribution of skin tones than initially desired.

Additionally, efforts were made to maintain a relatively narrow age range, with a minimum age of 18. While this reduces the generalizability of the dataset in terms of age, it facilitates the analysis of other data factors by decreasing variability and increasing homogeneity. Beyond these imposed conditions, there was no target population, as the intention was to avoid being bound by the socio-psychological or physiological characteristics of participants. If the age criterion was met, individuals were eligible to participate in the study. The goal was to obtain a randomized sample to validate the proposed methods.

3.1.2 Retrieved Information

The primary objective was to develop a database for conducting Remote Photoplethysmography (rPPG) and Eulerian Video Magnification (EVM) analysis. With this in mind, each participant was required to record a video of their face. This was accompanied by an Electrocardiogram (ECG) recording, which serves as the ground truth. These two represent the most important part of the dataset. Additionally, for each subject, several characteristics were collected beyond the video footage. Among these characteristics, analytical information such as:

- Age and sex, ensuring the subject falls within the specified age range and that gender balance is maintained;
- Height and weight;
- Skin tone, assessed according to the Fitzpatrick scale.

Moreover, other factors that could influence Heart Rate (HR) or its variability were noted [71], including:

- Presence of any medical condition/disease;
- Regular physical exercise, defined as a minimum of 3 hours of weekly physical activity;
- Recent consumption of alcohol or drugs within a 12-hour window. Coffee and certain medications were also noted due to their potential impact on HR, but considered within a 4-hour window;
- Whether the subject is a smoker.

Furthermore, characteristics that could affect rPPG estimation were also recorded, including:

- Skin obstructions such as sunscreen or makeup;
- Lighting conditions, measured at the beginning and end of the recording. Some videos may have lighting outside the noted range due to limitations in measurement.

Lastly, environmental data was collected, including:

- Temperature and humidity, as these factors may influence HR;
- Date and time of the recording.

3.1.3 Acquisition

Equipment

The equipment used for each measurement was as follows:

- ECG: BITalino with a 3-electrode system and corresponding software (OpenSignals), using disposable Ag/AgCl foam electrodes with semi-liquid gel;
- Video Recording: A standard Microsoft webcam with a 1280x720 resolution and 30 fps, using OBS Studio as recording software;
- Environmental Data: Temperature and humidity were measured with a Sensibo Air Pro;
- Illumination Level: Measured with a YFE Digital Light Meter, model YF-1065.

The remaining characteristics were assessed via participant questionnaires or observation.

Set up

The recordings were conducted in a room with only natural light. Subjects were positioned approximately 2 meters from the window. Recordings took place between 10 a.m. and 4 p.m. to ensure adequate lighting, although on some days, this condition was compromised due to cloudy weather. The camera was positioned roughly 1 meter from the subjects' faces and at approximately the same height. Recordings were done against a white background, with participants seated.

The device used to measure room conditions was placed to the right of the subjects, and illumination measurements were taken twice, at the start and end of the recording, with the light sensor positioned in front of the subjects' faces. An image of the setup can be seen in Figure 3.1.

Figure 3.1: Set up for the acquisition

The electrodes were positioned on the wrists and the left hip. The ECG setup also included an input button that, by generating a pulse in the recorded signal, allowed for synchronization between the video and the biomedical signal. Since the synchronization was done manually, there may be a delay between the recording and the signal, but it is expected to be no more than 0.5 seconds. This button was also used in case participants felt unwell or if something occurred that required them to stop or disrupted the recording, allowing that segment to be later discarded.

Procedure

The acquisition process was systematically carried out as follows:

1. Upon entering the room, subjects were seated to begin a relaxation period, allowing their HR and its variability to decrease;
2. Participants were given an informational leaflet containing important details about the procedure, as well as information regarding their data and how it would be used. They were also provided with a consent form to authorize the use of their data as described. Any additional questions regarding the experiment were addressed at this point. All these steps were performed according to ethical guidelines;
3. A brief questionnaire was then administered to collect analytical information and other physiological factors that could influence HR;
4. After these steps, the electrodes were placed, and the ECG measurement system was prepared for recording;
5. Room conditions (temperature and humidity), as well as the date and time, were recorded. The first illumination measurement was also taken;
6. At this point, the subjects should have had time to relax, and the recording process began;
    - The ECG recording was started first in order to confirm it was functioning correctly;
    - After that, the video recording was initiated simultaneously with the press of the input button on the ECG system for synchronization purposes;
    - The recording lasted for 10 minutes, during which subjects were instructed to remain as still as possible, keeping their faces directed toward the camera.
    - During this time, participants were left alone in the room to help them feel more comfortable and to avoid any distractions during the recording period;
    - After recording was complete, the integrity of the recordings was checked, and a final illumination measurement was taken.
7. Finally, the data were saved for processing.

3.1.4 Data Processing

All data processing for the dataset was carried out using Python with Google Colab.

The data to be processed can be divided into three main groups: ECG signal, facial video recording, and lastly, the information obtained from the questionnaires. Two classes were developed to handle both signal and video processing, and a function was created to apply the processing iteratively to each subject. Conversely, the observational and questionnaire data were organized manually.

Electrocardiogram Signal Processing

The ECG is an exam that detects the heart's electrical activity [12]. It is measured by attaching a set of electrodes to the surface of the subject's body, allowing the electrical potential to be recorded. Measuring the electrical activity results in a signal, which is typically used to monitor HR in a clinical

setting. This activity underlies the heart's muscle contractions and creates a recognizable pattern in the measured signal that repeats itself over time. Its analysis is a standard in medical research and clinical diagnosis. The typical shape of an healthy ECG signal is exemplified on Figure 3.2.

Heart Rate calculation is usually performed by measuring the time between two different and consecutive R-peaks. These peaks can be identified in the QRS complexes of the signal, shown in Figure 3.2. Since the time between consecutive peaks may vary throughout the recording, an average of the differences is taken, which is then used to calculate the average HR over a specific time interval. The measurements in this dataset were conducted following this procedure to align with medical standards [12].

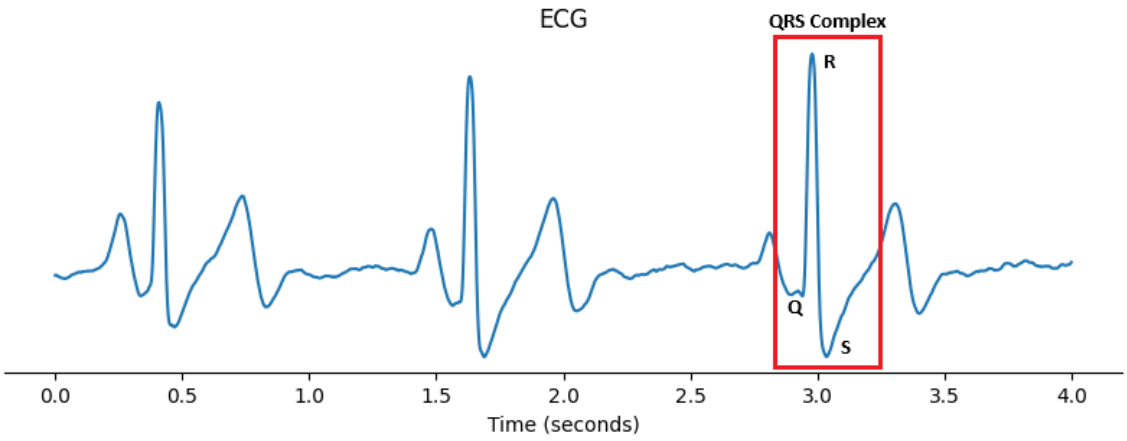


Figure 3.2: ECG signal: QRS complex

$$HR = \frac{60}{RR} \tag{3.1}$$

Equation 3.1 outlines the standard process for calculating the HR, where *HR* represents the mean HR for a given interval in beats per minute (bpm), and *RR* represents the mean of the differences between the consecutive R-peaks in that same interval.

In the processing of the retrieved signals, they were firstly imported and separated into the two different source signals: the ECG signal and the Input signal produced by the button. After that, the synchronization step was done by removing the initial part of the ECG signal before the pulse, which signalizes the start of the facial recording. The remaining signal was also limited to 10 minutes of length. It was then divided into segments. In the project associated with the dataset, it was decided that rPPG estimation would be performed in 20-second intervals, so this was the defined split time. The ECG signal for each 20-second interval is provided in the dataset.

A text file for each subject summarizing the average HR values for each interval is also provided. It was calculated for each split using the ECG signal corresponding to that same interval. The signal was first detrended by subtracting its average value, and then scaled by dividing it by the signal's maximum peak, so that the data falls within a standardized range. A threshold was applied to eliminate low amplitude noise, since only the R-peaks are mandatory to measure the HR. A peak detection algorithm was employed in order to calculate the average difference between consecutive peaks, and using Equation 3.1, the average HR for a given interval was determined. More detailed information can be found in the repository containing the code associated with the dataset.

Video Recording Processing

All the video processing was mainly carried out using the Open Computer Vision Library (OpenCV) for Python. The primary objective during this processing was to reduce the video's dimensions to focus on the ROIs and to segment the video into intervals, similar to the ECG signal processing. It was decided to include only the Regions of Interest in the dataset to maintain the anonymity of the participants, in line with ethical considerations.

For face detection and tracking, Mediapipe Face Mesh was employed. It is a machine learning approach developed by Google that uses a series of models to detect face landmarks and facial expressions in images and videos [72]. One model detects faces and a second model locates landmarks on the detected faces. It shows consistently on the literature to outperform other means of face tracking and therefore was employed to perform this task [73].

The ROIs were defined as the forehead and both cheeks, as these areas are not only the most commonly used for rPPG tasks according to the review but also yield the best results due to being large flat areas with high vascularization, enhancing the Signal-to-Noise Ratio (SNR). Hence, three feature points were identified in the face, corresponding to central points within each one of them. Based on the selected landmarks, bounding boxes were created around them to delimit the Regions of Interest. Subsequently, three separate videos were reconstructed, one for each Region of Interest (ROI). These videos were limited to a 10-minute duration and divided into 20-second intervals, following the same rationale applied to the ECG signal. The division, not only aligns with the estimation time window criteria, but also serves as a data augmentation tool, since the number of subjects was rather low.

The original video had dimensions of 1280×720 and a duration slightly exceeding 10 minutes, while the resulting videos were resized to 64×64 pixels and segmented into 20-second intervals, each one featuring the forehead or one of the cheeks. The rationale behind the 64×64 resolution was simply that it was empirically determined to cover the entire defined ROI without exceeding it.

A common issue with the collected signal and videos is that some intervals could not be properly used due to severe signal or video artifacts, or interruptions in the recording caused by factors related to the subject. Thus, both processing developed algorithms were designed to take timestamps of problematic intervals as input. This parts of the video/signal, that had to be removed, were excluded resorting to those timestamps.

Questionnaire Data

All the data resulting from the questionnaires made to each subject, as well as the observational data, were organized into an Excel file, which is shared with the dataset. A comprehensive list of the data can be found in Subsection 3.1.2. All specifications, units of measurement, and an explanation of the file's organization are also provided alongside the dataset.

3.1.5 Dataset Description

This dataset consists of various facial videos with the purpose to support research in rPPG applications, developing innovative approaches to measure biomedical signals from videos. The dataset consists of 27 subjects (13 male, 14 female) with a mean age of $22.5(\pm 1.2)$, each with up to 90 videos of 20 seconds. The original videos were recorded in 10-minute segments under natural lighting conditions, divided into 30 intervals of 20 seconds each. For each interval, three regions were extracted: Forehead,

left and right Cheek, resulting in a total of 90 videos per subject. Each subject has a corresponding text file containing the mean HR data for each interval of recording. Further details are provided in the file header. Also, the ECG signals associated to each interval are provided. Some of the intervals were not used due to severe image/signal artifacts and there may be a slight desynchronization (no more than half a second) of the signal with the videos due to it being done manually. The videos have a resolution of 64×64 and a frame rate of 30 fps, while the ECG signal has a sample rate of 1000 Hertz. A Microsoft Excel Worksheet accompanies the dataset, containing additional details about each subject.

3.2 Task Breakdown

The methods applied in the study can be grouped into two different tasks: Video Magnification and HR Estimation. The aim of combining these tasks is to develop a more robust rPPG approach, providing increased explainability through EVM, which serves as a visualization algorithm. Although the pipeline is integrated and the tasks interact with one another, their analysis will be conducted separately.

3.2.1 Underlying Theory

As blood circulates through the body, it induces periodic changes in the volume of blood within the skin's capillaries. When a camera captures video footage of the skin, minute variations in skin color can be detected. These variations correspond to the heartbeats, as the skin absorbs light differently depending on the volume of blood flowing through it, which is influenced by the concentration of hemoglobin. Hemoglobin absorbs light in the green wavelengths more effectively than other tissue components. As the volume of blood fluctuates with each heartbeat, the amount of light reflected from the skin's surface changes, particularly in the green channel of the camera due to its strong absorption properties. In addition to color changes, subtle body or facial movements resulting from the heart's pumping action, a phenomenon known as ballistocardiography, can also be detected. These micro-movements are captured by the camera as small shifts in pixel intensities and can be used to estimate the pulse.

By measuring these small changes, both in color and the subtle movements caused by the heartbeat, it is possible to infer the blood volume within the body's micro-vascular system, which is represented by the Blood Volume Pulse (BVP) signal. This signal can be used to calculate physiological parameters such as HR, Heart Rate Variability (HRV), and Respiration Rate (RR). The collection of this signal is known as Photoplethysmography (PPG), or, when done remotely, Remote Photoplethysmography.

Although these physical phenomena are not perceptible to the naked eye, video processing techniques like EVM can amplify these slight temporal variations, making them visible. This technique is based on the same principles as rPPG, and therefore the two are easily grouped together.

3.2.2 Eulerian Video Magnification

The EVM algorithm is a powerful visualization technique that enhances subtle temporal variations in a video sequence, such as changes in skin color due to blood flow or small movements caused by ballistocardiography. The algorithm takes a video as input and returns an amplified version of the same video, making previously imperceptible changes visible.

This algorithm can be classified into two categories depending on the type of changes being amplified: motion magnification and color magnification. Motion magnification highlights tiny movements

such as those caused by ballistocardiography. Motion magnification is older than color magnification, with langragian-style applications dating back to 2005 [74]. The big difference between the lagrangian and eulerian styles is the type of motion tracking. While the former tracks the translational movement of structures in the video, the latter tracks the color changes along the contours of the structures. Thus, color magnification only exists for the eulerian approach. The color magnification focuses on amplifying subtle color changes such as those in skin caused by heartbeats. While both eulerian methods analyze pixel color over time, they differ in focus and processing. Motion magnification targets subtle displacements along facial edges, where small movements are more noticeable. To better capture these movements, it preserves spatial definition by employing Laplacian pyramids, rather than the simpler Gaussian pyramids used in color magnification for spatial decomposition [75].

In this study, the focus is on skin color magnification, amplifying the slight changes in the skin tone. As such, the methodology described here revolves around color magnification, with the eulerian approach. The specific details of motion magnification, which is more concerned with edge detection and structural movement, can be found in greater depth in the original EVM paper [1].

Algorithm Implementation

The implemented code is based on the open-source code provided with the original paper [1]. The EVM pipeline was adapted to process the videos in alignment with the specifications of the videos in the created dataset, but is prepared to handle other configurations. In the original paper, the first step in the workflow involves converting the video from RGB to the YIQ color space before applying further processing, which is deemed to be advantageous for the signal retrieval. Following this approach, the initial step in the data processing pipeline here was also the conversion to the YIQ color space to maintain consistency with the original methodology.

Then, in order to achieve its task, EVM first applies spatial decomposition to the video, frame by frame, typically using Gaussian pyramids. A Gaussian pyramid consists of multiple layers of images, with the base being the original image at full resolution. Each subsequent layer is created by applying a Gaussian filter (or Gaussian blur) to smooth the image, followed by subsampling the smoothed image by a factor of 2 in both the horizontal and vertical directions. As a result, each higher level in the pyramid has a lower spatial resolution and smaller image size compared to the previous level. The purpose of constructing a Gaussian pyramid is twofold: First, it reduces the spatial resolution, making the algorithm more efficient by allowing it to operate on smaller representations of the image. Second, it performs a local weighed averaging of the pixels intensity, which helps in suppressing noise or minor variations at higher resolutions.

The highest level of the pyramid to be processed can be specified. In this case, it was strategically set to 7, which effectively reduces the image dimensions to 1×1 pixel. Given that the Regions of Interest were already relatively small, there was no need to preserve additional spatial information. Therefore, this reduction to such compact dimensions was deemed beneficial for the estimation task and to reduce the memory usage of the algorithm, improving efficiency. However, the magnification result heavily depends on this parameter and should therefore be tested to achieve the best outcomes. If a lower lever is chosen the magnification pipeline is prepared to further reduce the dimensions of the video in order to fit the input requirements of the estimation algorithm.

Next, the algorithm performs temporal filtering on the highest level image of the pyramid by applying a Fast Fourier Transform (FFT) to each pixel's intensity over time, isolating the frequencies

that correspond to the physiological signals. These signals, often hidden in the original video, are then amplified by a chosen factor, making the small variations significantly more pronounced. The filtering was done between 0.7 and 2.5 Hertz, which corresponds to the range $42 - 150$ bpm, and the signal was amplified by a factor of 75.

At this stage, the video, which now contains only the amplified subtle variations, is resized and converted back from the YIQ color space to RGB. Afterwards, a moving average of the temporal dimension is done in order to eliminate low-amplitude noise. In this case it was done with a window of 10 points. Finally, the algorithm reconstructs the video, combining the amplified signals with the original spatial structure, resulting in a version where previously invisible phenomena, such as the pulse induced color changes in the skin, become clearly visible.

This algorithm not only saves the amplified output video but also provides the Gaussian video in RGB format before it is resized and merged with the original video, i.e. the video which contains only the amplified variations. This Gaussian video is a crucial component for estimating the heartbeat in later steps.

Model Evaluation

The assessment of the magnification model's performance is somewhat subjective, as there is no objective measurement parameter available to quantify its effectiveness. The evaluation in this case focused specifically on its ability to reveal subtle color variations in the regions of interest. With this in mind, the video magnification algorithm was tested empirically on all videos in the created dataset in order to evaluate its overall performance. An analysis of the parameters was also carried out later. These were analyzed separately in order to understand their impact on the final result. Finally, it was tested whether there were differences between the ROIs.

In order to exemplify the results obtained, in some cases frames of the video were shared, corresponding to different phases of the BVP. Additionally, Spatio-Temporal Map (ST Map) was created, helping to visualize the color changes over time throughout the video. To create this ST Map, the horizontal dimension of each frame was averaged, resulting in a single column that preserves all vertical spatial information while sacrificing the horizontal axis. These columns were then stacked horizontally to form an image that displays the temporal variation. This method provides a clear visualization of the color changes over time within the selected region.

3.2.3 Estimation Task

The objective of this task is HR estimation, which can be broken down into two key stages. First, the extraction of the BVP signal, and second, determining the HR from this signal.

In most common rPPG applications, the BVP signal is typically obtained by averaging the pixel intensity within the ROI, producing a noisy 3-channel signal (R, G, and B). To improve the SNR, this raw signal undergoes various signal processing techniques. Common methods include temporal filtering using FFT, detrending, normalization, and advanced approaches such as Independent Component Analysis (ICA). Then, one of the resulting channels is used as the BVP signal. The primary challenge here is to accurately extract the BVP signal while filtering out noise from non-heartbeat sources. The major issue with these approaches is that, despite employing various techniques to mitigate it, they remain susceptible to noise problems related to movement or lighting. Recently, more robust methods for performing remote

rPPG using deep learning have been proposed, offering better resilience to such noise. However, these methods are computationally intensive and inefficient, making them impractical for real-time monitoring.

As for the HR estimation, the processed BVP signal is usually analyzed either by peak counting, detecting peaks corresponding to individual heartbeats and estimating the HR based on their time differences, or by applying FFT to identify the dominant frequency, which directly corresponds to the HR. These techniques, while sufficient, are also vulnerable to the signal's noise and can produce poor results if the signal is not properly pre-processed. Combined with the previously mentioned rPPG methods, this makes the estimation process highly sensitive to even the slightest increase in noise.

Proposed Approach

To overcome some of the limitations of common models, this study proposes a two-phase approach. In the first phase, it combines standard signal processing with the EVM technique, keeping computational demands low. In the second phase, a computationally efficient deep learning model was developed to capture the temporal dynamics of the signal and make predictions.

Additionally, predictions in this study were made using 20-second windows. While most models typically use larger windows, at least 30 seconds, for greater consistency in results, a shorter window can be more beneficial for real-time monitoring. Shorter windows allow for predictions closer to the instantaneous HR, which is crucial in scenarios that require quick updates.

Baseline Models

To evaluate the proposed approach, two baseline models were developed. The first model, Base-rPPG, is an adaptation of the method proposed by Poh et al. [76], where ICA is applied. This model was implemented by following the processing steps and using the parameters outlined in the original paper, with two key modifications: the prediction window was set to 20 seconds, and the moving average window was adjusted to 10, in line with the model proposed in this study.

The process is similar to the one described earlier. After importing the video, spatial averaging of pixel intensity is performed within the ROI. Several signal processing techniques are then applied in sequence, including detrending (based on smoothness priors), Z-Score normalization of each RGB channel (Equation 3.2), and blind source separation using ICA. Then, one of the resulting components of ICA is chosen as the BVP signal and a moving average is applied to smooth the signal. Finally, temporal filtering, using FFT, to isolate the desired frequencies, corresponding to the HR. The resulting BVP signal is then used to estimate the HR through a peak-counting algorithm, which measures the time difference between successive peaks. The HR is calculated using Equation 3.1. Further details on this approach can be found in the original paper.

$$x_{normalized} = \frac{x - mean(x)}{\sigma(x)} \tag{3.2}$$

The second baseline model, Base-EVM, also employs the peak-counting algorithm for HR estimation but retrieves the BVP signal in a slightly different way. In this model, the signal retrieval is done in the same manner as in the proposed one. Specifically, the Gaussian video produced by the EVM algorithm is used. Each RGB channel is normalized using min-max normalization to create data that is better suited for the deep learning algorithm (Equation 3.3), followed by the application of a moving average.

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)} \tag{3.3}$$

By comparing these two baseline models (control), it will be possible to evaluate the results at each stage and determine whether the final outcome was more influenced by the rPPG signal retrieval method combined with EVM, or by the proposed deep learning model for HR estimation.

Algorithm Implementation

The BVP signal for the proposed model was obtained in the same way as in the second baseline model. The Gaussian video generated by the EVM algorithm was used, with the Gaussian pyramid levels set so that it produces a 1×1 image. This output no longer has spatial definition and essentially represents a time series of values with three channels (RGB channels). This allows for the development of a model that works with time series instead of images, reducing the computational load. Another key distinction between this estimation model and more traditional methods is that it takes all three RGB channels as input for HR prediction, whereas most models typically rely on a single channel. Although the HR information is strongest in the green channel, which is commonly used, both the blue and red channels can still contain valuable data, even though the red channel primarily carries residual information.

These signals were then normalized using min-max normalization (Equation 3.3), scaling the values to a range of 0 to 1, making them more suitable as input for the deep learning model. A moving average was applied to further reduce low-amplitude noise, as done in the magnification task. The processed signal was then used as the BVP signal for HR estimation. The deep learning model is designed to learn meaningful data patterns even in the presence of some noise, reducing the reliance on extensive signal processing techniques for recovering the BVP signal.

The next step was data preparation. The recovered signals were paired with the initially calculated HR provided with each one of the videos in the dataset, with the average HR values serving as the target variable for prediction. The data were then split into three groups: training, validation, and testing. Each group has a specific role: the training set is used to train the model, the validation set evaluates model performance during training, and the test set is reserved for assessing the model's overall performance on unseen data. The split was done with a ratio of 60% for training, 20% for validation, and 20% for testing.

The division was done based on subject, which meant each subject must be in only one of the train, validation or test groups. It was implemented this way in order to avoid the introduction of subject bias, as the model could learn subject-specific traits. A subject-based split, where each subject is assigned to only one group, generally leads to better generalization for new subjects compared to a random split of all samples. However, this approach reduces the available training data, limiting the model's ability to capture the full variance of the dataset and resulting in slower convergence during training. Given the limited data available, the split was carefully designed to ensure a similar distribution across the different groups. To achieve this, the mean HR for each subject was assessed, and the subjects were organized into bins based on the quartiles of the distribution of this value. The division aimed to ensure that the resulting groups maintained a consistent distribution of the mean HR values, further supporting the balance between training, validation, and testing sets. This approach helps the model in capturing data variation.

The developed model then requires training on a portion of the dataset to learn the data variance, for which a cycle was prepared. This training phase incorporates both training and validation sets, serving

as an initial assessment of the model's performance. It is particularly useful for hyperparameter tuning, ensuring that by the time the model is evaluated on the test set, it represents the most optimized version possible. The training process is conducted iteratively, allowing for adjustments to the model's parameters and the evaluation of whether these modifications lead to improved performance.

Model Architecture

As mentioned earlier, the HR estimation approach relies on deep learning. The architecture of the model plays a key role in its performance. Several key variables were tested and fine-tuned to optimize the model's architecture. Identifying the most effective configuration, both in terms of layers and variables, required empirical testing, as small changes can greatly influence outcomes. Additionally, other parameters related to the learning process, such as loss functions and optimizer learning rates, were explored. The optimization process was conducted under the assumption that individual parameters contributed independently to performance, as testing all possible interactions would be excessively time-consuming. Multiple iterations of the model were evaluated, with only the best iteration used for final evaluation on unseen subjects in the test group. Detailing every iteration would be impractical and redundant, given the iterative and empirical nature of the process.

The proposed architecture is mainly composed of 1D Convolutional Blocks and a Long Short-Term Memory (LSTM) layer. The combination of these layers is effective at capturing the temporal behavior of the data. 1D Convolutional Blocks, commonly used in CNNs, are well-suited for sequential data, such as time series data, where the input has one spatial dimension. These are particularly useful for detecting local patterns. On the other hand, LSTM layers, a type of Recurrent Neural Network (RNN), are specifically designed to capture long-term dependencies in sequential data, making them particularly effective for handling temporal relationships in time series data. Thus, the developed model is capable of predicting HR based on both short-term and long-term patterns in the data. An ablation study was conducted, revealing that removing either the temporal convolution or the LSTM layer resulted in complete underfitting. This finding underscores the importance of both components in enabling the model to learn and perform effectively.

The final model architecture consists of a temporal CNN followed by an LSTM layer and a fully connected layer. The CNN component is composed of two 1D Convolutional blocks, each performing two consecutive convolutions. The first convolution in each block doubles the number of channels and uses a kernel size based on specific HR frequencies: 12 in the first block (matching the maximum HR of 150 BPM in the dataset) and 24 in the other (approximating the average HR of 75 BPM). This design helps the model capture local temporal features related to intervals between peaks in the BVP signal. The second convolution in each block is a point-wise convolution, which maintains the same number of channels. Each Convolutional block includes batch normalization, the ReLU activation function, and a dropout layer after the two convolutions.

Following the CNN, an LSTM layer is used to capture long-term dependencies in the sequence. After the LSTM, the model includes one fully connected layer with a sigmoid activation function, ensuring the output is within the continuous range of 0 to 1. This is followed by a custom denormalization layer that maps the output values to the range of 40 to 150 bpm, corresponding to the target HR range. The model architecture has been organized in Table 3.1.

The model training process was conducted using the Adam optimizer with a learning rate of 0.02, and the data was fed to the model in batches of 32. The loss function used was the smoothed L1 loss. It

blends L1 and L2 loss, offering robustness to outliers and smooth optimization. The model was trained using backpropagation for weight adjustment. In order to mitigate the risk of overfitting to training data, some regularization techniques were included in the learning process, such as early stopping, which monitors the model's performance on the validation set and halts training when performance starts to degrade, indicating overfitting. Early stopping was implemented with a patience of 30 epochs, meaning training would stop if the model did not improve after 30 consecutive cycles. Both learning process parameters and architecture can be checked in more detail in the repository.

Table 3.1: Model Architecture

Model Architecture		**Channels**
Input	BVP Signal	3
Convolution Block	1D Conv (kernel = 12) Pointwise Conv (kernel = 1) Batch Normalization ReLU Activation Function Dropout	8
Convolution Block	1D Conv (kernel = 24) Pointwise Conv (kernel = 1) Batch Normalization ReLU Activation Function Dropout	16
LSTM	(hidden size = 16)	16
Dense Block	Fully Connected Layer (16×1) Sigmoid Activation Function Denormalize Layer	1
Output	Heart Rate	1

Model Evaluation

In addition to the model's loss function, three error metrics were used to assess its performance: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) (Equations 2.1, 2.2, 2.3). These metrics were calculated alongside the loss function during training to track the model's performance and help determine the best version. In addition to this metrics, certain graphics of the error residuals were plotted resorting to Matplotlib in Python, which is a visualization library.

In order to assess the performance, the model was always run 10 times, and the results were averaged to obtain a more representative value. This approach was necessary because deep learning models are non-deterministic, meaning their results can vary significantly between runs. By averaging the outcomes, a more accurate understanding of the model's true performance distribution was achieved. In contrast, the baseline models, which do not involve deep learning, are deterministic and do not require multiple runs to generate their results.

Chapter 4

Results and Discussion

This chapter presents and discusses the results, evaluating the performance of the proposed approaches and identifying their limitations. As in the previous chapter, the analysis is divided by task, starting with video magnification, followed by the evaluation of average Heart Rate (HR) prediction.

4.1 Eulerian Video Magnification

To assess the model's integrity, an initial evaluation was performed to verify its ability to replicate the results presented in the original paper [1]. This step is crucial for validating both the model itself and the credibility of any enhanced outcomes derived from it. The source video was withdrawn from [77]. Figure 4.1 demonstrates the results achieved by the algorithm, where the two foremost images show distinct frames, corresponding to a minimum (Figure 4.1a) and a maximum point (Figure 4.1b) of the Blood Volume Pulse (BVP) signal. Additionally, a Spatio-Temporal Map (ST Map) was generated (Figure 4.1c). The replicated results match those of the original paper exactly, validating the created model. This successful replication confirms that the model is prepared for further performance testing on the developed dataset.

(a) Frame 1 (Minimum) (b) Frame 2 (Maximum)

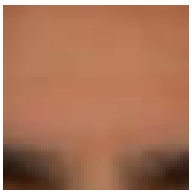
(c) ST Map

Figure 4.1: Replication of Original Paper's Results [1]

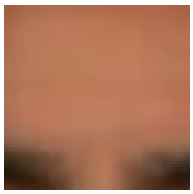
4.1.1 Performance Assessment

The video magnification algorithm was then tested on the created dataset. One example was picked to show the potential impact of such technique, comparing two video frames and the video's ST Map

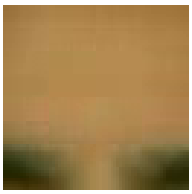
before and after the magnification was performed (Figures 4.2 and 4.3, respectively). The ST Maps of the developed dataset have double the width, given that the original paper's video was only 10 seconds.



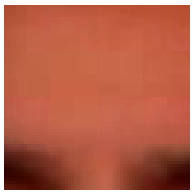
(a) Frame 1 before augmentation



(b) Frame 2 before augmentation



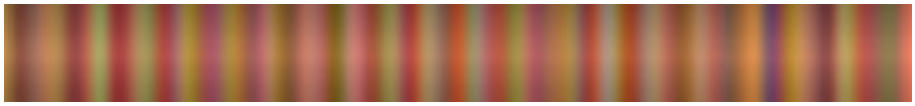
(c) Frame 1 after augmentation



(d) Frame 2 after augmentation

Figure 4.2: Frame Examples (Subject 1)

(a) Input ST Map



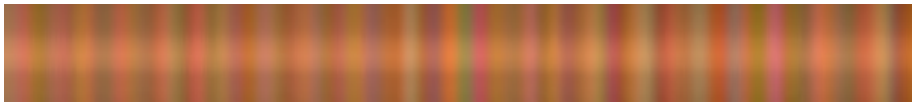
(b) Augmented ST Map

Figure 4.3: ST Maps (Subject 1)

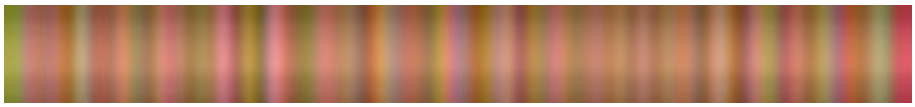
In the frames shown in Figure 4.2, the effect of image magnification is clearly visible. This color variation reflects the underlying blood flow in the skin. The ST Map in Figure 4.3b reveals a cyclical pattern in the color change, but upon closer analysis, some noise is noticeable. The color bands, which should ideally be purely red or greenish hues, also appear in other colors. This occurs due to noise present in the videos, which causes horizontal shifts or alterations in the magnitude of the peaks, thereby distorting the color change pattern.

It is unlikely that this noise was introduced by the recording equipment. It is most likely attributed to the skin's own reflectance, which is expected to be primarily influenced by the lighting conditions the subjects were exposed to (natural light, in this case). Natural illumination exhibits greater variability over time and could also create differences across various regions of the face, both introducing noise in the video. Another variable that may have impacted the skin reflectance on some subjects is the skin tone. A darker skin tone has a lower reflectance level, contributing to a reduced Signal-to-Noise Ratio (SNR). Additionally, it was observed that the SNR degraded as HR increased, suggesting that individuals with higher bpm were more susceptible to the noise present in the videos. Another factor that may have contributed to the noise present in the videos is the compression that they underwent. Since the dataset was formed based on videos with length over 10 minutes, its compression was mandatory in order to save them online. The compression of videos is often times lossy, i.e. information, particularly of the hue, may be lost during compression to generate lighter videos, introducing noise to sensitive tasks that rely on such small variations of color.

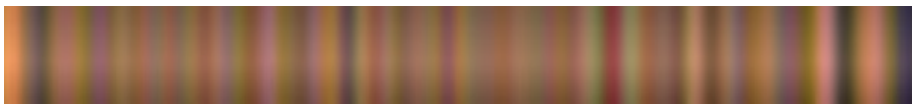
Unfortunately, some of the magnification tests experienced strong noise effects and the achieved results were unlike Figure 4.3b. Some extreme examples of these noise conditions are illustrated in Figure 4.4. The caption for each image has the skin tone (Fitzpatrick scale), Mean HR of the interval, and the illumination level. The light measurements are available with the dataset. An interval with a low light level is assumed to have an illumination of 500 lux or less.



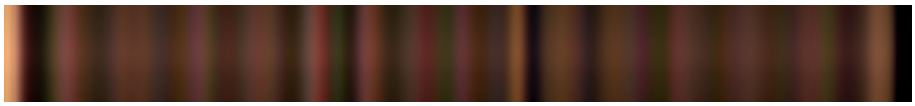
(a) Subject 17 (Skin Tone: 4, HR: 59 bpm, Light: Normal)



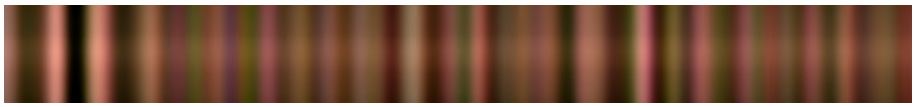
(b) Subject 6 (Skin Tone: 2, HR: 105 bpm, Light: Normal)



(c) Subject 11 (Skin Tone: 1, HR: 83 bpm, Light: Medium to Low)



(d) Subject 2 (Skin Tone: 1, HR: 97 bpm, Light: Low)



(e) Subject 22 (Skin Tone: 4, HR: 142 bpm, Light: Normal)

Figure 4.4: Noisy ST maps

In the first image (Figure 4.4a), for an individual with a darker skin tone (level 4 on the Fitzpatrick scale), the ST Map shows less discernible color bands. This is likely due to a lower magnitude of the captured signal caused by the reduced reflectance of this type of skin. In the next ST Map, Figure 4.4b, the effects of increasing the signal frequency (i.e., higher HR) are highlighted, as more information is lost to noise. It is harder, in this case, to identify a signal pattern when compared to the previous, even with the remaining conditions relatively controlled. This suggests that signals with a higher HR frequency may be more vulnerable to interference with noise, as these require more information in order to be reconstructed. Another example is the reduction of the illumination, which can further reduce the SNR due to the sharp decline in skin reflectance. The effect is similar to what happens with darker skin tones, as can be seen in Figure 4.4c. The last two images, Figures 4.4d and 4.4e, show the effect of stacking the previous factors. As seen in the latter, noise levels are significantly higher, making it even more difficult to extract useful information from the signal.

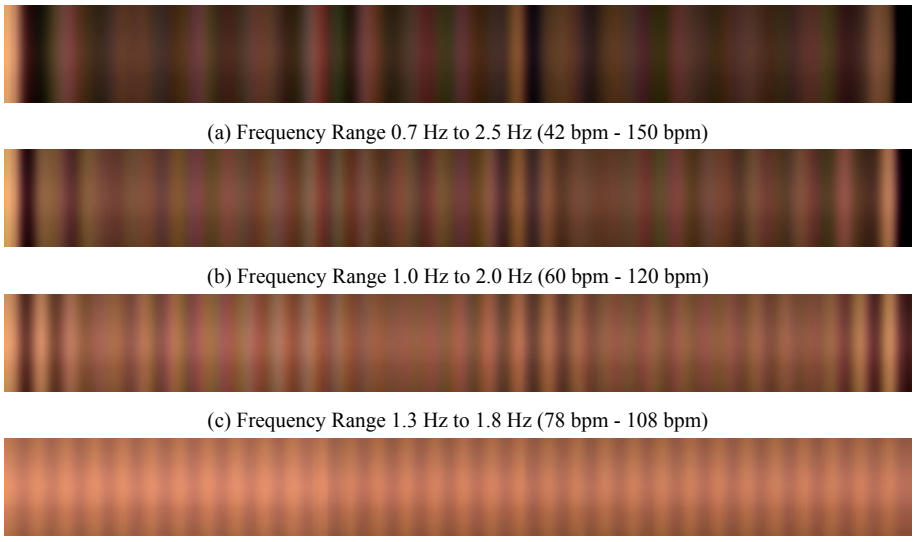
Although the exact contribution of each factor cannot be fully assessed, as they were not isolated in the analysis, it is assumed that lighting had the greatest impact, in line with the general findings and with what was reported in the literature. The dataset was made with these conditions in mind, aiming to replicate a common, everyday situation, to build a more robust model. However, this magnification technique is highly sensitive to noise, and the results were severely affected by the choices made.

4.1.2 Parameter Analysis

Despite high noise levels, it is possible to improve result quality to some extent, as the technique is highly sensitive to both the parameters used and the noise present in the videos. These parameters can play a crucial role in boosting the SNR, either by reducing noise impact or by enhancing the signal itself. To understand their influence on the final results, each parameter was analyzed individually to assess its specific contribution.

Amplification Bandwidth

One of the most critical parameters in noise regulation is the frequency range selected for amplification. A key limitation of this magnification technique is the risk of amplifying not only the desired signal but also the surrounding noise, which will not improve the SNR. A wider frequency range for amplification increases the likelihood of keeping the noise in the final output. A potential solution is to restrict the amplification to a narrower frequency range, specifically targeting the frequencies closer to the actual average HR during each interval. For example, in Figure 4.4d, where one of the highest noise levels is observed, progressively narrowing the frequency range to match the subject's average HR leads to a significant improvement in the signal quality (as illustrated in Figure 4.5).



(a) Frequency Range 0.7 Hz to 2.5 Hz (42 bpm - 150 bpm)

(b) Frequency Range 1.0 Hz to 2.0 Hz (60 bpm - 120 bpm)

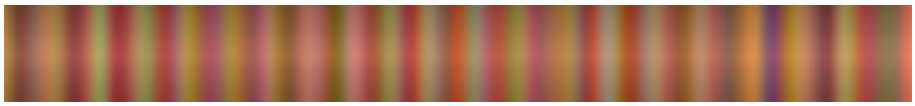
(c) Frequency Range 1.3 Hz to 1.8 Hz (78 bpm - 108 bpm)

(d) Frequency Range 1.55 Hz to 1.65 Hz (93 bpm - 99 bpm)

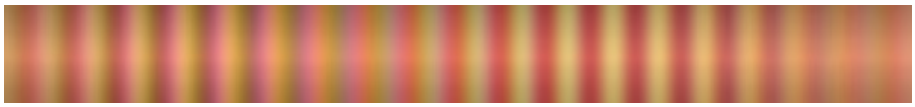
Figure 4.5: Frequency Range Effect on Noise Level (Subject 2, 97 bpm)

Obviously, Figure 4.5d was an extreme case, and applying such a narrow frequency range wouldn't make sense, as HR varies over time and involves multiple frequencies, rather than maintaining a fixed average value (which is only used as the ground truth). Ideally, a frequency range similar to the one used in the second and third images would be more appropriate. However, it's important to consider the trade-offs that using filters with different bandwidths can introduce. As seen in these examples, the bands in the ST Map are more clearly distinguishable, which means there is some signal content in that bandwidth, yet they still lack the typical alternating reddish and greenish coloration expected in a clean signal. This discrepancy may be due to the loss of signal information during noise removal. The remaining information in these bandwidths could even be attributed solely to the noise due to the lack of color variation but it is assumed to be from the signal since it is around the mean value of the HR for this interval. In cases like this one, it is important to keep in mind that the original image already contains

substantial noise, making accurate signal retrieval difficult. However, part of the information may still be recovered. For images with lower initial noise levels, it's possible to recover the signal with a much higher SNR (Figure 4.6).



(a) Frequency Range 0.7 Hz to 2.5 Hz (42 bpm - 150 bpm)



(b) Frequency Range 0.9 Hz to 1.1 Hz (54 bpm - 66 bpm)

Figure 4.6: Frequency Range Effect on Noise Level (Subject 1, 62 bpm)

A major limitation of this parameter, one that is evident but not yet explicitly mentioned, is that it only proves truly advantageous when some form of prior information is available. This is because its effectiveness depends heavily on the specific video in question, and the optimal values for this parameter can vary significantly from one video to another. This poses a critical limitation for algorithms designed to predict or perform magnification on an unknown HR under unpredictable conditions. Specifically for the prediction task, manually adjusting this parameter would make the problem trivial. The model would easily achieve good results as it would be essentially given clues of what the prediction should be. Since both the real HR and the noise level are unknown, the ability to predefine or manually adjust this parameter becomes impractical and even counterintuitive. There is also considerable variability not only between different subjects but also between different intervals within the same subject. As a result, generalizing the parameter to a "one size fits all" approach becomes impossible.

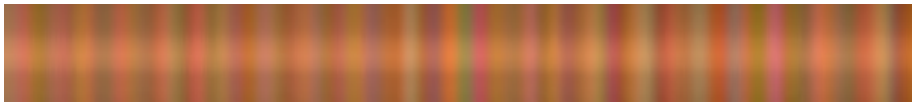
In order to get around this problem, two possible solutions emerge:

1. Controlled conditions: Either video capture conditions need to be tightly controlled to reduce noise to a manageable level, ensuring that it doesn't compromise the analysis. This includes managing illumination, camera quality, and other environmental factors that might introduce noise.

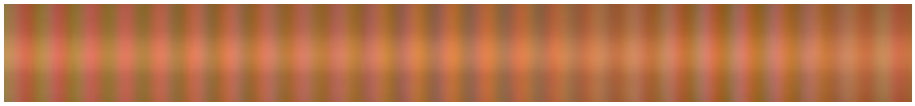
2. Dynamic adaptation: Alternatively, new strategies must be developed to dynamically adjust the parameters based on the real-time conditions in the video.

In this case, since it is not feasible to reduce the amount of noise in the videos (doing so would even contradict the study's objectives), alternative approaches must be considered that can dynamically adjust this parameter. These methods often rely on making initial, rough predictions of the HR to fine-tune the frequency bandwidth, matching the actual real value. Such techniques are particularly effective at reducing noise in reconstructed videos by narrowing the bandwidth used for amplification. However, predictive models can be prone to errors, especially when working with noisy data, as exemplified in some earlier images. If the predicted HR frequencies are significantly inaccurate, the narrowed frequency range may exclude the true HR signal. In such cases, the amplification process ends up amplifying the wrong frequencies, capturing information unrelated to the HR, or perhaps a smaller part of the phenomenon likely due to harmonic frequencies. As a result, the reconstructed video may become misleading. For example, Figure 4.7b shows about 26 red strips, which allows to reach a rough estimate of 78 bpm (each ST Map represents 20 seconds), which is significantly higher than the actual value of 59 bpm. In such scenarios, the technique may become counterproductive by amplifying the wrong frequencies,

thus making it impossible to infer the truthfulness of the result. This creates a delicate balancing act. Narrowing the amplification window can enhance the SNR, but it also increases the risk of missing the true signal if the prediction is inaccurate. On the other hand, using a wider frequency range to mitigate this risk can undermine the benefits of the approach, as it would have a smaller impact.



(a) Frequency Range 0.7 Hz to 2.5 Hz (42 bpm - 150 bpm)



(b) Frequency Range 1.167 Hz to 1.5 Hz (70 bpm - 90 bpm)

Figure 4.7: Effect of Amplifying the wrong Bandwidth (Subject 17, 59 bpm)

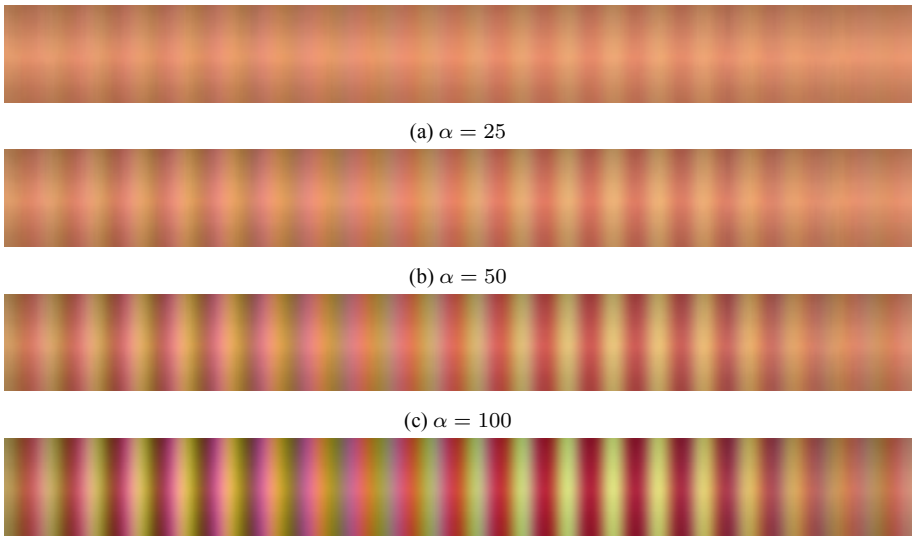
In any case, this method could prove beneficial in the majority of instances, as only a small fraction of cases are significantly impacted by noise to the extent that it causes the model's predictions to fail considerably. Two studies have applied similar methods, though in an iterative fashion [33, 54]. This iterative approach provides a slower convergence, which yields a greater accuracy, thereby increasing the overall effectiveness of the technique. However, both these studies and the simpler method discussed here would substantially increase the computational requirements, as certain parts of the pipeline would need to be run at least twice. This hypothesis was not pursued further in the current study, as it somewhat deviates from one of the primary goals, which was to develop the most efficient model possible. Therefore, this part of the discussion served mainly to highlight the sensitivity of the magnification technique. Nonetheless, this approach presents a valuable avenue for future research.

Amplification Factor

Another parameter that plays a role in signal recovery is the amplification factor (α). It determines the factor by which the Gaussian video, generated by the algorithm prior to reconstructing the final output, is amplified. In other words, it determines how much of the magnified signal is added back to the original video. This parameter can be impactful when utilized in images with a weak signal in the reconstructed video, but it poses some risk, as it will raise both the signal and the noise levels. In instances like Figure 4.4a, where the magnitude of the signal is hindered by the low reflectance of the skin, increasing the amplification factor may be advantageous, but if the signal is noisy it won't further increase SNR. It is mandatory to ensure the image quality is not being further degraded. On the other hand, a signal such as in Figure 4.6b does not pose a risk since it has a low noise level, but is often times not useful because it does not require a higher amplification factor, as the signal is already evident.

In short, it basically gives a higher saturation and sharpness to the color of the magnified video. As a result, it generates an ST Map with more defined and distinguishable bands, as can be seen in Figure 4.8. Although it does not address noise issues, it may be useful in increasing the clarity of results.

Similar to the previous parameter, this one is most effective when there is some prior knowledge of the signal characteristics. The key difference, however, lies in its reliance on the outcome itself; it is adjusted empirically through trial and error, as it is impossible to determine the signal's magnitude within the video without first reconstructing it. Additionally, this parameter minimally impacts the magnification outcome, influencing only the visual quality of the result. Setting a fixed, balanced value, somewhere



(a) $\alpha = 25$

(b) $\alpha = 50$

(c) $\alpha = 100$

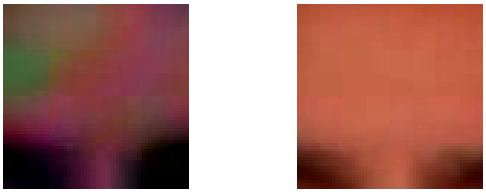
(d) $\alpha = 200$

Figure 4.8: Amplification Factor Effect (Subject 1)

around 100, has been found to produce satisfactory results, with no further fine-tuning of this parameter appearing necessary. A level below 50 tended to produce a faint signal in some cases, while levels significantly above 100 resulted in excessive saturation. Within this range, however, there appears to be an optimal level that balances visibility without over-saturation.

Pyramid Levels

The pyramid level, as previously discussed, determines the Gaussian pyramid level used for reconstructing the video, after filtering. In essence, it sets the spatial resolution of the temporally filtered image and thus controls the resolution of signal induced variations. Lower pyramid levels (higher spatial resolution) capture more fine-grained differences across image regions. In controlled settings, using an intermediate level may be valuable, as it could enable the observation of distinct regional variations on the face, potentially aiding in the identification of facial blood flow patterns (Figure 4.9), an area for future study.

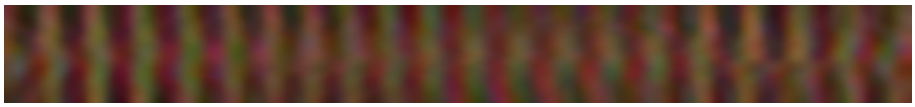


(a) 4^{th} Pyramid Level (b) 7^{th} Pyramid Level

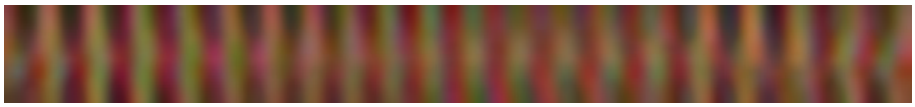
Figure 4.9: Pyramid Level's Effect on Spatial Resolution (Subject 1)

However, in cases with high noise, reducing spatial resolution by increasing the pyramid level suppresses smaller variations, helping to mitigate noise effects. Levels 5 and above, were determined empirically to perform well at reducing the low amplitude noise to satisfactory amount, defining a typical result target. A level of 3 or lower leads to a lot of small amplitude variations in the video reconstruction and the signal clarity becomes lower.

Figure 4.10 illustrates the typical results obtained at each pyramid level, from levels 3 to 6. Level 7



(a) 3^{rd} Pyramid Level



(b) 4^{th} Pyramid Level

(c) 5^{th} Pyramid Level

(d) 6^{th} Pyramid Level

Figure 4.10: Pyramid Level Effect (Subject 1)

was used as the standard in this study, as it fully reduces the low amplitude variations of the image, besides making it optimal for the developed model. All examples prior to this parameter analysis correspond to results achieved using this 7^{th} pyramid level. Using the highest level of spatial reduction entirely sacrifices the spatial resolution of the reconstruction, which in some instances could be a drawback. It may also increase processing demands due to the computational load of maximum reduction, for input images with high resolutions. However, given the already small size of the Region of Interest (ROI), these limitations do not affect the procedure employed in this study. Also, this parameter does not require fine-tuning if a higher level is chosen, as the only factor that could have an effect on it is the spatial dimension. Since it is maintained consistently across all images used, the results will be consistent.

4.1.3 Different ROIs

As anticipated, none of the tested Regions of Interest displayed an inherent advantage over others for magnification purposes. However, differences arose due to varying noise levels in each region. Natural lighting created uneven illumination patterns across regions, which, as previously noted, can significantly impact magnification quality. A general trend observed was that the forehead region typically exhibited less noise than other areas, producing the most consistent results. This may be because, for most subjects, the room's window was positioned above head height (with subjects seated during data collection), providing more stable lighting in the forehead region. For the other two regions, identifying a clear visual pattern was difficult, as both contained some noise.

Another important factor was that regions with larger skin coverage produced better results. In other words, a closer recording distance, results in a ROI with more pixels, which helps reducing noise. This trend was consistent within the current dataset as well. For example, Subject 20 had part of the forehead obscured by hair, resulting in better outcomes for the cheeks compared to the forehead. Notably, the use of glasses or makeup did not appear to affect magnification outcomes

4.2 Estimation Task

4.2.1 Baseline Models

The first step was to assess performance in the created dataset using baseline models. This helps to establish a reference for evaluating the proposed approach. This initial evaluation focused on the forehead region, with other regions of interest tested later. This strategy aimed to reduce variability in results during the hyperparameter tuning of the proposed model, as it was also done using the forehead region. Baseline techniques were thus evaluated in a comparable manner to ensure consistency. This step is crucial, because if the proposed model underperformed or yielded similar results to the baselines, the added complexity of a more advanced model would not be justified for a problem potentially solvable with simpler methods.

The evaluation began with the Base-rPPG model, representing a widely used approach for signal extraction. Next, the Base-EVM method was assessed. Although similar to the Base-rPPG model, the Base-EVM method primarily differs in its signal extraction process, utilizing the magnification pipeline for this purpose. Comparing these methods helps identify whether preprocessing the signal through the magnification algorithm influences the results. The obtained values are summarized in Table 4.1. The results of the Base-EVM model were calculated individually for each color channel (RGB). Conversely, for the Base-rPPG model, following the methodology in the relevant study [76], the ICA component with the highest peak magnitude within the target frequency range in the FFT (0.7 to 2.5 Hz) was selected.

Table 4.1: Baseline Models Results (Forehead)

Model	MAE (bpm)	MAPE (%)	RMSE (bpm)
Base-rPPG	10.86	14.07	13.47
Base-EVM (R)	10.76	14.32	13.64
Base-EVM (G)	6.63	8.15	9.93
Base-EVM (B)	10.38	13.83	13.08

As anticipated from the literature and underlying theory, the green channel contains the most HR information. Among the baseline models, the Base-EVM algorithm, when applied to the green channel, achieved the best results. Does this imply that the preprocessing of the Base-EVM model with the magnification pipeline is superior to the one of the Base-rPPG model? Not necessarily. Abstracting from the specific methods, we can identify four primary steps integral to both approaches: spatial reduction, temporal filtering, normalization, and noise regulation techniques, with ICA serving as one example of a noise regulation method. If anything, the Base-rPPG model is expected to achieve greater results since it employs more techniques in this regard. Additionally, what happens is that when the Base-rPPG algorithm is executed without ICA, the results are nearly identical to those obtained with the Base-EVM, with only negligible differences. The achieved results are unexpected and somehow suggest that the noise removal technique did not perform as expected, given the fact that it was the major difference between the two approaches. Furthermore, given the similarity between the processing on both algorithms it is not expected that EVM has any positive influence on the estimation task and its advantages stem solely from the visualization aid, contrary to what was suggested in previous literature.

Although ICA is commonly effective for noise reduction, it has a key limitation: it assumes each source is independent, which may not be valid in this context. Given that the noise overlaps with essential features of the true signal and that the SNR is already very low, ICA may struggle to isolate the signals

effectively. Consequently, the signal, primarily concentrated in the green channel, could even become dispersed across multiple source signals estimated by ICA, further reducing the SNR and resulting in less accurate HR estimations.

There is also a remote possibility that the technique was applied differently than in the original article. One of the major challenges in this field, as highlighted by the reviewed literature, is the scarcity of open-source algorithms and specific information needed to replicate methods precisely. Nonetheless, the method was reconstructed as carefully as possible with the available details, and it is not to be expected that this error would have occurred.

From here on, the Base-EVM model with the Green channel served as a performance benchmark for the proposed deep learning model.

4.2.2 Proposed Model

The results obtained with the validation and test datasets are presented in Table 4.2. As mentioned, this initial assessment was carried out with the forehead region. Validation metrics were calculated during training, with values recorded for the iteration showing the lowest loss. Testing metrics were then derived by evaluating the fully trained model on the test set.

Table 4.2: Proposed Model Training Results (Forehead)

Proposed Model	**MAE (bpm)**	**MAPE (%)**	**RMSE (bpm)**
Validation Set	4.69 ± 0.38	5.70 ± 0.43	6.34 ± 0.49
Test Set	4.57 ± 0.87	5.93 ± 1.21	6.34 ± 1.07

The proposed estimation method outperformed the baseline technique, demonstrating particular resilience against dataset noise, an issue that often hindered signal analysis in previous models. By incorporating information from all three color channels, the model more effectively recovered signals obscured by noise. The greatest improvement was observed in the RMSE metric, with a reduction of approximately 36% compared to the Base-EVM model, suggesting that the proposed model successfully minimized large errors in challenging cases (since RMSE emphasizes larger errors). Additionally, general performance improvements were evident, with MAE and MAPE reduced by roughly 29% and 27%, respectively. The results obtained, while not equally accurate when compared with medical grade equipment, have a competitive performance with wearables present on the market, as determined by some articles [78–80].

As shown in Figure 4.11, illustrating one of the training runs, both the training and validation loss curves converged steadily to a similar point, indicating effective training without signs of underfitting or overfitting. Its behavior demonstrates stable learning across epochs. The early stopping technique proved effective, halting training when the validation loss began to increase, thereby preventing overfitting. Additionally, the close alignment of validation and test metrics, with non significative differences, suggests that the model generalized well and did not overfit to subject-specific features, yielding strong results for unseen subjects.

4.2.3 Different ROIs

To evaluate estimation performance across the other Regions of Interest, the same procedure was applied to each cheek. Both baseline models and the proposed deep learning model, using the same ar-

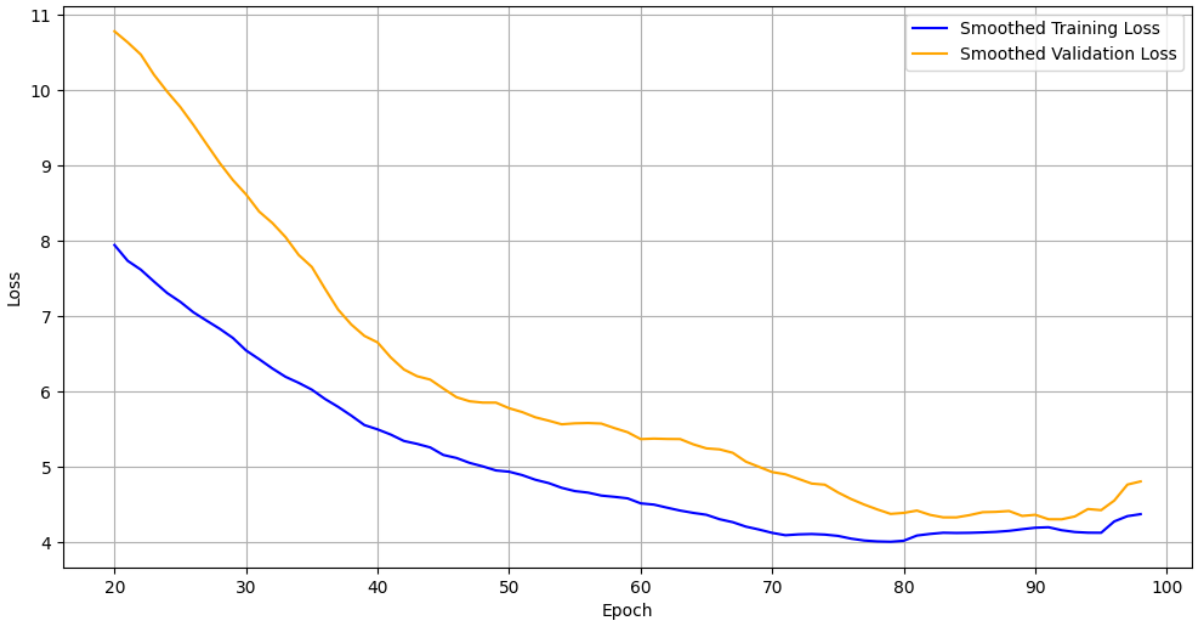


Figure 4.11: Training Loss Curves (Forehead)

chitecture, were tested individually on each cheek. Model training and testing were done on the dataset of the respective region. Tables 4.3 and 4.4 summarize the results for the right and left cheeks, respectively.

Table 4.3: Right Cheek Results

Model	MAE (bpm)	MAPE (%)	RMSE (bpm)
Base-rPPG	10.40	13.48	12.96
Base-EVM (R)	10.55	14.15	13.22
Base-EVM (G)	**10.03**	13.31	12.56
Base-EVM (B)	10.65	14.37	13.27
Proposed Model	10.09 ± 0.48	**13.20** ± 0.49	**11.72** ± 0.62

Table 4.4: Left Cheek Results

Model	MAE (bpm)	MAPE (%)	RMSE (bpm)
Base-rPPG	10.87	14.04	13.37
Base-EVM (R)	10.44	13.92	13.15
Base-EVM (G)	**8.75**	**11.26**	**11.46**
Base-EVM (B)	10.51	13.97	13.12
Proposed Model	9.72 ± 0.51	12.83 ± 0.66	11.66 ± 0.65

The Base-EVM model yielded similar results on the red and blue color channels for both regions, as expected. In the case of the green channel, while still yielding the best values, it provided significantly less information in these Regions of Interest, thus achieving worse results. This outcome supports the hypothesis that these regions have a lower SNR than the forehead, likely influenced by lighting conditions during capture. Additionally, the right cheek exhibited higher noise levels than the left counterpart, having only slightly better values in the green channel when compared to the other color channels and the Base-rPPG model. This elevated noise level was unexpected, and its exact cause remains unidentified. It

may relate to light incidence on the face, since natural lighting, as previously mentioned, can introduce considerable variability across regions. The consistent results in this area suggest a systematic issue, assumed to have caused uneven lighting between the two sides of the face. It was potentially due to the room setup (a wall to the right and a larger window area on the left) or the angle of sunlight during data capture. However, the source of this noise remains unknown.

Interestingly, the consistency of results from the red and blue channels with previous findings suggests that the advantage of using all three channels may stem less from additional information in the red and blue channels and more from the model's ability to identify noise patterns across channels. This capability appears to allow the model to better separate useful information in the green channel, even in noisier conditions. Explainability methods could help further investigate this behavior. The Base-rPPG model results were also relatively consistent across regions, although this consistency is of limited relevance given the persistently poor performance.

This analysis highlights a distinct pattern in data quality across Regions of Interest. The forehead enabled the Base-EVM (Green) model to achieve the highest performance, followed by the left cheek, with the right cheek showing results nearly indistinguishable from the Base-rPPG model or the Base-EVM with other channels. A similar outcome was witnessed while testing the proposed deep learning model on these regions. Performance on them was notably lower than on the forehead. For the right cheek, the proposed model's results matched those of the Base-EVM (Green), which were already only slightly better than the other baseline models. On the left cheek, while performing better than on the right counterpart, the proposed model could not even achieve Base-EVM (Green) performance, even though it still managed to outperform the other baseline models in this region.

Training convergence of the proposed model was different for this regions as well, generally showing suboptimal results, suggesting that the noise increase in these two regions hindered learning performance. The increased complexity of the problem due to added noise impaired the model's ability to converge effectively, making pattern identification in the data more challenging. The training loss curves (see Figure 4.12) illustrate this, with the validation curve often failing to improve and diverging from the start. Overall, the model tended to overfit as it attempted to capture data variance or, with a slight reduction of the architecture complexity, began to underfit. The parameters and architecture, optimized for the forehead, were unable to adapt well to the varying noise levels in other regions, underscoring the "No Free Lunch" principle. In short, the cross region generalization of the model did not yield great results.

To address these challenges, an alternative approach was tested: the model was trained initially on the forehead region and subsequently tested on other regions, with the entire dataset from the new region acting as the test set. Table 4.5 presents the results of this cross-regional training strategy.

Table 4.5: Cross Region Training Results

Region	**MAE (bpm)**	**MAPE (%)**	**RMSE (bpm)**
Right Cheek	10.34 ± 0.60	13.96 ± 0.90	13.46 ± 0.89
Left Cheek	8.26 ± 0.56	11.02 ± 0.78	11.16 ± 0.42

For the left cheek, a region with less noise than its counterpart, this cross-region training approach proved advantageous, allowing the model to achieve improved results. This improvement likely stems from the similar conditions to what was seen for the forehead, with the green channel showing a rel-

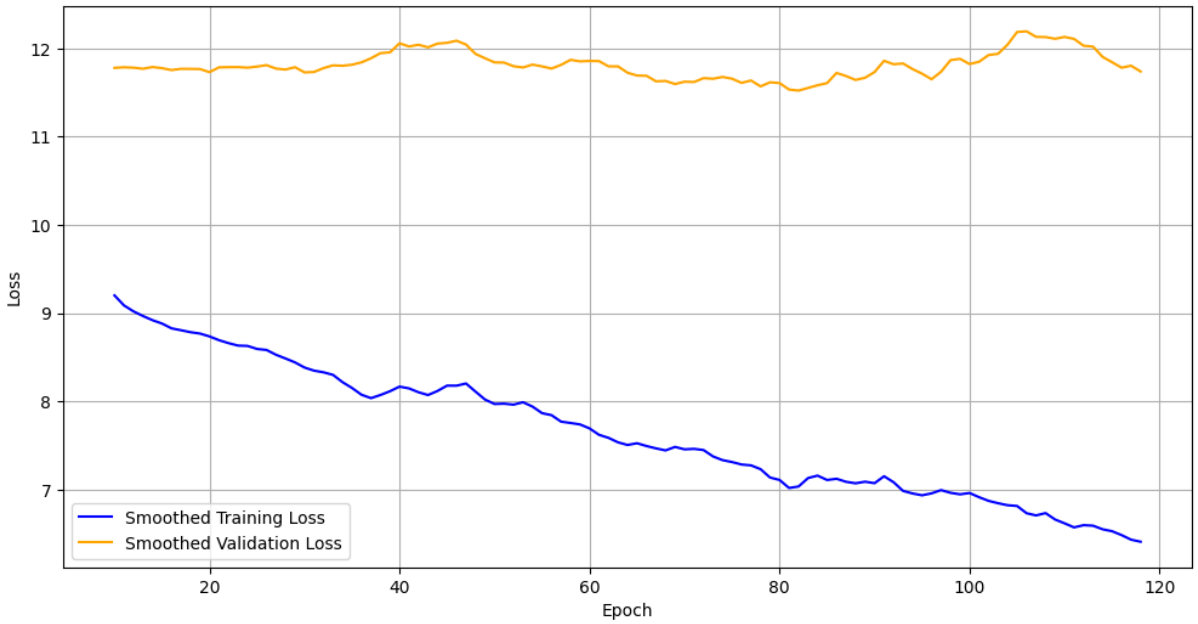


Figure 4.12: Training Loss Curves (Cheeks)

atively high SNR. Although training directly on the left cheek dataset may present some noise related challenges, the forehead dataset training leads to a good performance due to the similarity between both problems. Applying a transfer learning strategy, where specific network layers are fine-tuned to adapt to the left cheek data, could further enhance these results. While this is beyond the current study's scope, it represents a promising avenue for future research.

Conversely, the cross region approach did not yield similar gains for the right cheek, where results were slightly worse with this training method. In this case, noise significantly compromised the green channel's SNR, making the problem notably more complex than for the forehead or left cheek. This increased noise diminished the model's efficacy, highlighting the limitations of cross region training when SNR is critically low and variability in signal quality between regions is high.

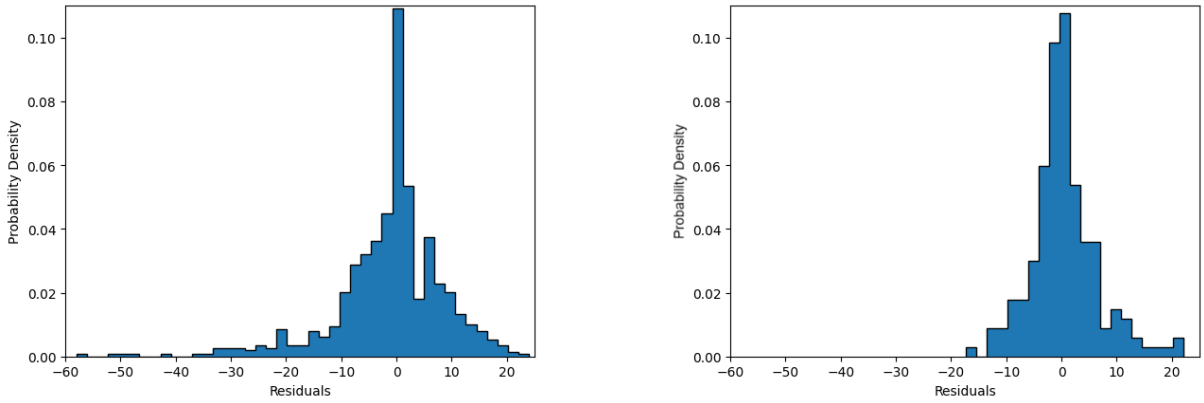
The estimation task is evidently more affected by noise levels than by the specific region itself. Achieving high-quality results would be challenging from regions with noise profiles comparable to that of the right cheek. However, with appropriate parameter tuning and/or architectural adjustments, regions like the forehead and left cheek show promise for producing reliable results, as the noise in these areas remains within a manageable range. Focusing on the forehead region during model tuning likely introduced a region-specific bias, leading to significantly better performance on the forehead compared to other areas. This highlights how targeted tuning can optimize model performance for particular regions, although it may restrict the model's generalizability to regions with different noise characteristics.

4.2.4 Residual Analysis

An error residual analysis was conducted to detect any potential trends in the model's prediction errors. This analysis compared the Base-EVM (Green) model with the proposed model, specifically focusing on the forehead region, as these conditions demonstrated the most promising results. Notably, a limiting factor in this comparison is the data subset size: the proposed model's analysis used only 20% of the data, while the baseline model was evaluated on the full dataset. However, since the data

was split to ensure consistent distribution across groups, this should minimally impact the comparative findings. Furthermore, given the non-deterministic behavior of the deep learning models, the examples of the proposed model shown in this analysis are the result of an average run.

The initial analysis examined the distribution of residuals, shown in Figure 4.13 for both models. Key differences emerged, notably the narrower spread in the proposed model's distribution, which aligns with its capacity to reduce larger errors. Errors for the proposed model were more densely concentrated within an absolute error of 10, with nearly none exceeding 20. Additionally, the residual distribution for the proposed model showed a reduced tail compared to the baseline's.



(a) Base-EVM (Green)

(b) Proposed Model

Figure 4.13: Residual Distribution

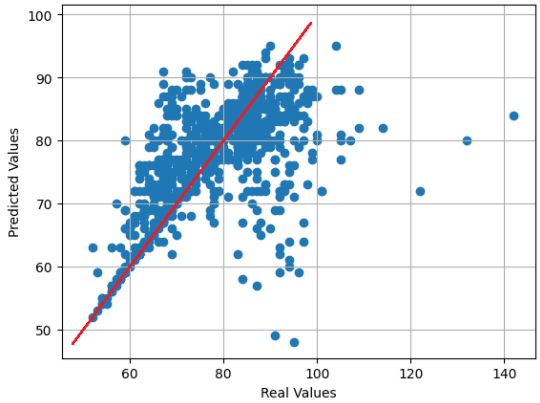
To further validate these observations and better assess the normality of the error distributions, skewness and kurtosis metrics were calculated (Table 4.6). The proposed model showed a skewness value closer to zero, suggesting a more symmetrical error distribution, while the Base-EVM model exhibited a leftward tail, indicating a tendency to underpredict relative to the actual values. Kurtosis, which measures tail heaviness compared to a normal distribution, on average was higher for the baseline model, supporting of a greater presence of extreme values or outliers. Even so, the proposed model has a kurtosis value far from zero, mainly due to the strong presence of errors equal to zero in most of the runs.

Table 4.6: Skewness and Kurtosis of the Residuals

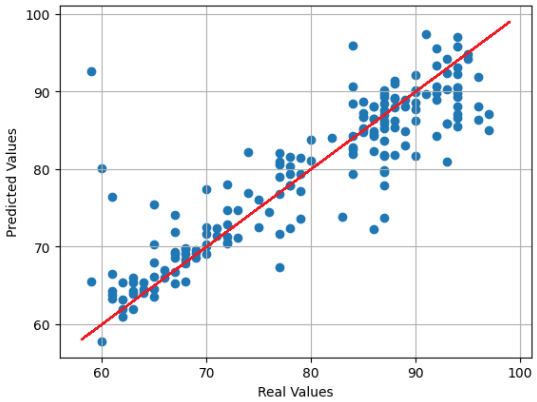
Model	**Skewness**	**Kurtosis**
Base-EVM (Green)	-1.36	4.41
Proposed Model	0.48 ± 0.67	2.87 ± 2.78

To evaluate this tendency in the baseline model's predictions, residuals and predictions were also plotted against the real HR values (Figure 4.14). The two upper plots (Figures 4.14a and 4.14b) display scatter plots comparing predictions to actual values for both models. Ideally, points would cluster along a diagonal line with a slope of 1, representing perfect prediction. In reality, a normal distribution around this line is expected due to prediction errors. The proposed model's predictions show a normal distribution along this line, indicating balanced performance (Figure 4.14b). In contrast, the baseline model exhibits a noticeable deviation, with more points falling below the line, confirming its tendency to underpredict (Figure 4.14a). Additionally, larger errors appear to have a correlation with higher actual HR values. To further investigate this trend, residuals were plotted against actual values (Figures 4.14c and 4.14d). Points in this plots are expected to align closely around zero, representing more accurate predictions.

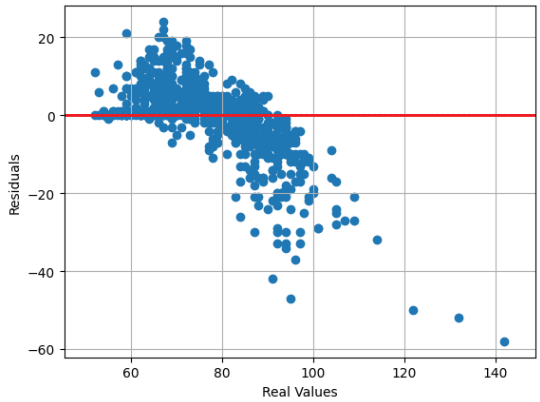
For the baseline model (Figure 4.14c), a clear residual trend emerges: it tends to overestimate at lower Heart Rates and underestimate at higher ones, with the largest residuals occurring at higher Heart Rates. This further confirms that elevated Heart Rates are more vulnerable to noise interference, impacting both signal quality and estimation accuracy. The proposed model addresses some of these baseline limitations handling the noise levels, thereby reducing errors, particularly in higher HR ranges, and resulting in an overall performance improvement. Still, some trend can be identified visually, although it doesn't have such a steep slope (Figure 4.14d). It occurs because the proposed model is also somewhat affected by the noise.



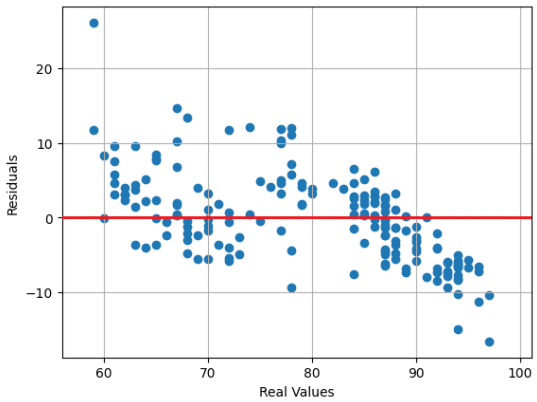
(a) Actual vs Predicted Values (Baseline)



(b) Actual vs Predicted Values (Proposed)



(c) Residuals by Real Values (Baseline)



(d) Residuals by Real Values (Proposed)

Figure 4.14: Errors Scatter Plot

The influence of other noise factors that hindered magnification, such as lighting and skin tone, could not be properly assessed. In order to isolate this variables, the already small subset of data (20%) would become even smaller, which would make it difficult to draw conclusions.

Chapter 5

Conclusion

Addressing magnification first, although it functioned correctly, the results did not meet expectations, primarily due to the significant influence of noise on this technique. Moderate noise does not drastically affect the visual output, as changes in facial color (though sometimes inaccurate) and brightness variations can still be observed in the reconstructed video. However, while the phenomenon remains visible, the results fail to accurately reflect reality and thus remain unsatisfactory. This noise stems from multiple factors, making it challenging to isolate each one's individual impact. Lighting is likely the primary factor, as it is the most extensively studied in the literature, yet any element affecting skin reflectance critically impacts result quality. Additionally, higher Heart Rate signals appeared particularly susceptible to noise interactions, leading to information loss.

Adjusting the magnification bandwidth could serve as an initial step in noise reduction; however, as this parameter cannot be set in advance, tuning it involves a trade-off between efficiency and accuracy. While promising, the techniques discussed here suggest future research potential. Other parameters mainly adjust the visual quality of results and can be effectively handled with a generalized, "one size fits all" approach, but are not particularly useful in noise regulation.

Even if the results were suboptimal, the challenges cannot be attributed solely to the algorithm or dataset but rather to the experimental conditions. These conditions were designed to mimic real-world acquisition, yet the model is not equipped for entirely uncontrolled scenarios; it requires either some level of regulation or new techniques that allow it to adapt in real-time. It is also important to give the disclaimer that, contrary to the beliefs in some studies, EVM is just a visualization technique and should have no inherent characteristic that aids in estimation, given the similarity between typical rPPG and EVM processing. The advantage of incorporating a method like this in the pipeline still stands, as the technique still has potential applications linked to its spatial visualization. For example, in clinical practice and beyond, future research could explore whether facial blood flow patterns are discernible, a possibility that might hold diagnostic value for certain conditions.

As for the estimation task, the proposed model outperformed the baselines but not consistently across all cases. When trained and tested within each specific region, only the forehead demonstrated superior performance, while both the left and right cheeks exhibited relatively poorer results. This disparity appears to stem not from intrinsic characteristics of these regions but rather from varying noise levels, as evidenced by similar performance variability in the best baseline model, and in the magnification task. Testing a cross-training approach, in which the proposed model was trained on the forehead

and subsequently tested on other regions, showed promising outcomes, though further investigation is needed.

Future research should include testing with different measurement instruments and varied environments, as increased diversity could reduce biases introduced by specific data collection setups and improve generalization of the proposed model.

The results achieved in this study, while not always great, were rather insightful. Both the magnification and estimation tasks share several limitations, with noise being a crucial factor impacting their performance, especially for magnification, though it affects estimation as well. Considering the efficiency importance, the current pipeline relies on multiple deep learning algorithms (for face tracking and estimation), which prioritize accuracy over some efficiency, a trade-off that may benefit from further exploration. The code could probably be further optimized, which would also reduce computational demands. Lastly, an important direction for future research would be developing a real-time algorithm. Real-time capability was not pursued in this study given the substantial work still needed to achieve broad generalization.

Bibliography

[1] Hao-Yu Wu et al. “Eulerian Video Magnification for Revealing Subtle Changes in the World”. In: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31.4 (2012).

[2] Adam Bohr and Kaveh Memarzadeh. “The rise of artificial intelligence in healthcare applications”. In: *Artificial Intelligence in Healthcare*. Elsevier, 2020, pp. 25–60. ISBN: 9780128184387. DOI: 10.1016/b978-0-12-818438-7.00002-2. URL: http://dx.doi.org/10.1016/B978-0-12-818438-7.00002-2.

[3] Srecko Joksimovic et al. “Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review”. In: *Computers and Education: Artificial Intelligence* 4 (2023), p. 100138. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2023.100138. URL: https://www.sciencedirect.com/science/article/pii/S2666920X23000176.

[4] Charu Krishna, Dinesh Kumar, and Dharmender Singh Kushwaha. “A Comprehensive Survey on Pandemic Patient Monitoring System: Enabling Technologies, Opportunities, and Research Challenges”. In: *Wireless Personal Communications* 131.3 (June 2023), pp. 2125–2172. ISSN: 1572-834X. DOI: 10.1007/s11277-023-10535-9. URL: http://dx.doi.org/10.1007/s11277-023-10535-9.

[5] Roxana Filip et al. “Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems”. In: *Journal of Personalized Medicine* 12.8 (2022). ISSN: 2075-4426. DOI: 10.3390/jpm12081295. URL: https://www.mdpi.com/2075-4426/12/8/1295.

[6] Idar Johan Brekke et al. “The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review”. In: *PLOS ONE* 14.1 (Jan. 2019), pp. 1–13. DOI: 10.1371/journal.pone.0210875. URL: https://doi.org/10.1371/journal.pone.0210875.

[7] Sirina Keesara, Andrea Jonas, and Kevin Schulman. “Covid-19 and health care’s digital revolution”. In: *New England Journal of Medicine* 382.23 (2020), e82.

[8] Li Ran et al. “Risk factors of healthcare workers with coronavirus disease 2019: a retrospective cohort study in a designated hospital of Wuhan in China”. In: *Clinical infectious diseases* 71.16 (2020), pp. 2218–2221.

[9] Esther Monica Pei Jin Fan et al. “Factors to Consider in the Use of Vital Signs Wearables to Minimize Contact With Stable COVID-19 Patients: Experience of Its Implementation During the Pandemic”. In: *Frontiers in Digital Health* 3 (Sept. 2021). ISSN: 2673-253X. DOI: 10.3389/fdgth.2021.639827. URL: http://dx.doi.org/10.3389/fdgth.2021.639827.

[10] Estela Kristal-Boneh et al. “Heart rate variability in health and disease”. In: *Scandinavian journal of work, environment & health* (1995), pp. 85–95.

[11] Robert E. Kleiger, Phyllis K. Stein, and J. Thomas Bigger. “Heart Rate Variability: Measurement and Clinical Utility”. In: *Annals of Noninvasive Electrocardiology* 10.1 (Jan. 2005), pp. 88–101. ISSN: 1542-474X. DOI: 10.1111/j.1542-474x.2005.10101.x. URL: http://dx.doi.org/10.1111/j.1542-474X.2005.10101.x.

[12] Ali A Mehdirad Tarek Ajam. *Medscape: Electrocardiography*. Mar. 2019. URL: https://emedicine.medscape.com/article/1894014-overview?_gl=1*1watoww*_gcl_au*MTY0ODQ5MDk2MS4xNzI3NDUwNjY2#a1.

[13] Bruce M Lo. *Medscape: Pulse Oximetry*. Jan. 2021. URL: https://emedicine.medscape.com/article/2116433-overview?_gl=1*1uugj5o*_gcl_au*MTY0ODQ5MDk2MS4xNzI3NDUwNjY2.

[14] Pireh Pirzada et al. “Remote Photoplethysmography (rPPG): A State-of-the-Art Review”. In: *medRxiv* (2023), pp. 2023–10.

[15] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. “Remote plethysmographic imaging using ambient light”. In: *Optics Express* 16.26 (Dec. 2008), p. 21434. ISSN: 1094-4087. DOI: 10.1364/oe.16.021434. URL: http://dx.doi.org/10.1364/oe.16.021434.

[16] F. P. Wieringa, F. Mastik, and A. F. W. van der Steen. “Contactless Multiple Wavelength Photoplethysmographic Imaging: A First Step Toward “SpO2 Camera” Technology”. In: *Annals of Biomedical Engineering* 33.8 (Aug. 2005), pp. 1034–1041. ISSN: 1573-9686. DOI: 10.1007/s10439-005-5763-2. URL: http://dx.doi.org/10.1007/s10439-005-5763-2.

[17] Mona Alnaggar et al. “Video-based real-time monitoring for heart rate and respiration rate”. In: *Expert Systems with Applications* 225 (2023). DOI: 10.1016/j.eswa.2023.120135.

[18] Javaan Chahl Ali Al-Naji and Sang-Heon Lee. “Cardiopulmonary signal acquisition from different regions using video imaging analysis”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 7.2 (2019), pp. 117–131. DOI: 10.1080/21681163.2018.1441075.

[19] Mavlonbek Khomidov, Deokwoo Lee, and Jong-Ha Lee. “A Novel Contactless Blood Pressure Measurement System and Algorithm Based on Vision Intelligence”. In: *Electronics (Switzerland)* 12.24 (2023). DOI: 10.3390/electronics12244898.

[20] Matthew J. Page et al. “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews”. In: *International Journal of Surgery* 88 (2021), p. 105906. ISSN: 1743-9191. DOI: https://doi.org/10.1016/j.ijsu.2021.105906.

[21] Matthew J Page et al. “PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews”. In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n160.

[22] *Inclusion and exclusion criteria - Systematic Reviews for Health Sciences and Medicine - Library Guides at University of Melbourne*. URL: https://unimelb.libguides.com/c.php?g=492361&p=3368110.

[23] Microsoft Corporation. *Microsoft Excel*. Version 2019 (16.0). Sept. 24, 2018. URL: https://office.microsoft.com/excel.

[24] Gaganjot Kaur. and Jeff Kilby. “Contactless Heart Rate Measurement using Image Processing”. In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - BIOSIGNALS*. INSTICC. SciTePress, 2022, pp. 111–116. ISBN: 978-989-758-552-4. DOI: 10.5220/0010761400003123.

[25] Neal R. Haddaway et al. “PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis”. In: *Campbell Systematic Reviews* 18.2 (2022), e1230. DOI: https://doi.org/10.1002/cl2.1230.

[26] Ali Al-Naji, Asanka G. Perera, and Javaan Chahl. “Remote monitoring of cardiorespiratory signals from a hovering unmanned aerial vehicle”. In: *BioMedical Engineering Online* 16.1 (2017). DOI: 10.1186/s12938-017-0395-y.

[27] Ali Al-Naji and Javaan Chahl. “Simultaneous tracking of cardiorespiratory signals for multiple persons using a machine vision system with noise artifact removal”. In: *IEEE Journal of Translational Engineering in Health and Medicine* 5 (2017). DOI: 10.1109/JTEHM.2017.2757485.

[28] Asawari K. Chinchanikar and Manisha P. Dale. “Blood Volume Pulse Extraction Method of Heart Rate Estimation”. In: 2023. DOI: 10.1109/ICCUBEA58933.2023.10392039.

[29] Diane Elhajjar et al. “Assessing Confidence in Video Magnification Heart Rate Measurement using Multiple ROIs”. In: *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. 2023, pp. 1–6. DOI: 10.1109/I2MTC53148.2023.10175914.

[30] Pallavi Genu Pansare and M.P. Dale. “Heart Rate Measurement from Face and Wrist Video”. In: 2018. DOI: 10.1109/ICCUBEA.2018.8697722.

[31] Bauyrzhan Aubakir et al. “Vital sign monitoring utilizing Eulerian video magnification and thermography”. In: vol. 2016-October. 2016, pp. 3527–3530. DOI: 10.1109/EMBC.2016.7591489.

[32] Kian Hamedani, Zahra Bahmani, and Amin Mohammadian. “Spatio-temporal filtering of thermal video sequences for heart rate estimation”. In: *Expert Systems with Applications* 54 (2016), pp. 88–94. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2016.01.022.

[33] Yasmina Souley Dosso, Amente Bekele, and James R. Green. “Eulerian Magnification of Multi-Modal RGB-D Video for Heart Rate Estimation”. In: 2018. DOI: 10.1109/MeMeA.2018.8438741.

[34] Bo-Yu Huang and Chi-Lun Lin. “Improvement of Environment and Camera Setting on Extraction of Heart Rate Using Eulerian Video Magnification”. In: vol. 74. 2020, pp. 381–388. DOI: 10.1007/978-3-030-30636-6_52.

[35] Bin Lin et al. “Estimation of vital signs from facial videos via video magnification and deep learning”. In: *iScience* 26.10 (2023). DOI: 10.1016/j.isci.2023.107845.

[36] Ennio Gambi, Manola Ricciuti, and Susanna Spinsante. “Sensitivity of the contactless videoplethysmography-based heart rate detection to different measurement conditions”. In: vol. 2018-September. 2018, pp. 767–771. DOI: 10.23919/EUSIPCO.2018.8553167.

[37] Ennio Gambi et al. “Heart rate detection using microsoft kinect: Validation and comparison to wearable devices”. In: *Sensors (Switzerland)* 17.8 (2017). DOI: 10.3390/s17081776.

[38] Ghulam Abbas et al. “Scope of Video Magnification in Human Pulse Rate Estimation”. In: *2017 International Conference on Machine Vision and Information Technology (CMVIT)*. 2017, pp. 69–75. DOI: 10.1109/CMVIT.2017.28.

[39] Steven Lawrence Fernandes et al. “A novel nonintrusive decision support approach for heart rate measurement”. In: *Pattern Recognition Letters* 139 (2020), pp. 148–156. DOI: 10.1016/j.patrec.2017.07.002.

[40] Sachin M. Karmuse, Arun L. Kakhandki, and Mallikarjun Anandhalli. “A Robust rPPG Approach for Continuous Heart Rate Measurement Based on Face”. In: *Journal of The Institution of Engineers (India): Series B* (2022). DOI: 10.1007/s40031-022-00817-4.

[41] Xiujuan Zheng et al. “Remote measurement of heart rate from facial video in different scenarios”. In: *Measurement: Journal of the International Measurement Confederation* 188 (2022). DOI: 10.1016/j.measurement.2021.110243.

[42] Karim Alghoul et al. “Heart Rate Variability Extraction from Videos Signals: ICA vs. EVM Comparison”. In: *IEEE Access* 5 (2017), pp. 4711–4719. DOI: 10.1109/ACCESS.2017.2678521.

[43] Ali Al-Naji, Kim Gibson, and Javaan Chahl. “Remote sensing of physiological signs using a machine vision system”. In: *Journal of Medical Engineering and Technology* 41.5 (2017), pp. 396–405. DOI: 10.1080/03091902.2017.1313326.

[44] Ying Qiu et al. “EVM-CNN: Real-Time Contactless Heart Rate Estimation from Facial Video”. In: *IEEE Transactions on Multimedia* 21.7 (2019), pp. 1778–1787. DOI: 10.1109/TMM.2018.2883866.

[45] Tashfiq Rahman, Worarat Krathu, and Chonlameth Arpnikanondt. “Heart Rate Measurement on PC and Phone using Facial Videos”. In: 2023. DOI: 10.1109/KST57286.2023.10086729.

[46] Tiago Palma Pagano et al. “Remote Heart Rate Prediction in Virtual Reality Head-Mounted Displays Using Machine Learning Techniques”. In: *Sensors* 22.23 (2022). DOI: 10.3390/s22239486.

[47] Lijia Liu and Gang Wang. “Remote Heart Rate Estimation in Low-Light Environments Based on Eulerian Video Magnification”. In: 2022. DOI: 10.1109/CISP-BMEI56279.2022.9979977.

[48] Leen Yassin Kassab et al. “Effects of Lighting and Window Length on Heart Rate Assessment through Video Magnification”. In: 2022. DOI: 10.1109/SAS54819.2022.9881347.

[49] Leen Yassin Kassab et al. “Effects of region of interest size on heart rate assessment through video magnification”. In: 2021. DOI: 10.1109/MeMeA52024.2021.9478596.

[50] Ahmed Alzahrani et al. “Reducing Motion Impact on Video Magnification Using Wavelet Transform and Principal Component Analysis for Heart Rate Estimation”. In: vol. 2021-May. 2021. DOI: 10.1109/I2MTC50364.2021.9460026.

[51] Ruqiang Huang et al. “Remote Measurement of Vital Signs for Unmanned Search and Rescue Vehicles”. In: 2020, pp. 164–168. DOI: 10.1109/ICCRE49379.2020.9096468.

[52] Cheng Wang et al. “Non-contact measurement of heart rate based on facial video”. In: 2019, pp. 2269–2275. DOI: 10.1109/PIERS-Fall48861.2019.9021402.

[53] Kai Cai et al. “A passive heart rate measurement method using camera”. In: 2018, pp. 68–72. DOI: 10.1145/3232829.3232841.

[54] Abhijit Sarkar et al. “Evaluation of video magnification for nonintrusive heart rate measurement”. In: 2016, pp. 494–498. DOI: 10.1109/CMI.2016.7413797.

[55] Meshram Rohit Pramod et al. “Remote Heart Ailment Detection using Eulerian Video Magnification”. In: 2023. DOI: 10.1109/ICACTA58201.2023.10392739.

[56] Guillaume Heusch, Sébastien Marcel, and André Anjos. *COHFACE*. 2016.

[57] Ewa Magdalena Nowara et al. "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 1272–1281.

[58] Serge Bobbia et al. "Unsupervised skin tissue segmentation for remote photoplethysmography". In: *Pattern Recognit. Lett.* 124 (June 2019), pp. 82–90.

[59] *MMSE-HR dataset (Multimodal Spontaneous Expression-Heart Rate dataset)*. URL: https://binghamton.technologypublisher.com/tech/MMSE-HR_dataset_(Multimodal_S%20pontaneous_Expression-Heart_Rate_dataset).

[60] Paul Viola and Michael J. Jones. "Robust Real-Time Face Detection". In: *International Journal of Computer Vision* 57.2 (2004), pp. 137–154. DOI: 10.1023/B:VISI.0000013087.49260.fb.

[61] Carlo Tomasi and Takeo Kanade. "Detection and tracking of point". In: *Int J Comput Vis* 9.137-154 (1991), p. 2.

[62] *MediaPipe - On-device machine learning for everyone*. URL: https://developers.google.com/mediapipe.

[63] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[64] B Ans, J Hérault, and C Jutten. "Architectures neuromimétiques adaptatives: Détection de primitives". In: *Proceedings of Cognitiva* 85 (1985), pp. 593–597.

[65] HAROLD HOTELLING. "RELATIONS BETWEEN TWO SETS OF VARIATES*". In: *Biometrika* 28.3-4 (Dec. 1936), pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321.

[66] ZHAOHUA WU and NORDEN E. HUANG. "ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD". In: *Advances in Adaptive Data Analysis* 01.01 (2009), pp. 1–41. DOI: 10.1142/S1793536909000047.

[67] Shiqian Chen et al. "Detection of rub-impact fault for rotor-stator systems: A novel method based on adaptive chirp mode decomposition". In: *Journal of Sound and Vibration* 440 (2019), pp. 83–99. ISSN: 0022-460X. DOI: https://doi.org/10.1016/j.jsv.2018.10.010.

[68] María E. Torres et al. "A complete ensemble empirical mode decomposition with adaptive noise". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 4144–4147. DOI: 10.1109/ICASSP.2011.5947265.

[69] *Ethics Committee of the Faculty of Sciences at the University of Lisbon*. URL: https://ciencias.ulisboa.pt/pt/comissao-etica-ciencias.

[70] Vishal Gupta and Vinod Kumar Sharma. "Skin typing: Fitzpatrick grading and others". In: *Clinics in Dermatology* 37.5 (Sept. 2019), pp. 430–436. ISSN: 0738-081X. DOI: 10.1016/j.clindermatol.2019.07.010. URL: http://dx.doi.org/10.1016/j.clindermatol.2019.07.010.

[71] Julien Fatisson, Victor Oswald, and François Lalonde. "Influence Diagram of Physiological and Environmental Factors Affecting Heart Rate Variability: An Extended Literature Overview". In: *Heart International* 11.1 (Jan. 2016), heartint.500023. ISSN: 2036-2579. DOI: 10.5301/heartint.5000232. URL: http://dx.doi.org/10.5301/heartint.5000232.

[72] Google. *MediaPipe Face Mesh*. URL: https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker.

[73] J.A.S.Y. Jayasinghe, Stamos Katsigiannis, and Lakmini Malasinghe. "Comparative Study of Face Tracking Algorithms for Remote Photoplethysmography". In: *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. Vol. 23. IEEE, Nov. 2023, pp. 1–7. DOI: 10.1109/icecet58911.2023.10389182. URL: http://dx.doi.org/10.1109/ICECET58911.2023.10389182.

[74] Ce Liu et al. "Motion magnification". In: *ACM Transactions on Graphics* 24.3 (July 2005), pp. 519–526. ISSN: 1557-7368. DOI: 10.1145/1073204.1073223. URL: http://dx.doi.org/10.1145/1073204.1073223.

[75] OpenCV. *Image Processing in OpenCV - Image Pyramids*. URL: https://docs.opencv.org/4.x/dc/dff/tutorial_py_pyramids.html.

[76] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam". In: *IEEE Transactions on Biomedical Engineering* 58.1 (2011), pp. 7–11. DOI: 10.1109/TBME.2010.2086456.

[77] Hao-Yu Wu et al. *Massachusetts Institute of Technology (MIT), Computer Science and Artificial Intelligence Laboratory (CSAIL) - Eulerian Video Magnification for Revealing Subtle Changes in the World*. URL: https://people.csail.mit.edu/mrub/evm/.

[78] Benjamin W. Nelson et al. "Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research". In: *npj Digital Medicine* 3.1 (June 2020). ISSN: 2398-6352. DOI: 10.1038/s41746-020-0297-4. URL: http://dx.doi.org/10.1038/s41746-020-0297-4.

[79] Philipp Helmer et al. "Accuracy and Systematic Biases of Heart Rate Measurements by Consumer-Grade Fitness Trackers in Postoperative Patients: Prospective Clinical Trial". In: *Journal of Medical Internet Research* 24.12 (Dec. 2022), e42359. ISSN: 1438-8871. DOI: 10.2196/42359. URL: http://dx.doi.org/10.2196/42359.

[80] Benjamin W Nelson and Nicholas B Allen. "Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study". In: *JMIR mHealth and uHealth* 7.3 (Mar. 2019), e10828. ISSN: 2291-5222. DOI: 10.2196/10828. URL: http://dx.doi.org/10.2196/10828.

Appendix A

Appendix

Table A.1: Reported Performance Metrics (RR)

Study	MAE (bpm)	MAPE (%)	RMSE (bpm)
[18]	1.3	-	1.39
[26]	-	0.18	0.2
[27]	-	-	0.32
[31]	0.91	5.74	1.65
[17]	1.727	0.158	2.099
[43]	0.9797	-	1.0693
[51]	1.62	-	1.82

Table A.2: Reported Performance Metrics (HR)

Study	MAE (bpm)	MAPE (%)	RMSE (bpm)
[18]	0.73	-	0.78
[26]	-	0.29	0.31
[27]	-	-	0.32
[28]	8.91	11.53	-
[29]	2.46	3.57	12.36
[30]	4	5.84	4.1
[31]	6.14	7.47	7.81
[32]	5.96	7.54	7.57
[33]	-	16.6	-
[34]	-	Between 2 and 3 (Not exact)	-
[35]	0.2	2.85	2.13
[36]	-	6.19	-
[37]	-	2	-
[17]	2.034	2.49	3.316
[38]	-	6.8	-
[40]	-	-	1.89
[41]	3.75	5	5.09
[42]	0.0006	-	-
[43]	1.69	-	1.75
[44]	1.17	6.55	6.95
[45]	-	1.03	-
[46]	2.63	11.46	9.68
[47]	8.5373	-	-
[48]	-	4.9	-
[49]	-	4.4	-
[50]	0.25	7.59	10.48
[51]	1.38	-	1.66
[52]	2.975	3.808	-
[53]	0.9	-	0.7
[54]	1.35	-	-
[55]	1.74	-	2.32