Facial video based Physiological Parameters estimation in dark environment

Supervised by:
Prof. Fernando Morgado Dias
Prof. Antonio G. Ravelo-Garcia

Committee Members:

Submitted by:
Ankit Gupta
Enrollment: 2118169

Abstract

Physiological parameters estimation plays a significant role in determining the health status of an individual. These parameters are blood pressure, temperature, Heart Rate (HR), oxygen saturation (SpO2), and Breathing Rate (BR). HR and SpO2 have been extensively used for regular medical checkups and scenarios such as surgery, sleep disorders diagnosis, and Intensive Care Units (ICU). The gold-standard technique for measuring HR is the electrocardiogram (ECG), whereas HR and SpO2 can be measured by photoplethysmography. Both are contact based techniques that need sensors or electrodes attached to the skin using adhesive gels.

Furthermore, their placement in proximity to the skin is crucial for accurate estimations. However, these sensors or electrodes may cause discomfort or allergies to the subject and therefore cannot be used in a few scenarios such as sensitive or burnt skin, unobstrusive monitoring. Although ECG is accurate, it is pretty complex for designing portable systems for HR estimation. Therefore, realizing the simplicity of PPG, its non-contact variant was introduced, which is also called iPPG or rPPG. This technology does not need physical contact but can be based on a camera to record videos, followed by blood volume pulse extraction for HR estimation. On the other hand, SpO2 estimations are being carried out by processing the color channels only using the "*ratio of ratios*" (ROR) method. However, almost all estimation methods were developed considering ambient light conditions, which does not resemble the light conditions used in clinical conditions. Moreover, limited studies have conducted parameters estimation under dark environments using infrared channels, which have poor pulsatile strengths. It was found that RGB has comparatively better PPG pulsatile strength than any other imaging modality. Furthermore, conventional methods consider certain impractical assumptions to reduce the effect of prominent artifacts, which may not hold in real-time conditions. Considering these factors, this research aims at designing the non-contact parameters estimation methods in dark environments. Furthermore, the potential of RGB in dark environments would be explored by developing image processing methods for extracting the relevant spatial and temporal information for PPG extraction. It can be done using a two-step approach: 1) the methods will be designed for an ambient environment to reduce the assumptions considered by conventional studies 2) the developed methods would be optimized for estimations under dark environments suitable for monitoring under clinical conditions. Appropriate performance metrics will be identified and used to ensure the developed methods' generalizability, reliability, and clinical relevance. Finally, the performance of all the developed methods will be analyzed to propose the best non-contact HR and SpO2 estimation method for future research.

Keywords: Blind Source Separation, Blood Volume Pulse, Deep Learning, Non-contact approaches, Physiological Parameters Estimation.

Table of Contents

List of Tables

List of Figures

1. Introduction

Physiological parameters are quantitative measures that relate to the physiology of the human body to identify health issues. Applications of investigation using these parameters include disease diagnosis, tracking immediate or long-term effects of surgery or medicinal therapy, early identification of fatal disorders, sleep analysis [1]. Furthermore, activities occurring inside the body such as metabolism, respiration, oxygen levels can be measured using physiological parameters. Five physiological parameters used for determining the individual's health status are Blood Pressure (BP), Heart Rate (HR), oxygen saturation (SpO2), body temperature, and Breathing Rate (BR) [2]. HR and SpO2 have been used in various scenarios like intensive care units, surgery. Furthermore, due to the simplicity and portability of photoplethysmography, HR and SpO2 are the most common vital signs measured by physicians [3, 4].

Heart rate is defined as the number of synchronous contractions and relaxation of the heart while blood pumping. A healthy individual's heart beats between 60 to 100 times per minute (bpm). Heart rate less than 60 or more than 100 bpm are considered abnormalities, termed as bradycardia and Tachycardia, which must be identified by continuous HR monitoring at the earlier stage to avoid situations such as cardiac arrest or sudden death [5]. Furthermore, HR can also reveal valuable information in conditions such as high-level stress, emotion elicitation. Therefore HR estimation has been an active area of research in biology and medicine [6].

On the other hand, SpO2 values indicate the percentage of oxygen bonded hemoglobin in the blood, which is necessary for the survival of body cells or tissues. SpO2 value for healthy individuals ranges between 96-99% and not less than 88% during sleep. It has been used in various scenarios such as intensive care units (ICU), surgery, neonatal care, sleep disorders identification, anesthesia. It also reflects the cardiorespiratory health of the individual. Moreover, continuous monitoring of SpO2 helps in the early detection of hypoxemia conditions.

HR can be measured by electrocardiogram, while pulse oximetry can be used to measure HR and SpO2. Both techniques need sensor's or electrode's contact with the particular body organ's skin using adhesive gels [7]. Monitoring accurate parameters are therefore dependent on the placement of these sensors or electrodes to the correct positions. The common limitation of these techniques is that incorrect positioning of sensors or electrodes may intrude on the accurate estimation. Furthermore, electrodes or sensors may cause allergies or irritation at the binding spot in the scenarios such as prolonged monitoring, sensitive skin. Additionally, this may also disturb the sleep of the subject if used for parameters estimations.

Unlike these gold-standard techniques, the non-contact approach does not need physical contact with the subject's skin. This approach utilizes various imaging modalities for estimating vital physiological parameters. Essentially, this overcomes the limitations of contact based methods for the scenarios mentioned above. Currently, non-contact physiological parameter estimations are being done using the short time interval videos of body organs such as the face, wrist, palm, and chest. A study conducted by van der Kooji and Naber concluded that face is a reliable Region Of Interest (ROI) for this

estimation approach [8]. Existing face based non-contact techniques are based on two conventional principles: Ballistocardiography (BCG) and Photoplethysmography (PPG), followed by frequency or time-based methods for physiological parameters estimations.

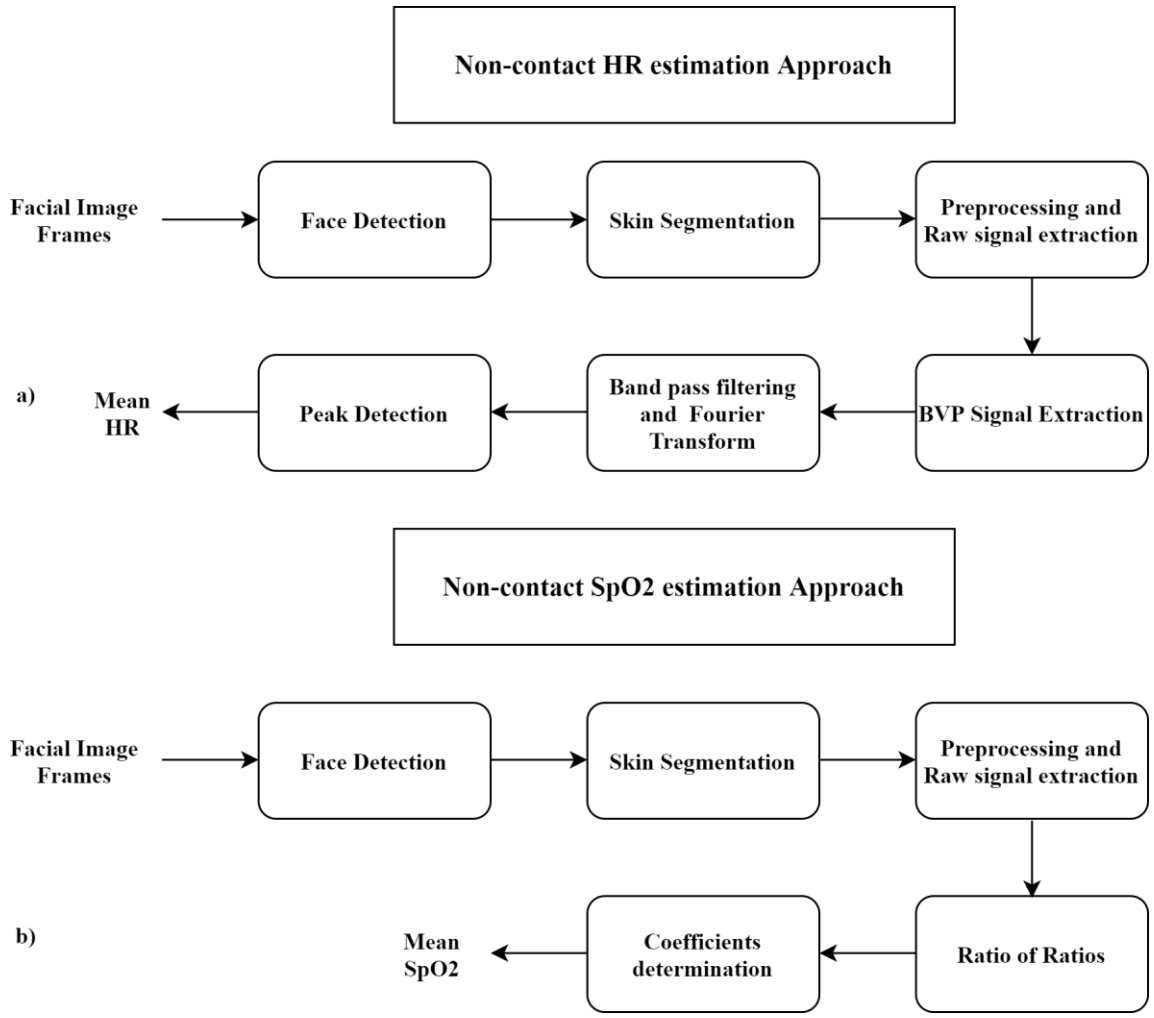

Fig 1. Conventional Non-contact HR and SpO2 estimation approach: a) Heart Rate estimation; b) SpO2 estimation.

BCG tracks the periodic head movement synchronous with the flow of blood from the heart to the head or vice-versa. Although this method works well with the occluded face and is illumination resistant, the method could not perform correctly under voluntary movements, which is the case in real-time scenarios [5]. On the other hand, PPG works on measuring the subtle color changes from the face due to the movement of blood through arteries, which cannot be seen through naked eyes. Non-contact PPG is also called remote PPG (rPPG) [9] or Imaging PPG (iPPG) [10]. The limitation of rPPG or iPPG is that the acquired raw signal is corrupted with noise due to various factors such as motion, uneven illumination variations, and camera quantization. Despite this, several studies have proved that it is possible to reduce noise due to these factors for a cleaner PPG signal followed by an estimation task [11]. Furthermore, this type of approach is

suitable for scenarios such as prolonged monitoring, or sensitive skin. Therefore, this proposal aims at focusing on the non-contact approach. A typical example of the rPPG estimation approaches for HR and SpO2 are shown in Fig 1.

1.1. Motivation

The non-contact estimation approach involves recording the subject's face video in the presence of a light source. Ambient light is predominantly used for physiological parameters estimations, while the other sources are fluorescent, ceiling, and incandescent lights [12]. Furthermore, the majority of the non-contact estimation studies used Red, Green, and Blue (RGB) color channels, whereas only a few studies have used Infrared (IR) channels [13]. The majority of the estimation methods have been tested in a well-controlled laboratory environment, but deploying these methods for clinical settings such as dim light or darkness is not feasible since these methods work well in good lighting conditions. Because accurate estimation needs an appropriate selection of ROI, and poor choice leads to inaccurate values. Secondly, infrared light can be used for dealing with such scenarios, but the pulsatile strength of the corresponding PPG signal is much weaker than RGB videos [14]. Moreover, if the distance is larger between the face and the camera, it will likely have significantly less PPG information, leading to false estimations. The majority of studies have used self-created databases for presenting the effectiveness of their non-contact methods. On the contrary, publicly available databases can also be used, which provides relatively challenging conditions for parameter estimations,

However, limited estimation studies have been conducted in dark environments. These studies have used infrared (IR) spectra for video recording besides relatively weaker PPG pulsatile strength than RGB spectra. Various research developments in the image processing domain provide tools to extract information from the images, which can be recorded using an RGB camera [15]. Despite this, no study has attempted physiological parameters estimations in the dark environment using visible light spectra. Therefore, the unavailability of a suitable dataset and color model for non-contact estimations for HR and SpO2 in the dark is a significant research gap, which this proposal will address. This facilitates the applicability of non-contact approaches in clinical conditions because these conditions include dim or dark light conditions instead of bright light, as used by almost all non-contact estimation studies.

1.2. Research Questions

This proposal aims at addressing the following research questions:

- Is it possible to use RGB videos to estimate physiological parameters in darkness?
- Is it possible to achieve relatively similar performance with the RGB modality as an infra-red modality in a dark scenario?
- Among Deep Learning and Conventional estimation methods, which can estimate physiological parameters accurately?
- Is it feasible to combine image processing methods with state-of-the-art conventional methods to estimate HR and SpO2 in a dark environment?
- Is it possible to use the deep learning models trained on RGB image frames in normal light for HR and SpO2 estimations in darker environments?

1.3. Objectives

The study primarily aims at designing the non-contact based physiological parameter estimation approaches in the darker environment. The estimation method's main component will be PPG extraction, from which physiological parameters will be extracted using frequency or time-domain methods. The methods will be first designed using normal light conditions and tested using publicly available datasets; then, they will be tested for darker environments after improving them using image processing methods. This study also aims at creating a database by recording facial videos without an external light source. The objectives of the study can be summarized as:

1. To review state-of-the-art non-contact methods for HR and SpO2.
2. To create a database by recording the facial videos in a darker environment synchronized with a pulse oximeter for HR and SpO2 ground truth values collection.
3. To identify the best performing conventional and deep learning PPG extraction methods using different publicly available datasets.
4. To develop novel conventional and deep learning based PPG extraction methods under normal lighting conditions followed by implementing them for darker environments.
5. Performance analysis of best deep learning model with the conventional non-contact methods developed during this study.

1.4. Outline

This report consists of 5 chapters and is organized as follows:

Chapter 2 deals with presenting the state-of-the-art methods for non-contact HR and SpO2 estimations using facial videos, followed by a literature summary depicting significant improvements and limitations of the existing methods.

Chapter 3 describes the tentative methodology and workflow to attain the desired objectives by addressing the research questions and answering them.

Chapter 4 presents a tentative timeline to conduct the study.

Chapter 5 presents the results attained so far to attain the desired study objectives.

Chapter 6 describes the conclusion of the study.

2. Literature Review

This proposal aims at designing novel non-contact estimation approaches for calculating HR and SpO2 by extracting the PPG information from facial videos acquired without an external light source. As discussed in the previous chapter, extracting PPG information from facial videos is challenging due to three types of noise sources: motion, illumination variations, and camera quantization. These artifacts are primarily responsible for color distortions in the video, which poses a challenge in PPG information extraction, resulting in poor physiological estimates. Additionally, the amplitude of these artifacts is relatively higher than actual PPG information, which causes PPG/ Blood Volume Pulse (BVP) signal task tedious and challenging. Literature suggested potential solutions to mitigate the effect of these artifacts but needs considerable attention. Existing methods were developed under certain assumptions, which are not always feasible in real-time scenarios. Hence, there is a need to study existing methods for designing reliable estimation methods by emphasizing less on the assumptions considered by the currently existing methods. Therefore, this chapter aims at presenting a comprehensive review of all non-contact HR and SpO2 estimation approaches. This chapter is divided into two sections: the first part presents state-of-the-art HR estimation studies, while the second part corresponds to SpO2 estimation studies. It is important to note that the few estimation studies have reported other parameters, such as breathing rate, eye blink, step count, and heart rate variability. Two studies have reported HR and SpO2 simultaneously, which will be presented in the separate sub-section.

2.1. Heart Rate Estimation

2.1.1. Blind Source Separation

Blind Source Separation (BSS) has been used in HR estimation since Poh et al. [11] utilized it for PPG extraction followed by HR estimation. Blind Source Separation (BSS) techniques separate the PPG signal from raw RGB signals by maximizing the statistical dependence between PPG and noise components. The advantage of using BSS based methods is that they do not need any apriori information for signal extraction. With the evolution of PPG extraction methods, few other variants of BSS methods such as semi-BSS [16] and Joint-BSS [17] have also been proposed to extract cleaner PPG/Blood volume pulse (BVP) signals. This subsection presents state-of-the-art BSS studies for non-contact HR estimations.

Poh et al. [11] extracted the PPG signal using Joint Approximation Diagonalization of Eigen-Matrices (JADE) from the R, G, and B signal traces acquired by processing the facial ROI captured using a webcam. Consequently, three independent components (ICs) were extracted from each color channel, followed by selecting an appropriate IC as a PPG signal for heart rate estimation. The IC corresponding to the Green channel was chosen as the BVP signal. To avoid the artifacts, the historical HR estimates were considered, i.e., if the difference between the heart rate from the previous peak and current peak is greater than 12 bpm, then the current estimated value is rejected. If none of the frequencies met this criterion, the current pulse frequency was considered as HR. The

proposed method did not work well with large movements, although small movements were simulated in the study.

Furthermore, the proposed method could not show its performance under dark or relatively low light conditions. The face detector may halt the detection of the face in case of a high degree of rotations or movement. This study was further extended by adding a temporal filtering component, consisting of de-trending and signal smoothing using a moving average filter for accurate Blood Volume Pulse (BVP) signal extraction. The artifacts were removed using the non-causal of variable threshold (NC-VT) filtering algorithm with a 30% tolerance threshold, finally calculating HR using the mean of InterBeat Intervals (IBI) [18]. The above methods mainly used the green component of the ROI since it was considered to have maximum PPG information.

Additionally, this method used kurtosis optimization, which does not have descent statistical properties to support statistical independence among components. Gill et al. [19] addressed the problem of unsorted ICs of Independent Component Analysis (ICA), which is challenging to select the appropriate independent component as BVP signal. They proposed constrained-ICA, which uses negentropy as an optimization function, avoiding local minima convergence. It is important to note that Negentropy possesses better statistical properties and symmetric decorrelation than kurtosis to ensure statistical independence. With an effort to overcome the illumination variations during video recording in the ambient light, Kado et al. [14] used the combination of Near InfraRed (NIR) and green channel (G-NIR) of the RGB videos, thereby combining the spatial, spectral, and temporal domains. The study aims at recording the subject videos using double CCD having a beam splitter for RGB and NIR sensors without misalignment. Furthermore, the potential of a single-sensor RGB-NIR camera was also assessed. The proposed method splits the video into short windows for motion and illumination artifacts free samples for HR estimation. Subsequently, paired face patches extraction was carried out using a pair of green, NIR, and G-NIR. BVP signal extraction was performed using FastICA, followed by taking the ratio of dominant and second dominant frequencies and its comparison with the threshold for the face patch pair's reliability for the corresponding HR histogram bin voting. Successively, histogram fusion and parabola fitting was performed for maximum peak extraction corresponding to HR frequency. The proposed method was tested under different motion and illumination variation scenarios. The method's performance degrades if there is a large HR spread or head movements for longer durations in the video. Purucuhe et al. [20] analyzed the effect of using different facial region parts on HR estimation and divided the facial region into three ROIs: forehead, eye, and nose surrounding area, and mouth area. The heart rate was computed using each ROI, applying ICA for PPG signal extraction, followed by Fourier transform.

Exploiting the apriori information of the PPG signal, i.e., periodicity, Macwan et al. [16] proposed a semi-blind source separation method named "*multi-objective optimization using Autocorrelation and ICA*" (MAICA), which utilizes a combination of negentropy and signal autocorrelation at different time lags, as an optimization function for extracting the BVP signal. Kalman filter was also used to address motion and illumination artifacts. The bottleneck of this method is the assumption that BVP exhibits the highest periodicity, which does not hold during periodic movements such as rhythmic exercises.

Utilizing the variations in different facial regions for HR estimation, Qi et al. [17] presented the method constituting Connectivity Multiset Canonical Correlation Analysis (C-MCCA) and Joint Blind Source Separation (JBSS) for heart rate estimation. C-MCCA aims to find correlations among the multi-face ROI regions by introducing a Connectivity Design Matrix (CDM) followed by training Max-Margin Multi-Label (M3L) classifier to find correlations by predicting optimal CDM. The CDM would then be used to find maximum correlating regions for feeding the raw signals JBSS, thereby extracting the BVP signal. JBSS used Independent Vector Analysis (IVA) for finding Source Component vectors (SCV) with an assumption that SCVs extracted from different facial regions are correlated, and SCVs from different color channels are uncorrelated. The resultant SCVs were then spectrally clustered based on a similarity matrix using the Normalized cut algorithm. Then, FFT was applied to select the BVP signal by estimating the heart rate from each source signal in the cluster. Subsequently, adaptive peak detection using a regression model was used to detect peaks from the signal for HR estimation. A similar study was conducted by Cheng et al. [21] in which single channel Near Infrared (NIR) videos were utilized for non-contact heart rate estimations using Joint-BSS. The study employed three ROI, cheeks with nose, forehead, and chin. Single-channel signals from three ROIs were multi-channeled using delay coordinate transformation (DCT), fulfilling the prerequisite for JBSS methods which needs multichannel signals for pulsatile component extraction. Independent vector analysis was applied as JBSS resulting in Source Component Vectors (SCV). Mostly SCV1 was utilized for HR estimation, but SCV2 and SCV3 were also tested using 20s and 30s video segments. The proposed method did not consider motion or illumination artifacts. Its performance degraded for rigid motions and lower frame rate. As, DRONZY database have performed relatively poor than MR-NIRP database. This was due to the presence of rigid motions during drowsiness in videos corresponding to DRONZY database.

2.1.2. Color Subspace Transformations

This subsection presents the methods that work on projecting the motion and illumination affected signals to orthogonal projection planes to separate pulse information and other noisy components. These methods are called color subspace transformations because they aim at eliminating the color distortions due to artifacts by transforming the color subspace to other projection planes. Most color subspace transformation methods have used two commonly used color subspace projections proposed by De Haan et al. [22] and Wang et al. [23] as base methods. The studies corresponding to these methods and their variants are presented in this sub-section.

De Haan et al. [22] proposed the first color subspace transformation approach utilizing the chrominance features of R, G, and B spectra, which mitigates the effect of motion on HR estimation. The method extracted the two chrominance vectors, orthogonal to each other, from the RGB color spectra. Specifically, it starts with using the normalized G/R ratio for nullifying the effect of the stationary light source and brightness level, followed by adding white light to eliminate the effect of specular reflection using color difference channels. The method also used skin tone standardization to reduce the impact of different skin tones. RGB to chrominance vector transformation was carried out by empirically known coefficients. Finally, the ratio of the two vectors was used for PPG signal

extraction. Two scenarios were considered to check the efficacy of the proposed method: the first estimation using stationary position and the second with the participants performing gym activities. Under both scenarios, the proposed method after skin standardization performed better than ICA methods. This method assumed that the distortions were due to non-local intensity variation and specular reflection. To mitigate the assumptions during this study, De Haan et al. with different co-workers than the previous study [11] further improved this method by employing the absorption spectra changes of the RGB spectra for BVP signal extraction, where the Hulsbusch noise-free spectrum model was used to develop a normalized BVP vector. The problem with this method is the sensitivity of the BVP vector which, when deviates slightly, may lead to inaccurate estimation results.

The limitation of CHROM is gathering abundant accurate BVP information due to different skin tones. Therefore, Wang et al. [23] proposed a method named "*Plane orthogonal to skin tone*" (POS), which is based on a skin reflection model and consists of two steps: removing the intensity variations; and specular components elimination. The model starts with intensity variation elimination by projecting the temporarily normalized RGB traces to a projection axis, consisting of a vector of ones and a projection matrix. This projection matrix was found by calculating the blood volume pulse vector for each color channel that maximizes the pulsatile strength after projection. The projected signals will be calculated by projecting the raw RGB traces using a projection plane orthogonal to the temporarily normalized skin tone, with one vector corresponding to the pulse signal while the other corresponds to the specular reflection.

Furthermore, the projected signals will always have in-phase pulsatile components and antiphase specular components. The limitation of this method is that its performance degrades in heterogenous illuminance conditions. Additionally, all experiments done in this study have used a single source of light (frontal fluorescent lamp) at one time and a fixed camera source, which is impractical for real-time applications. Finally, if the amplitude of specular and pulsatile components are the same, the method might not perform well.

However, the dimensionality of motion was not discussed in CHROM or POS. Therefore, Wang et al.[24] addressed the limitation of removing a restricted number of motion distortions from three-channel RGB videos. Specifically, the proposed work addresses the scenario when the number of distortions is more than the number of color channels and argues that in any case, the number of motion distortions eliminated is less than the number of color channels. To resolve this problem, the RGB signal traces were bandpass filtered and separated into orthogonal frequency bands, followed by suppressing the effect of motion from every band and finally extracting the cleaner PPG signal. First, the RGB traces were temporally normalized to eliminate the DC component. Successively, Discrete Fourier Transform (DFT) was used to convert the time domain signal to the frequency domain, dividing it into orthogonal frequency bands and selecting the frequency components within the human heart rate range. Using inverse DFT, these components corresponding to each frequency band were converted back to the time domain signals (RGB signal traces). The POS method was used for each component to extract the BVP signal, followed by concatenation using a weighting scheme. The method was tested using various scenarios and concluded that it worked well for all skin tones,

single light sources, and longer processing windows. However, the method would not work if heart rate frequency lies near motion artifacts frequency and under low-intensity light conditions for darker skin tones due to less light penetration. Furthermore, increasing the window will create problems with latency; therefore, a processing window of 128 frames@20 fps was suggested.

Furthermore, another problem with CHROM and POS is their fixed projection planes which were vanquished by the adaptive pulsatile plane (APP) method proposed by Tran et al. [25]. This work also eliminates the fixed ROI selection and the inability to deal with skin tones using fixed thresholds. Facial skin detection was performed using Deeplabv3+ trained on Celeb HQ dataset, followed by testing the performance of different facial tracking algorithms. It was empirically found that Moose facial tracker worked well with HR estimation. Subsequently, projection vectors for PPG signal extraction were extracted using singular vector decomposition (SVD) followed by translation for projecting the RGB plane to the new transformation plane, with mean of the RGB values as origin, followed by two rotations by keeping the R channel vector fixed, resulting in APP plane. The red channel was chosen due to the highest specular reflection, followed by green and blue channels. Therefore, the APP plane was further rotated to 360° followed by 8^{th} order Butterworth filtering to identify the region with the best pulsatile strength. Consequently, the best pulsatile areas are found within 0-100° with 90% accuracy with signal S containing (S1 and S2) corresponding to pulsatile and motion components, which were further optimized using alpha tuning. Finally, the gradient-based peak detection method calculates interbeat intervals and averages them for heart rate calculation.

2.1.3. Empirical Mode Decomposition based methods

Empirical mode decomposition is a wavelet processing method that decomposes the signal into intrinsic mode functions of varying frequency and amplitude. However, it suffers from a mode aliasing problem [26]. Hence, Ensemble empirical mode decomposition for PPG extraction was introduced by Chen et al. [27], which proposed an HR estimation method by addressing the problem of facial illumination variations. The method used reflectance decomposition, which was based on weber law for discriminating the reflectance from illumination. The resultant signal traces were then applied Ensemble Empirical Mode Decomposition (EEMD) followed by EMD, for splitting it to Intrinsic Mode Functions (IMFs). Out of 10 IMFs, IMF4 was used for heart rate detection using a peak location algorithm. The reason behind using EMD after EEMD is that the latter did not give real IMF due to averaging and summation of noise to each IMF. The termination criteria used for EEMD was the predefined standard deviation threshold, while for EMD, it was S-number which is the number of consecutive iterations until zero crossing, and extrema number is equal or differ by 1. Lin, Chen, and Tsai [7] proposed a similar approach which deals with using the reflectance signal from the green channel of the RGB videos. The proposed method separates the reflectance from the illuminance based on weber law, and the Alternating Direction Method of Multipliers (ADMMs) based optimization algorithm. The resultant reflectance signal was then applied Hilbert-Huang Transform [27]. BVP signal extraction was performed by applying EEMD followed by EMD, thereby extracting two IMFs: IMF3 and IMF4. The termination criteria used for EEMD was standard deviation less than threshold and S-

number, i.e., 10. Heart rate estimation was performed using multiple linear regression. The bottleneck of this method is the extraction of a reflectance signal due to distortions in facial appearance during recording. Therefore, it is infeasible to implement this algorithm under real-time conditions. Although the studies mentioned above have successfully utilized EEMD for BVP signal extraction, signal integrity is a primary concern while using EEMD, i.e., it may halt the signal integrity. Therefore Yue et al. [26] proposed a rigid motion resistant HR estimation method combining Complementary ensemble empirical mode decomposition (CEEDMAN) and permutation entropy. This work addressed the face shake or head movement problem and used a subspace rotation algorithm (Affine Transformation) to deal with it. The raw signal was extracted using spatial correlation followed by non-negative matrix decomposition. Furthermore, the first Eigenvector from the decomposition i.e. skin vector was projected to an orthogonal projection plane subsequent to alpha tuning, for initial BVP extraction. Furthermore, BVP was denoised using CEEDMAN algorithm with permutation entropy, as an objective function with an upper and lower mode threshold as 0.3 and 0.7. Permutation entropy greater than 0.7, were further decomposed using wavelets. Finally, the denoised BVP signal was used for heart rate estimation. The problem with this method is the use of manual parameters set empirically, which may not work for all.

Following the approach of using multiple ROIs, Song et al. [28] proposed a combination of EEMD with multiset canonical correlation analysis. EEMD was used to extract the pulse signal, while MCCA aimed at maximizing the correlation among resultant signals to make them free from noise. The proposed method split the ROI into multiple squared ROIs followed by calculating light intensity and variations as mean and standard deviation using the L channel of LAB color space. The light intensities were sorted in ascending order while the variations were sorted in descending order, subsequently to the calculation of SNR of the green channel for the top 10 patches, followed by using high-quality patches for further processing. A two-step procedure was performed to deal with outliers during HR estimation: identifying HR outlier and replacing it with valid HR. First, the whole video was divided into seven segments of 30 seconds, each with an overlap of 5 seconds. For each segment, absolute HR differences were calculated with the other segments, followed by calculating the frequencies of minimum differences by comparing them with a threshold. If the frequency for a particular segment was greater than another chosen threshold, the corresponding HR candidate was considered reference HR. Furthermore, the HR outliers were removed by following the same approach and replacing the outlier with the new HR candidate closest to reference HR.

Finally, to suppress the effect of motion for HR estimation, Zhang et al. [29] presented an approach, which uses LAB color space for HR estimation. Specifically, the study uses the A and B channels of the LAB color space followed by a smoothness prior approach for stationary artifacts elimination. Then the difference of A and B channels is bandpass filtered, which was considered as raw signal. BVP signal extraction was performed using EEMD. The potential IMFs were selected based on peak and power ratios.

2.1.4. Neural Networks based HR Estimation

Neural Networks, specifically deep learning technology, have shown immense potential in predicting the HR or BVP from shorter facial videos. Unlike conventional methods,

the methods employing these networks do not need assumptions and possess good generalization ability for the samples [9]. Literature suggested two types of deep learning approaches for HR estimation. The first approach predicts the BVP waveform followed by signal processing for HR estimation, while the second predicts the HR values directly from the face videos. Furthermore, HR estimations using the second approach were predominantly performed using face video to HR value prediction as a regression task, while few studies have considered it a classification task. Most of the studies employed conventional computer vision deep learning architectures utilizing transfer learning, while few have proposed their architectures for HR estimation tasks.

Yu et al. [30] proposed a deep learning framework to estimate HR from compressed videos. The proposed framework consists of a SpatioTemporal Video Enhancement Framework (STVEN) and a deep learning architecture rPPGNet. STVEN was used for video enhancement, whereas PPG signal extraction was accomplished using rPPGNet. STVEN consisted of two downsampling convolutional layers followed by four spatiotemporal blocks and two upsampling convolutional layers. Furthermore, STVEN consisted of two loss functions: reconstruction loss and cycle loss. The rPPGNet consists of three modules spatiotemporal blocks, skin segmentation, and partition constraints. The Spatio-Temporal (ST) block calculates relevant features, covering both tasks of projecting the RGB color space to other color spaces for rPPG extraction and temporal normalization for irrelevant information elimination. The loss function for this module is the combination of negative Pearson correlation coefficients from the middle spatiotemporal block and at the network's end. The third component, the partition constraints module, was used to learn specific PPG features. rPPGNet was trained on high-quality images, and compressed videos of different bitrates were used to train STVEN separately. Full architecture STVEN followed by rPPGNet training was performed using the weighted sum of rPPGNet, STVEN, and perceptual loss. The framework achieved the highest accuracy with MPEG-4 and x265 compression formats. However, the study did not use compression related metrics, which could investigate the effect of compression on HR estimations.

Qiu et al. [9] proposed EVM-CNN for HR estimation. As the name suggests, the proposed method constitutes Euler Video Magnification (EVM) and Convolutional Neural Networks (CNN). EVM aimed at removing the irrelevant information by spatial decomposition using Gaussian pyramid decomposition and temporal filtering followed by FFT. The resultant feature images extracted using EVM were fed to 15 layered CNN, consisting of a convolution layer, five sets of depthwise convolution followed by pointwise convolutions, one average pooling layer, one dropout layer, and two fully connected layers. Depthwise convolutions were used for fast execution time, while pointwise convolutions combined the output of depthwise convolutions. Average pooling, dropout, and fully connected layers have performed their usual functions. The effect of HR diversity was not analyzed in this work.

Treating HR estimation as a classification task, Wu et al. [31] proposed an artificial neural network ensemble consisting of 31 models with each model predicting the different HR range intervals such as 5, 10, and greater than 30 bpm. The HR prediction was carried out in four stages. Furthermore, the problem of high degree motion affected videos in which motion spectra resemble the pulse spectra were also addressed. To resolve this problem,

a method was proposed consisting of the harmonic product spectrum and power spectrum entropy. The framework developed during this study was tested under driving conditions in nine scenarios: Sunny, Overcast, Night, Rainy day and Night, driving in the city, driving on the highway, and Sunset. The proposed method is very sophisticated to use during the driving scenario, hence needs significant improvements in terms of faster executions. Moreover, suitable thresholding methods could further increase the proposed method's estimation accuracy. The study conducted by Bousefsaf, Pruski, and Maaoui [32] presented a pilot study depicting the potential of CNN for Heart rate estimation. Solving the problem of limited labeled samples, a synthetic pulse waveform generator method is proposed and trained with a CNN having one convolution, one max-pooling, and two dense layers with labels as heart rates ranging from 55-240 bpm with a difference of 2.5 bpm. The study treated the HR estimation problem as a classification task with 75 heart rate ranges and 1 "No-PPG" class. The model was tested with the facial videos taking the first 60 green color channel image frames of the video with each frame of size 25×25 pixels with one pixel overlap. In other words, the method was tested using volumes of 25×25×60 with a spatial overlap of one pixel. The final pulse rate is calculated by taking the weighted average of pulse rates over the whole HR range (55-240bpm). As mentioned before, the study proposed a synthetic data generator to solve the problem of limited labeling samples. The proposed synthetic data generator consisted of 5 steps: pulse waveform modeling from PPG signal using seven mathematical models followed by their selection using data fitting and Fourier series; two second signal formation from waveform using Fourier model; Adding linear, cubic or, quadratic trends; video creation by signal repetition; and addition of random noise to image frames. The study did not consider the effect of motion and color channels for the estimation task. Furthermore, no effort has been done to improve the accuracy of the proposed CNN.

Exploring the potential of conventional deep learning architecture, Niu et al. [33] presented a transfer learning based framework named RhythmNet, which utilized YUV color space based spatiotemporal representation from 25 ROIs to feed to deep learning architecture for HR estimation. ResNet-18 was used as a backbone network with GRU for temporal modeling. The study considered the larger head movement, different types of devices, and varied illumination conditions. This also considers analysis with video compression formats. Furthermore, a benchmark database VIPL-HR was also proposed considering nine different conditions and three types of cameras with 3130 videos from 107 subjects. The proposed architecture was trained on RGB videos, but its performance was also tested for NIR videos. Although the model did not perform well since it was trained on RGB videos only, the proposed model also showed its potential for HR estimation using NIR videos. Furthermore, it was revealed that all kinds of movements or illumination variations significantly affect the estimation accuracy. Another transfer learning approach proposed by Hsu et al. [34] used a time-frequency representation as input and predicted mean HR. The study also proposed customized deep learning based facial landmark selection method, which consisted of a Single Shot multi Detector (SSD) and Double Conv-Dropout (DCD) framework. SSD utilized multiple bounding boxes of different aspect ratios and scales and calculating a score depending on the object's presence within the bounding box. Subsequently, regression was used for mapping the bounding box to the bounding box of the ground truth object. DCD is a 15-layered network consisting of seven convolution layers, two pooling layers, two dropout layers

and, three fully connected layers. The resultant ROI will be applied three-stage filtering for illumination rectification, detrending and, signal amplification. Finally, the signal is converted to its 2D time-frequency representation for feeding to the Visual Geometry Group (VGG)-16 network. This work also proposed a benchmark database named Pulse from Face videos (PFF). The method did not consider motion artifacts and complex backgrounds for skin and no-skin pixels classification.

Yu et al. [35] proposed a deep learning framework AutoHR with neural architecture search (NAS), which depicts the self-exploring ability of neural networks. The proposed method has three components: gradient-based NAS methods for backbone search using temporal difference convolution to gather relevant PPG information; a loss function comprising time and frequency domain constraints; and spatiotemporal data augmentation. Backbone search deals with finding the optimal neural network architecture for HR measurement using Temporal Difference Convolution TDC, while two spatiotemporal data augmentation techniques used in this work were for better representation and increasing labeled samples. However, the second data augmentation strategy was dependent on HR values.

Song et al. [36] utilized conventional methods for PPG extraction along with the deep learning method. The proposed method includes feature images construction from BVP or ECG using Akima cubic Hermite interpolation. Furthermore, real PPG image features were constructed from CHROM methods followed by creating a spatiotemporal map through time-delayed way using Toeplitz representation. ResNet18 was first trained using synthetic images following the transfer learning approach, followed by training refinement using the feature images constructed from the facial video. The study has shown that color images, samples balancing, pre-training, and CHROM signals for creating synthetic feature images enhanced the HR estimation accuracy.

Exploiting the deep learning's attention mechanism for cleaner and accurate PPG/BVP extraction, Hu et al. [37] pointed out the problem of redundant spatial information in the facial videos, head movement and proposed a spatial-temporal attention network for HR estimation. Specifically, the proposed method has the following components: spatial-temporal facial feature extraction, BVP signal construction, and processing. Spatial-temporal facial features were extracted using 2DCNN and 3DCNN, then aggregated using adaptive average and max pooling functions. BVP signal construction was done using spatial-temporal convolutions consisting of cascading convolution filters. Successively spatial-temporal strip pooling dealt with head motions and accurate PPG information extraction from multiple ROIs. Furthermore, the spatial-temporal attention mechanism was used to extract the information from the long video sequences along with extracting the vital local and BVP information with no other physiological signal information,i.e., PPG signal processing. This component consists of spatial-temporal strip pooling for eliminating head motions, along with a spatial-temporal attention module. This attention module consists of global average pooling to avoid extreme values, max pooling for texture features preservation. The method's performance degraded under a higher degree of compression and non-rigid motions like talking or rigid movements like rotations and fast translations. As per their study, the adaptive average aggregation function performed better than the adaptive max pooling function. Hu et al. [38] followed the same approach and proposed an effective time domain-based attention deep learning framework (ETA-

rPPGNet) that addressed three problems and provided three components of the framework. The first problem dealt with the spatial redundancy due to dense image frames in a video; this problem was solved by dividing the video into 50 segments and feeding each segment into time-domain subnet segments. The time domain subnet segments include a three-step process: gathering the vital spatial features, assigning weights followed by independent updating of the corresponding weights using attention mechanism and, aggregation using adaptive average pooling. This component uses depthwise convolution, average pooling, and pointwise convolution with one filter, followed by a softmax gating scheme to create the attention map, which was used to create feature maps. The second problem was to remove the motion and illumination artifacts from the rPPG signal. Hence the result of the time domain subnet segment is passed to the convolution filters based backbone net for rPPG signal extraction. Global average pooling was performed on the constructed rPPG signal for spatial information aggregation. A time-domain attention model was used to remove artifacts, which utilizes 1-D convolutions to extract the local temporal changes without focusing on global temporal changes and using a self-attention mechanism by considering *k* members for each feature value at a time. The third problem was to only extract the rPPG signal, for which a two-part loss function was used. Finally, the resultant signal from ETA-rPPGNet was bandpass filtered and used for HR prediction using power spectrum density (PSD) analysis. The proposed framework did not cover all the HR ranges due to a lack of training samples. It did not work well for shorter videos of length < 4 secs. Selecting the number of frames in a segment for the time domain subnet is tricky and critical, as the lower number of frames may not filter redundant information, while a higher number of frames may eliminate vital rPPG information considering the redundancy; therefore, N=50 was used. The degraded performance of the framework was seen in compressed videos the uncompressed video.

2.1.5. Color channels using other color channels along with RGB

Realizing the need to add more channels for accurate BVP extraction, McDuff et al. [37] used a five-band lens camera to extract the orange and cyan spectra along with three traditional color spectra. It enables monitoring the absorption of light differences between Hb and HbO_2 by creating a bigger overlap between cyan, orange, and green spectra, for accurate heart rate estimation using ICA. Most importantly, this approach has shown the ability to estimate HR at larger distances, i.e., 3m using green, cyan, and orange channel.

2.1.6. Hybrid methods

Hybrid methods include those methods which utilize the combination of the methods as mentioned above for HR estimations. For instance, the study conducted by Cheng et al. [10] proposed an illumination variant HR estimation algorithm consisting of JBSS and EEMD. Two types of illumination, i.e., artificial and natural, were addressed in this work. The proposed work used JBSS for illumination extraction, while EEMD was applied to the raw signal from the green color channel. Specifically, independent vector analysis was used by keeping the order consistent between background and facial ROI. Following EEMD was used to decompose the resultant signal into IMFs. Subsequently, fifth IMF was used for BVP signal extraction and heart rate estimation. The method assumed that

illumination sources are the same for background and facial ROI, which is an impractical assumption for real-time scenarios. Secondly, the method did not address motion artifacts.

A similar approach with a different method combination was proposed by Zhao et al. [39], which aimed to provide a strategy to choose initialization parameters for pulse wave extraction and presented an application and extension of POS or skin reflection model. The proposed method analyses the probability distribution of model parameters in RGB space and assigns the distribution centers as model parameters, then utilizes the least mean square adaptive filter for optimization. The signal resulting from spatial averaging of ROI pixel values was subtracted from the mean to reduce the motion artifacts. To further reduce the motion artifact residuals, singular spectrum analysis (SSA) was applied, which aimed at removing the slowly varying trends followed by reconstructing the oscillatory components. The detrended traces were projected to RGB space to generate distribution centers for initial signal extraction, which was further optimized using PCA to extract motion and pulse vectors separately. Subsequently, alpha tuning was used for optimization, followed by the least mean square adaptive filter (motion artifact removal). Heart rate estimation was performed using a Gaussian distribution based weighting function for stabilization, i.e., assigning higher weight for lower deviation and lower weight for higher deviation, from the mean HR. Finally, a prototype, which controls the speed and incline of the Treadmill, was proposed. This was an automatic system dependent on the heart rate ranges set for fitness exercise. The proposed algorithm worked well in high motion situations and with compressed videos. The work also proved that distribution centers of motion and pulsatile vector formed a slightly overlapped cone, making these motion and pulsatile vectors the starting point for initial pulse extraction.

Another approach utilizing the SSA was proposed by Ryu et al. [40], which provided an illumination variation resistance for the HR estimation method. The method consists of two components: sub-band method and Singular Spectrum Analysis (SSA). The BVP signal was extracted using the sub-band method, which starts with converting the raw RGB signals to frequency domain signals using continuous wavelet transforms and applying POS to each frequency sub-band. On the other hand, singular spectrum analysis is an SVD based subspace decomposition method, which aims at separating the unrelated frequency components. In other words, SSA separated the frequency components of illumination variations from PPG information. The study did not address motion artifacts, but the authors claimed that since the POS method was used in sub-band analysis, it could tolerate limited head movements. However, complex rigid movements may degrade its performance.

Combining chrominance-based signals and ICA's advantages, Song et al. [41] extracted chrominance signals and applied Kernel Density ICA (KDICA) to them for BVP signal extraction. The kernel density ICA was used to address the problem of similar magnitude among illumination variation and PPG signal. In addition, the authors have also tested the effect of different shooting distances and image resolution for PPG signal extraction. The shooting distances of more than 1 meter would result in PPG information loss, leading to inaccurate estimations. On the other hand, high camera resolution enhances the rPPG signal quality and increases the execution time.

2.1.7. Miscellaneous

Almost all the HR estimation methods fall into the following categories: BSS, color subspace transformations, neural network based, or combination of two or more types. However, few have employed slightly different methods, which do not fall within the above categories. As an instance, Li et al. [42] proposed an HR estimation method that included the following components: Discriminative Response Map Fitting (DRMF), Discriminative Regularized Level Set Evolution (DRLSE), and nonlinear least square adaptive filter. DRMF was used for collecting the facial landmarks (66) and used only 9 points covering the face, excluding the forehead and eyes. The tracking process was carried out using "good features to track" [43], corresponding to ROI and tracking them using the KLT algorithm. This step eliminates the effect of rigid movement. Illumination rectification was performed by collecting the background region using Discriminative Regularized Level Set Evolution (DRLSE), extracting the pulse signal, and subsequently removing background noise. Furthermore, a nonlinear least mean square adaptive filter was used to estimate the motion suppressed pulse signal. Non-rigid movements were further removed by dividing the signal into *m* segments and calculating standard deviation, eliminating 5% of the signal possessing the highest standard deviation. Finally, a detrending filter, moving average filter, and bandpass filter followed by spectral power estimation using the Welch method was used to get the desired HR frequency. The proposed method did not work well with high degrees of head rotation, specifically more than 60°, due to the non-tracking of ROI features.

John Krishna and Galigekere [44] proposed an HR estimation algorithm that automatically selected the background and source ROI. The starting point for source ROI was the bounding box provided by the Viola-Jones detector, while the area close to the face was selected as the background ROI. Successively, the skin pixels were chosen using the YCbCr color model in which the ratio of Cb and Cr was used for skin pixels identification. Row and local variances were further used to refine the source ROI, whereas background ROI was selected as 45% of the face height. The green channel from source and background ROI was extracted, and normalized mean square adaptive filtering was applied for noise suppression. Finally, the resultant signal was transformed using Hilbert transform followed by power spectrum density (PSD) analysis for extracting the cardiac frequency. This work did not address the motion artifacts. Additionally, any strategy to deal with complex backgrounds was not addressed and eliminated.

An approach combining the Euler technique and Bayesian theorem for HR estimation was presented by Gupta, Bhomick, and Pal [45] in which the importance of removing artifacts from facial expressions and out-of-plane deformations was addressed for HR estimation. The proposed Monitoring using Modelling and Bayesian Tracking (MOMBAT) was divided into three stages: window extraction, window analysis, and HR tracking. Window extraction includes dividing the ROI into multiple overlapping windows. Window analysis deals with eliminating in-plane face movements from each window. Subsequently, BVP signal extraction was done using the Eulerian technique, which deals with extracting the eulerian temporal signals from the green color channel of each ROI, followed by bandpass filtering. BVP signal was then extracted using these signals with the approach given by [14]. Furthermore, out-of-plane movements were detected using a constrained local neural field from which 3D facial points near face

boundary were used to identify out-of-plane movements. Successively, the pulse signal was reconstructed by pulse modeling for the frames affected by the out-of-plane movements using Fourier basis based modeling. Finally, the Bayesian framework was used for removing false HR estimates, which utilized the prior information and likelihood information from the current window. MOMBAT does not work well if distortions persist for longer durations and compressed videos.

Woyczyk, Fleischhauer, and Zaunseder [46] highlighted the importance of ROI selection and proposed Gaussian Mixture Models (GMM) with level sets. This work extends the work by Chen and Vese in terms of using a distance of skin and non-skin pixels followed by normalization. In the proposed method, GMM was trained to model background and foreground pixels using the starting frame of the video. Furthermore, the trained model assigned each pixel foreground (skin) and background (other than skin) probabilities. These probabilities were calculated using an expectation-maximization algorithm. Finally, ROI selection was performed using these probabilities with additional constraints using a level set. The level set was defined by contour, approximated by image data, size, and shape of the contour. A contour defined by continuous energy function was intersected by a threshold plane to assign the skin pixel outside or inside the contour. The optimization of the contour was done to classify the skin and non-skin pixels for ROI selection. BVP extraction was performed using CHROM, POS, and green. The method failed to perform well for MPEG-4 compressed video due to rhythmic adaptation of ROI originating from forward and backward image data prediction. Secondly, the decision parameter for controlling the number of skin kernels was done empirically without proposing an estimation algorithm. A complete summary of HR estimation studies is depicted in table 1.

Table 1. A summary of HR estimation studies.

First Author (year)	Parameters	ROI Used	Method	Color channel	Limitations
M Poh [11], (2010)	HR	Face	ICA	RGB	The method would not work well under rigid movements and different illumination conditions.
G Tsouri [19], (2012)	PR	Face	ICA	RGB	The constrained ICA is 30 times slower than ICA.
G Haan [22], (2013)	HR	Face	CHROM	RGB	CHROM method uses skin standardization and fixed projection planes, which halts its generalizability.
X Li [42], (2014)	HR	Cheeks, Nose, Mouth	DRMF, DRLSE, NLMS	Green	Inaccurate ROI tracking due to high degrees of motion may degrade the method's performance.
D Chen [27], (2015)	HR	Forehead	EEMD	Green	The performance may degrade for longer interval videos.
K Lin [7], (2016)	HR	Forehead	EEMD, MR	Green	Extraction of reflectance signal was performed using the assumption of uniform light on each face region, which is impractical for real-time conditions.
W Wang [23], (2016)	HR	Face	POS	RGB	It does not work well with illumination variations and has fixed projection planes for pulsatile component extraction.

J Cheng [10] (2017)	HR	Cheeks	EEMD	Green	Motion artifact not addressed; The assumption used for the study that illumination sources are the same for face and background ROIs is impractical to consider in real-time scenarios
H Qi [17], (2017)	HR	Face	J-BSS, C-MCCA	RGB	The method did not address the motion and illumination artifacts.
W Wang [24], (2017)	HR	Face	Subband separation	RGB	If the cardiac frequency lies in the motion spectrum and dark skin subjects, the method will not work well.
Y Lin [47], (2018)	PR	Nose, Cheek	Adaptive filtering	RGB	ROI recapturing due to motion hindered the signal, which degrades the estimation accuracy.
Q. Tran [25], (2019)	HR	Face	APP	RGB	The method works poorly if heart rate frequency lies in the motion spectra.
B Wu [31], (2019)	HR	Cheeks	NN	RGB	The method is very complex, using 31 ANN models for HR prediction, which makes the method very slow.
Y Qiu [9], (2019)	HR	Cheeks	CNN, EVM	RGB	Heart rate distribution was not focused for this method.
R Macwan [16], (2019)	HR	Face	ICA	RGB	The proposed method uses periodicity as one of the criteria for BVP selection, which limits its applicability for estimation during periodic movements.
Bousefsaf [32], (2019)	HR	Forehead Cheeks	CNN	None	The proposed network was not fine-tuned for better estimation results.
X Niu [33], (2019)	HR	Face	CNN, GRU	YUV	RhythmNet did not work well on NIR videos.
C Zhang [48], (2017)	HR,	Eyes Area	SOBI	RGB	The method limits its applicability for higher degrees of freedom.
H Yu [26] (2019)	HR	Face	SSR, CEEDMAN	RGB	The method uses many thresholds that were set empirically; hence the method may not generalize well.
C Zhao [39], (2019)	HR	Nose, Cheeks	POS	RGB	Although the method reported the lowest standard deviation, the standard deviation reported in the study is too much for HR estimation. This makes the algorithm unstable.
R Song [41], (2020)	HR	Cheeks	KDICA	RGB	This study analyses the effect of resolution using KDICA did not propose their algorithm. Hence the corresponding limitations could not be realized.
J Ryu [40], (2020)	HR	Cheeks	CWT	YCbCr	The proposed method can not work well for rigid motions.
G Hsu [34], (2020)	HR	Face	CNN	Green	The study excluded motion artifacts.
Z Yu [35], (2020)	HR	Face	CNN	RGB	The data augmentation strategies were dependent on the HR values, which halts the method's generalizability.
R Song [36], (2020)	HR	Cheeks	CNN, CHROM	RGB	The method used a tiny processing window to mitigate the effect of artifacts; persistence of artifacts for a longer duration may degrade its performance.

Y Zhang [29] (2020)	HR	Face	EEMD	AB	The proposed method's performance degrades for shorter video sequences.
S Kado [14], (2020)	HR	Cheeks	FastICA	RGB, NIR	The method's performance will degrade in case of larger HR spreads and longer interval of motion or illumination variations.
R Song [28], (2020)	HR	Face	EEMD	RGB	The method's performance degrades for inaccurate ROI tracking or limited canonical variables in MCCV.
P Gupta [45], (2020)	HR	Face	NR	Green	MOMBAT did not perform well for compressed video and distortions for longer durations.
J John [44], (2020)	HR	Forehead	Cust-VJ	Green	Since ROI selection was carried out by discriminating the background with facial ROI, complex backgrounds were not considered; and motion artifacts were not addressed.
A Woyczyk [46] , (2021)	HR	Face	GMM	RGB	The proposed method did not perform well with compression videos; manual selection of decision parameters also halts its generalizability.
J Cheng [21], (2021)	HR	Face	IVA	NIR	The study did not consider motion and illumination artifacts.
M Hu [37], (2021)	HR	Face	CNN	RGB	Performance degradation was reported for rigid motions such as rotation and fast translations, and compressed videos.
M Hu [38], (2021)	HR	Face	CNN	RGB	The proposed framework could not be able to test for diverse HR ranges due to limited labeled samples.

2.1.8. HR and other parameters estimations

Few studies have estimated other vital signs along with HR such as BR, HRV, SpO2, eye blink, step count. HR and BR combination have been predominantly estimated in the literature, while few non-contact estimation studies also include SpO2, eye blink, step count. A comprehensive review of all such studies is presented in this subsection.

2.1.8.1 Breathing Rate and Heart Rate

Verkruysse et al. [49] demonstrated the extraction of multiple parameters under normal light conditions. The study used the PPG signal extracted using the green color channel of the ROI selected from the face video acquired by a digital camera. PPG signal was then processed using filtering techniques, subsequently heart, and respiratory rate calculation, respectively. Bousefsaf, Maaoui, and Pruski [50] proposed a Continuous Wavelet Transform (CWT) based method for instantaneous heart and breathing rate estimation. The method starts with detecting the face, followed by using Pan (P), Tilt (T), and Zoom (Z) parameters. PTZ parameters were used to convert the 320×240 pixel frames to HD frames preserving the framerate. Subsequently, RGB color space was converted to L* u* and v* using CIE XYZ space, followed by using u* for the detected skin pixels to eliminate the motion artifacts from the signal. The pulse signal was constructed using

continuous wavelet transform in which the raw signal after removing the DC component was convolved with the child wavelet to get the mother wavelet. CWT coefficient was filtered using wavelet energy curve followed by using them to reconstruct the PPG signal using inverse-DWT. Finally, the signal was interpolated to 256Hz using cubic spline interpolation, followed by peak detection and interbeat intervals calculation. The respiration signal is calculated by linearly interpolating the PPG signal to 30Hz, then identifying the respiratory cycles and subsequent respiratory interbeat intervals calculation. The performance of the proposed method was tested using two conditions stationary and motion (vertical and horizontal head rotation). Also, the breathing signal was deduced from the HR signal.

Another study estimation combination of Heart rate and breathing rate was proposed by Tarassenko et al. [51] in which autoregression models were utilized, based on the principle that the current value is calculated based on previous p-values with white Gaussian noise. Subsequently, time to frequency domain transformation was performed using z-transform, which includes extending a point z to complex planes of each pole. Poles were removed analyzing the background region, and the same poles were tracked in each ROI source. The poles corresponding to aliasing frequency components were removed, thereby extracting the relevant pole. Then the angle was calculated with a mathematical formula for cardiac frequency calculation. The processing window used for HR is 15 seconds and ROI of size 100×100 pixels. The BR signal was processed using 7^{th} order autoregressive model, followed by downsampling (2 Hz) to increase the angular resolution. Furthermore, the poles were identified as 95% of high magnitude poles, followed by using the lowest angle for breathing rate calculation. The processing window for breathing rate is 30 seconds and an ROI of size 25×25 pixels. SpO2 was calculated using ROR, but the method faced issues predicting the SpO2 range from 80-95%. Since the AC component amplitude was very near to camera quantization noise, a high-resolution camera would be required for accurate SpO2 estimations. Results have shown that the spectral distribution of HR and BR was done by dividing the whole ROI into 25×25 regions showing the angle of the pole and its radius as magnitude. The results have shown that for HR, the spatial distribution of the pole was on the entire face except for the area nearby the eyes and nose, whereas, for BR, it was mainly at the upper thorax region with a bit on the forehead. Furthermore, the study proved that the camera based BR estimation is more accurate and consistent than the chest sensor for BR estimation. The camera based PPG form contains cardio-synchronous color changes with blood volume pulsations.

Wei et al. [52] presented a second-order blind identification algorithm to estimate heart and respiratory rate using manually selected mouth and neck ROI. The mouth was used for the heart rate estimation, whereas the respiratory rate was calculated using the neck/throat region, resulting in six signals. Subsequently, Blind Source Separation (BSS) methods and SOBI were applied to extract the resulting independent components, followed by bandpass filtering in the range 0.8-2.3 for heart rate and 0.2-0.8 Hz for respiratory rate, respectively. The resultant signals were also high pass filtered using 8 Hz cutoff and low pass filtered using 0.15 Hz to filter out other noises. To select the appropriate channels, k-means clustering was used to cluster the channels, followed by selecting the appropriate channels using kurtosis. BVP signal was selected by choosing

the channel with the highest kurtosis value to ensure its non-gaussian nature, whereas the RR signal was selected based on the low kurtosis value due to its sub-gaussian nature. The performance of the proposed method was tested using low-quality (relaxed state) and high-quality (constrained situation) image data. The limitations of this method are that it did not use automatic ROI selection and tracking. Secondly, rigid motions were not addressed in this work.

The first deep learning based method for Heart rate and breathing rate calculation was presented by Chen and McDuff [13], in which Convolutional Attention Networks (CAN) were proposed for heart and breathing rate. The proposed method consisted of motion and attention modules. Motion module uses a motion representation using normalized frame difference based on skin reflection model. It has nine layers comprising convolutional, Average pooling, and two fully connected layers. The input to the motion module is the clipped normalized temporal difference between two consecutive frames based on the skin reflection model. The attention module follows the same architecture without fully connected layers and 1×1 convolutions between pooling and convolution layers paired set. These convolutions were also responsible for sharing feature maps between both modules, thereby providing spatial-temporal physiological representations. The input to attention modules were individual frames. The output to the CAN model was the first-order derivative of the pulse signal. The output was sampled at the same rate of the video using cubic Hermite interpolation. The final signal was then bandpass filtered with 6^{th} order Butterworth filter followed by power spectral density estimation for heart rate estimation. The loss function used for the method was a mean squared error (MSE) and frequency error (difference between the ground truth and estimated cardiac frequency). The frequency error was calculated by running the trained model to 16 extra epochs after convergence. This resulted in 16 models, which were then fed with training data. Subsequently, the model with minimum frequency error was selected as the final model. The proposed framework was tested with six tasks ranging from stationary to serious motions and considering patient dependence or independence. The trained models were generalized well, and visualizing attention weights gave an insight into the ROI regions focused on the trained models. Specifically, the HR model has focused on the earlobe(larger blood supply), forehead, and carotid arteries (most significant pulse information), while the breathing rate model focused on the nose suggesting the use of nose flaring, ultimately resulting in the extraction of strong physiological signals.

2.1.8.2. Heart Rate and Heart Rate Variability

Gupta et al. [53] presented a multicamera setup consisting of RGB, a monochrome camera with an additional magenta filter, and a thermal camera. This work addressed illumination variation and proposed that the green and red channels with a thermal camera can be used for accurate HR measurements under such conditions. Furthermore, ROI selection was performed using conditional regression forest followed by removing the abrupt intensity changes by thresholding the values having 99% of first SD and detrending. BVP extraction was performed using FastICA with negentropy as an optimization function. Cardiac frequency was extracted by FFT and maximum peak estimation. This work also presented HRV as a measure of stress level in which low frequency HRV component was

visualized as more red and HRV is shown as normalized LF/HF. A high ratio depicted a higher stress level.

Another multi-parameter estimation study for Pulse Rate (PR) and Pulse Rate Variability (PRV) was conducted by Kumar, Veeraraghavan, and Sabharwal [54], which addressed the challenges of PR estimation for the individual with dark skin tone with low illumination conditions, and motion during video acquisition. The study proposed a method named DistancePPG which consists of two modules: maximum ratio changing for pulse rate and pulse rate variability estimation; and motion tracking using deformable face fitting algorithm by detecting facial features "good features to track" followed by KLT for feature tracking. Maximum ratio changing aims at extracting the PPG waveform from multiple facial regions using weighted averaging in which weights were calculated based on blood perfusion and incident light illumination using the maximum ratio diversity algorithm. Motion tracking used the deformable face fitting algorithm for extracting facial landmark locations, followed by dividing the face using these locations. Subsequently, these planar regions were tracked by extracting features using the “good features to track” algorithm and tracking them using the KLT algorithm. These features were then used to find a rigid affine fit followed by applying the RANSAC algorithm [55] for rejecting outlier regions. The regions in which the affine fit model could not be searched were rejected. Finally, all the remaining regions were fed to the MRC step for pulse waveform extraction. The regions which possess the proposed goodness metric G_i between -6 to 2.5 dB only, were used for PPG waveform extraction. G_i worked better than SNR for G_i >- 3dB. For Pulse rate variability, the PPG signal was interpolated to 500 Hz using cubic spline interpolation similar to the sampling rate of a pulse oximeter. Peak finding was performed based on the threshold. The method did not consider non-rigid motions, which were avoided by excluding the area around the eyes and mouth. For the motion scenario, inaccurate PR and PRV estimation were due to excessive rejection of facial regions. Furthermore, DistancePPG performed well for the displacement <4 px, but for displacement >5px, the method did not work.

A similar study presented by Yu et al. [56] emphasized the clinical relevance of HR and HRV for geriatric patients and proposed an HR estimation method for nighttime unobtrusive monitoring. The method used a superpixel method for dividing the face into multiple ROIs. A bagged tree classifier was used to select ROI free from artifacts. Out of 19 extracted features corresponding to heart rate, motion, time, and frequency domain, only four features corresponding to heart rate (*HBratio1, HBdiff1*), one feature from motion (*spMotion*), and brightness was used for ROI selection. Finally, HR was estimated by dividing each raw signal into an interval of 10 secs with one second overlap, followed by identifying the valid HR estimates based on 3bpm difference criteria between estimated and reference HRs. For HRV estimation, a window of two heartbeat peaks with an overlap of one heartbeat peak was used. The proposed method was implemented on RGB videos and NIR videos. Motion artifact from RGB videos was eliminated using the CHROM method, while no artifact removal algorithm was used for green and NIR channel videos. The experimental setup consisted of acquiring the 10 and 15 minutes long videos before and after physiotherapy sessions for geriatric patients after admitting and before leaving the hospital. On the other hand, only two videos were acquired before and after leaving the hospital for healthy individuals. The results confirmed that ROI selection

using superpixel methods worked better than bounding box ROIs, although the difference lay in terms of temporal coverage. However, NIR videos have poor pulsatile component than visible light videos, but NIR is considered better because it does not cause any visual stimulation.

2.1.8.3. Eyeblink and step count with Heart Rate

Zhang et al. [48] proposed six channel Second-Order Blind Identification (SOBI) method for HR and eye blink estimation, simultaneously using the area near the left and right eye region. The proposed SOBI used kurtosis as an optimization function for extracting eye blink and BVP signals, respectively. Furthermore, it was tested for different channels, dynamic HR change, and motion scenarios to justify its effectiveness. The method has shown its efficacy for rotations with a tolerance of 30° yaw rotation. Furthermore, the proposed SOBI was also compared with three variants of ICA, namely JADE, infomax, and FastICA, and Principal Component Analysis (PCA).

Yu-Chen Lin and Yuan-Hsiang Lin [47] estimated pulse rate and step count during different exercises such as walking, biking, and treadmill running, at different speeds. The proposed method consists of two components: chrominance based adaptive filter and normalization (CADN) and Domain Selection Scheme (DSS). CADN consists of a chrominance algorithm cascaded with an adaptive filter followed by temporal normalization. The accuracy of pulse rate was further improved by DSS under which pulse rate was either computed using time-domain trough detection if temporal mean amplitude value is less than the threshold, or frequency domain using short-time Fourier transform if amplitude exceeds the threshold. Step count was estimated using the peak detection method with an adaptive threshold in time. Furthermore, ROI recapturing during high degrees of freedom may lead to abrupt signal generation. Although temporary, the permanent abruptness of the signal due to motion might hinder the pulse rate estimation accuracy. Table 2 presents the summary of multi-parameters estimations.

Table 2. A summary of multi-parameter estimations.

First Author (year)	Parameters	ROI Used	Method	Color channel	Limitations
X Yu [56], (2021)	HR, HRV	Face	CHROM	RGB, NIR	The method uses NIR channels that have poor pulsatile strength.
M Kumar [54], (2015)	PR, PRV	Face	MRD	RGB, Mono	The method tracks ROI using the KLT algorithm, hence for larger motion, features cannot be tracked, leading to PPG information loss.
O Gupta [53], (2016)	HR, HRV	Cheek, Forehead	Fast ICA	RGB, Magenta thermal	The study did not consider motion artifacts.
Z Yu [30], (2019)	HR, HRV	Face	CNN	RGB	They proposed a network for HR and HRV estimations using compressed videos but did not use any metric to measure the effect of compressions for estimations.
M Poh [18], (2011)	HR, HRV, RR	Face	JADE	RGB	The optimization function for JADE did not possess good statistical properties to ensure statistical independence between PPG signal and noise.

L Tarassenko [51], (2014)	SpO2, HR, RR	Face	AR, ROR	RGB	Lower SpO2 values could not be measured; Motion artifacts were not addressed.
F Bousefsaf [50] , (2013)	HR, BR	Face	CWT	RGB ,CIE	The motion was considered as the case of stress, which does not resemble the real-time scenario.
B Wei [52], (2017)	HR, BR	Throat, Mouth	SOBI	RGB	The study used manual ROI tracking, and rigid motions were also not addressed for estimations.
W Chen [13], (2018)	HR, BR	Face	CNN	RGB, NIR	Illumination variation for a longer duration may halt the accurate extraction of PPG signal and subsequently HR estimation.

2.2. SpO2 Estimations

As per the literature, all the studies have used the conventional *ROR* method for SpO2 calculations, except one study by Gastel, Stuijk & Haan [57]. However, the studies differ in the selection of color channels and color wavelengths. A total of five studies have been found by extensive literature search through multiple databases.

Guazzi et al. [58] addressed the effect of skin tones on SpO2 measurement accuracy and proposed a multiple ROI based SpO2 estimation. The proposed method consisted of the following steps: 1) ROI search area exploration using SNR, 2) dividing the ROI area into $n \times n$ regions; 3) splitting the video into 12 seconds segments with an overlap of one second; 4) HR, BR, and phase difference calculation for each ROI; 5) ROI selection based valid HR and BR estimates; 5) normalized AC calculation; 6) weighted averaging over selected ROIs using the ratio of SNR of each ROI and reaming ROIs; 7) green channel signal averaging, high pass filtering and peak identifications; and 8) SpO2 calculation using a log of *ROR* method using blue and red channels. Furthermore, to analyze the relationship between saturation and color change and eliminate the subject bias, the method used leave one out approach, in which the gradient (slope) for a particular subject would be the average of the gradient (slope) of all other subjects, the intercept was calculated from the first one minute of the video. Additionally, the thresholds for the inclusion criteria were chosen manually, which may exclude the ROI having vital information for SpO2 calculation and *vice-versa.*

Shao et al. [59] proposed a SpO2 estimation method using two wavelengths, 611 nm for orange and 880 nm for the near-infrared (NIR) region, considering the SpO2 range 83-100 % for the testing phase. The method starts with splitting the ROI video into segments of 10 seconds each. The AC and DC coefficients were calculated using peak-to-peak values and mean values of each window. Subsequently, a linear model was fitted between SpO2 measurements and *ROR* values for the 70-100% SpO2 range. The study also proposed a hardware design, which included a monochrome camera surrounded by two Light Emitting Diodes (LEDs) boards, with each board having NIR and orange LED in alternative rows. A microchip controller was used to generate the (LEDs and camera) trigger signals for switching the NIR and orange signals on and off.

Gastel, Stuijk & Haan [57] proposed a motion invariant SpO2 estimation method based on blood volume pulse signature, which was mapped to different SpO2 levels.

Specifically, the proposed method extracts the PPG signals for multiple ROIs, thereby selecting the PPG signal with the best SNR for SpO2 estimation using a five-point moving average. PPG signals were extracted using the proposed adaptive blood volume pulse vector method (APBV), based on the Blood Volume Pulse Vector (PBV) method. APBV is a generalized version of PBV, which aims at finding the pulse signal signature by estimating a weight matrix by equating the correlation of PPG signal and color channels with the pulse vector. The limitation of PBV is that if the correlation between pulse signal and color channels is not equal, it may add noise to make them equal. APBV overcomes this limitation. Specifically, adaptive PBV estimates the PBV vector as the sum of static PBV vector at 100% SpO2 value and PBV vector update weighted by a gain factor alpha, which adapts to overcome the noise artifact due to inequality of correlation between PBV vector with pulse signal and color channels. The new PBV vector was then mapped to different SpO2 values. The proposed framework consists of four steps: ROI selection, Pulse extraction; distorted regions pruning based on two spatiotemporal features; PBV selection, and SpO2 estimation. ROI selection was performed by CSK algorithm followed by calculating Q-score by the product of spatial features cross-spectral SNR ratio and spectral peak correspondence. The pulse signals were extracted from the selected ROIs using the APBV method. Finally, the pulse signal with the highest SNR would be utilized for SpO2 estimation. Furthermore, the proposed method uses two sets of three wavelengths (675,800,840) and (760,800,840), in which the former performed relatively well, proving the potential of this method under visible light by slightly adjusting the PBV vector and suggesting the better contrast between the Hb and HbO2 absorption differences. This method did not work well with low SpO2 values due to high PPG amplitude. Although the proposed method works well under infra-red light, visible light may create a challenge due to low temperatures. During low temperatures, the viscosity of blood increases, leading to reduced blood flow in the capillaries, but arterioles will remain unaffected due to their location depth within the body. Consequently, the low body temperature of the skin decreases the PPG amplitude due to high blood viscosity. Furthermore, this method will not perform well in the case of periodic movements like rhythmic exercises.

Rosa and Betini [60] proposed an Eulerian Video Magnification (EVM) based SpO2 estimation approach. The proposed method used EVM to detect changes in the facial skin followed by using *ROR* for SpO2 calculation using the red and blue color channels. Furthermore, a raspberry pi based hardware solution was proposed for real-time SpO2 calculation. SpO2 estimation was carried out under normal breathing and breath holding events for which the accuracy of equipment (A_{rms}) value was below 2%, making the prototype a suitable system to use. The proposed method did not prove its applicability for hypoxia patients and SpO2 values less than 92%. The study covers the subjects for skin type II-IV according to the Fitzpatrick scale. It also used image stabilization to select relevant image frames of the video with the threshold of 95% for comparing the similarity between two consecutive frames before applying EVM.

Moço and Verkruysse [12] hypothesized that a green and red channel combination could also be used for SpO2 estimation. Furthermore, this combination was tested with a conventional red and infrared combination. The green channel was tested since it has the strongest PPG amplitude. The study was conducted in three scenarios: (N) in which

normal conditions along with ambient temperature with the forehead as ROI; (H) hypoxia conditions along with ambient temperature with the forehead as ROI; and (C) normal oxygen conditions with ambient temperature using forehead and cheeks as ROIs, were used for collecting the video samples. The spatially averaged pixels forming the three channel traces were AC/DC normalized and denoised. Subsequently, FFT was applied, and peak detection was performed using a green channel. Inverse FFT was applied to detected pulse rate frequencies and included in "overlap and add fashion" using the Hanning window with 50% overlap for all three channels. Final signals were extracted by calculating the medians of peak and valley, thereby using the *ROR* method for SpO2 calculation. The study tested the effect of temperature on SpO2 estimation, concluding that red Over infrared (RoIR) and Red Over Green (RoG) indicated high errors in cold temperature. However, RoG showed significantly large errors than RoIR. Overall, RoIR performed better than RoG. The study discussed the effect of specular reflections and suggested using polarization filters for specular reflection suppression. Motion interferences were suppressed using the IR channel, but the effect could not be understood fully in the study, although it provided slightly better estimates. A summary of Spo2 estimation studies is presented in table 3.

Table 3. A summary of SpO2 estimation studies.

First Author (year)	Parameters	ROI Used	Method	Color channel	Limitations
M Gastel [57], (2016)	SpO2	Face	APBV	IR	The ABPV based SpO2 method will not work well for periodic movements; the PPG signal is also mapped to different SpO2 levels, which can only map to those levels, not other SpO2 values.
AR Guazzi [58], (2015)	SpO2	Face	ROR	RGB	The proposed method did not address the motion, illumination variations, effect of melanin concentrations, and breathing rate calculation could not be accurate due to the small processing window.
A Rosa[60], (2020)	SpO2	Forehead	ROR	R, B	The method was unable to show its applicability in low SpO2 values scenario.
A Moço [12], (2021)	SpO2	Forehead Cheeks	ROR	R, G, IR	The study proposed ratio of red over green for SpO2 estimations but could not perform better than red over IR.
D Shao [59], (2016)	SpO2	Lips	ROR	Orange, NIR	The study did not address motion artifacts.

2.3. Heart Rate and SpO2

Kong et al. [61] proposed two wavelengths, 520 and 660 nm, to acquire facial data using two monochrome CCDs. These wavelengths were chosen because, at 520 nm, the absorption coefficient of HbO2 and Hb are the same and different for 660 nm, respectively. Heart and respiratory rate and SpO2 were calculated by the standard approach (ROI identification, spatial averaging of both wavelength videos and FFT),

resulting in pulse signal extraction. Pulse signal was extracted using a bandpass filter of 0.7-3.0 Hz. This PPG signal was used for SpO2 calculation using a 10 sec window moving average followed by the *ROR* method. During the study, it was found that SpO2 recovery is faster in the mouth region than finger region. Furthermore, the wavelength chosen for this method remains unaffected by varying light intensity and temperature. However, the study did not address the motion artifacts and used diffusion lights to deal with illumination variations.

Bal [62] employed dual-tree complex wavelet transform for HR and SpO2 estimation in normal and clinical settings. Furthermore, the effect of hemoglobin was also checked in the anemic patients and found that low hemoglobin level leads to high HR error estimates. This method used three-level Dual-Tree Complex Wavelet Transform (DT-CWT) PPG signal extraction. The initial HR signal was created using facial recognition, skin segmentation, detrending, and chrominance signals. Subsequently, soft thresholding was applied, followed by signal reconstruction using inverse DT-CWT. The Fast Fourier transform was used to convert the signal to frequency domain followed by maximum peak estimation corresponding to HR frequency. SpO2 was estimated using the *ROR* method employing red and blue color channels using wavelengths used by Tarassenko et al. [51]. The studies presenting HR and SpO2 estimations are shown in table 4.

Table 4. HR and SpO2 estimations summary.

First Author (year)	Parameters	ROI Used	Method	Color channel	Limitations
L Kong[61], (2013)	SpO2, HR	Cheeks	FFT	Mono	The method did not address motion or illumination variations.
U Bal [62], (2014)	SpO2, HR	Face	DT-CWT	RGB	The study uses few subjects to test their method; hence, analyzing its performance could have taken a relatively higher number of subjects.

2.4. Review Summary

Non-contact estimation approaches presented in this chapter are at the proof of concept stage with a few shortcomings such as relatively constrained video acquisition settings, smaller sample size, limited clinical context. The reference devices used for most of the studies are pulse oximeters, while some have also used ECG. Some studies have also used other available devices whose accuracy can be questionable compared with a standard HR monitoring device (ECG or PPG). The selection of a valid reference device plays a crucial role in assessing the applicability of the proposed method in comparison to it. Additionally, a valid reference device could also address the limitations of ECG or PPG in interpreting the results.

Most studies have used lower camera resolution, making it a cost-effective solution for real-time monitoring such as driving, fitness exercises, and clinical monitoring. However, selecting the appropriate video resolution is challenging. It is also affected by the distance between the camera and the subject's face. A study conducted by Song et al. [41] tried to find the optimal resolution and camera shooting distance and concluded that higher resolution enhances the quality of PPG signal and distance more than 1 meter (m) will

deteriorate the HR estimations, which is consistent with the findings of this review. However, high-resolution cameras are computationally intensive for estimating physiological parameters, and a 1 m or less camera distance limits the applicability of non-contact approaches for clinical or sleep settings.

Also, frame rate plays a crucial role in tracking tiny variations in the image sequences, ensuring an accurate PPG signal. A few attempts have been made with higher sampling rates, but the performance of HR estimation with higher framerate in comparison to 30 fps did not show significant performance improvement [63]. Furthermore, a framerate of 30 fps worked well with most of the studies. Except for clinical and SpO2 estimation studies, most studies have acquired the data for about 1 min to 3 min, which may limit the legitimacy of the methods. Moreover, a shorter interval may hinder the robustness of the proposed method under different estimation conditions. SpO2 estimation studies have considered relatively longer duration subject's videos. The limitation of the prolonged video acquisition is the presence of artifacts due to movement and uneven illuminations. Besides, image quantization can also produce undesirable noise, but this effect can be mitigated by assuming constant light over the region of interest. The presence of these artifacts deteriorates the estimations of the physiological parameters, as the PPG signal (from the RGB color channel) is very weak and is challenging to extract from the artifacts' corrupted signals. An evident approach to mitigate the problem is to convert RGB to other movement or illumination artifact-resistant color channels such as YUV [33], LAB [29]. For dark scenarios, an infrared channel could be a better alternative, but the only problem is that the strength of the PPG signal is relatively weaker than the signal from the RGB channel. A combination of RGB with IR, similar to the one conducted by Kado et al. [14] or other color models, may produce promising results but increases the problem's complexity and is computationally intensive.

A vast category of non-contact PPG extraction has been used in the literature. Neural networks and their variants have been used extensively for HR estimation studies. The neural networks based methods performed relatively better than other conventional non-contact PPG estimation methods. Additionally, there has been extensive use of transfer learning for HR estimation methods. Most importantly, neural networks do not need assumptions to process the data, which was the case with the existing state-of-the-art methods. Furthermore, the neural network also possesses good generalization ability. A distribution of estimation methods presented is shown in Fig. 2. On the other hand, SpO2 studies have employed regression using ROR with an exception. The exception is the study conducted by Gastel et al.[57], which uses a blood volume pulse signature based method followed by PPG signal mapping to different SpO2 levels.

Furthermore, several studies have justified their method's clinical relevance by reporting the accuracy, which is calculated as the percentage of study samples having an error less than ±5 bpm. However, it is worth pointing out that the studies have predominantly used their self-created databases under well-constrained laboratory conditions to test their method in the normal HR and SpO2 ranges. The performance of these methods may deteriorate for abnormal HR parameter ranges. In addition, the accuracy in this scenario will not be sufficient to justify their clinical relevance and therefore needs further analysis. On the other hand, the performance of SpO2 estimation under extreme conditions is challenging to test since it needs multiple breath holding events, which is not always

possible for individuals. Consequently, developing a robust SpO2 estimation method is difficult due to the need to measure the subtle changes in the saturated blood. Therefore, there are limited SpO2 estimation studies in the literature. There is a need to devise methods to estimate SpO2 values from a single PPG signal extracted from the facial ROI, similar to other physiological parameters.

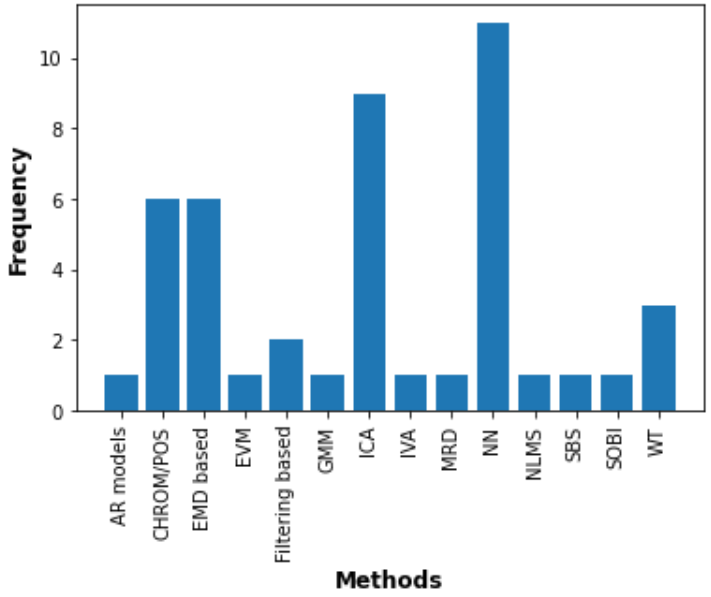

Fig. 2. Various estimation methods used for HR estimation studies.

Another limitation with almost all studies is that the parameters were estimated for healthy individuals, limiting the estimation methods' ability to measure for diseased people with conditions such as hypoxemia, bradycardia, or tachycardia.

Finally, the literature suggested a few common factors applicable for all non-contact approaches to estimate physiological parameters. Ethnicity, movement, illumination, and clinical relevance depicted by accuracy are the common factors used by all studies. Although none of the SpO2 estimation studies have reported movement and illumination artifacts, these factors are worth considering while developing non-contact approaches for physiological parameter estimations. Therefore, this proposal focuses on designing novel non-contact estimation approaches by considering the abovementioned factors to estimate two physiological parameters, such as HR and SpO2, using face videos. Furthermore, no studies have been reported that estimated these parameters under a darker environment to the best of our knowledge. Therefore, the methods developed would be tested in a dark environment to test their applicability for sleep and clinical scenarios, where the intensity may fluctuate from dim light to dark.

3. Methodology

A thorough literature review of non-contact state-of-the-art physiological parameters estimation methods provided a basis for designing non-contact estimation studies. This would help in developing standard and reliable non-contact estimation methods followed by testing them under various scenarios. For example, all face based physiological parameter estimation methods should address the skin ethnicity/color, motion, and illumination artifacts, assess their clinical relevance.

The study is divided into four tasks: data collection; best state-of-the-art methods identification, developing non-contact methods for HR and SpO2 methods under normal light conditions; testing the feasibility of these methods under dark environment, and performance analysis of methods developed during the study. The proposed workflow of activities carried out during this study is depicted in Fig. 3.

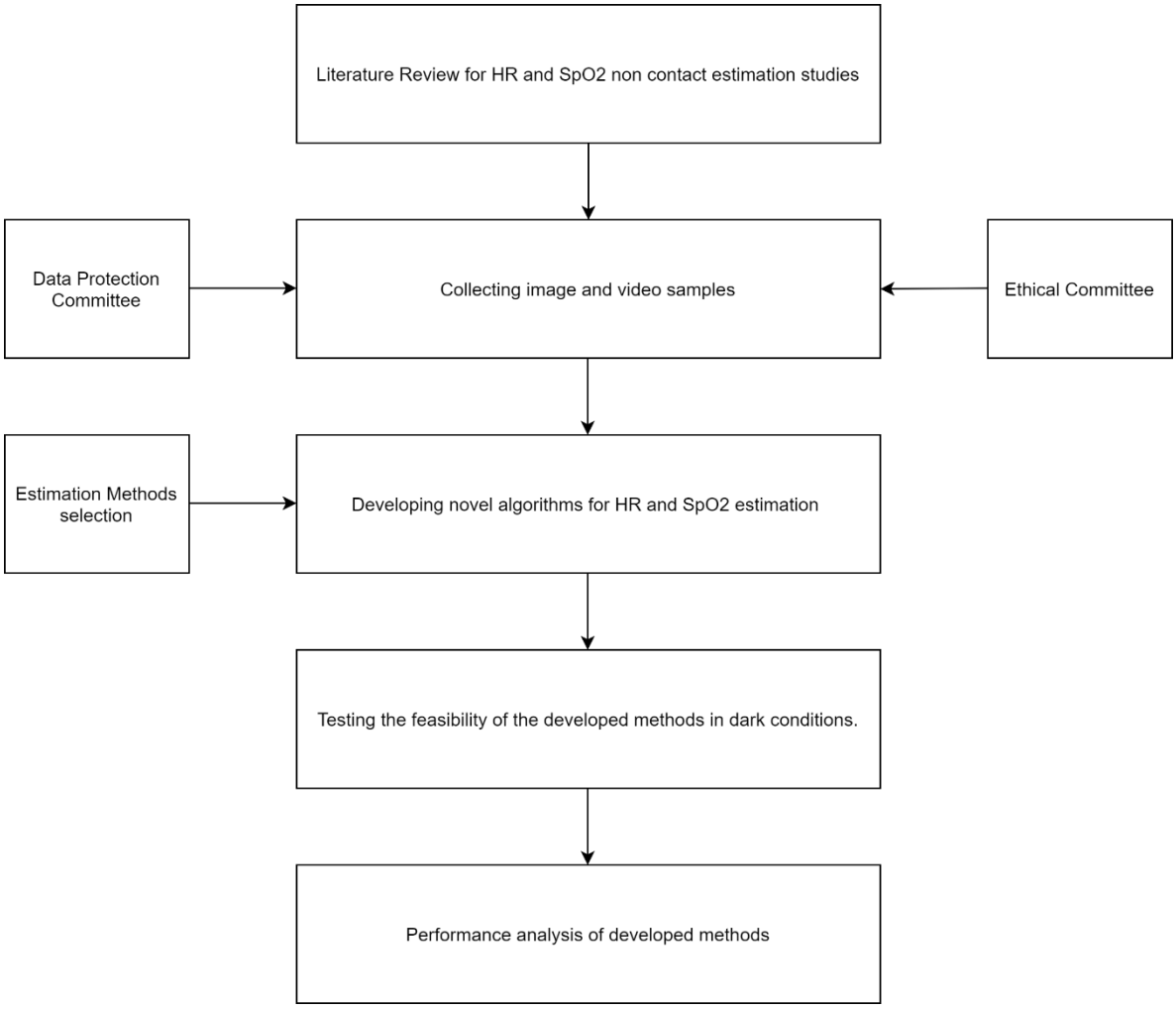

Fig. 3. The tentative workflow of the proposal.

3.1. Data Collection

To the best of our knowledge, all publicly available databases have been created using single-multiple external light sources. As this study explores the possibility of estimating

physiological parameters in a darker environment, a database is created by considering a few essential factors such as video resolution, frame rate, skin tones. For instance, high video resolution and framerate can measure detailed subtle color variations for accurate PPG information extraction. Additionally, the estimation accuracy varies for different skin tones. Therefore, the facial images in RGB format and video samples in mp4 format, synchronized with ground truth HR and SpO2 values using a pulse oximeter, would be collected and stored. The data collection process seeks approval from the ethical and data protection committee. Hence, the needed documents were prepared before submitting to these committees. Ethically approved informed consent forms were provided to volunteers before participating in the study so that they could understand the purpose, benefits, and potential risks associated with the study. All the documents sent to both committees are presented in the appendix.

Although there is limited availability of benchmark databases for physiological parameters estimations, the existing databases provide essentially similar real-time conditions for estimations. For example, the PURE database [64] covers seven types of head movements during heart rate acquisition, VIPL-HR database [65] collects the videos considering nine different kinds of conditions. Therefore, their contribution in analyzing the newly developed methods in due course of this study is non-trivial. Considering this, a formal application seeking access to these databases would be sent to the respective database owners.

3.2. Identifying the best estimation methods

The majority of the developed estimation methods have used their self-created databases. As a result, the methods might work well for their databases but not with other databases because different studies have recorded the videos in constrained laboratory conditions or according to their testing requirements. Moreover, the estimation methods used for HR estimations are highly diversified. The literature review identified 14 different types of methods, while for SpO2 estimation, the conventional ROR method has been used. Therefore, it is crucial to test the generalizability and reliability of the currently existing methods by testing them using various benchmark databases. Furthermore, it is also essential to identify the best performing estimation methods to explore their ability for other conditions and real-time applications. Additionally, limited studies have justified the proposed method's clinical relevance, an essential parameter for transforming the proof of concepts to real-time. It was found that ICA and deep learning networks are the prominently used methods for physiological parameters estimation. Therefore, publically available databases will be used to assess the generalizability and reliability of the deep learning (11 studies) and ICA (9 studies) based methods, followed by developing non-contact estimation methods using these types of methods.

3.3. Developing novel algorithms for HR and SpO2 estimations under different environments

3.3.1. ICA Based Method for HR estimation under ambient light conditions

As mentioned in the previous section, ICA is one of the predominant methods for BVP signal extraction, which will be processed to get an HR value. Therefore, this study aims at identifying the constraints and challenges faced by currently existing state-of-the-art ICA based methods, thereby proposing a novel method for HR estimation. A general assumption about ICA based methods is that the number of independent signals is equal to the number of mixed signals. In other words, the signal constructed by each color channel (mixed-signal) results in an Independent Component (IC). This assumption requires analyzing each IC as a potential candidate for the BVP signal and requires apriori knowledge. Moreover, there is no defined criterion for selecting the BVP signal from the independent components from different color channels. Conventionally, BVP signal extraction includes selecting the component with the highest periodicity, which may result in choosing the wrong IC as BVP signal in case of periodic motions by the subjects. However, most studies selected the second IC for BVP signal extraction by discarding the 1st and 3rd IC corresponding to the red and blue color channel. This results in loss of information present in the red and blue channels, which may be vital for HR estimation. Considering the limitations mentioned above, BVP extraction could be considered as an undercomplete problem. In other words, given three mixture signals corresponding to R, G, and B color channels, the task would be to extract one IC which corresponds to the motion and illumination variation resistant BVP signal. The workflow of the proposed method is presented in Fig. 4.

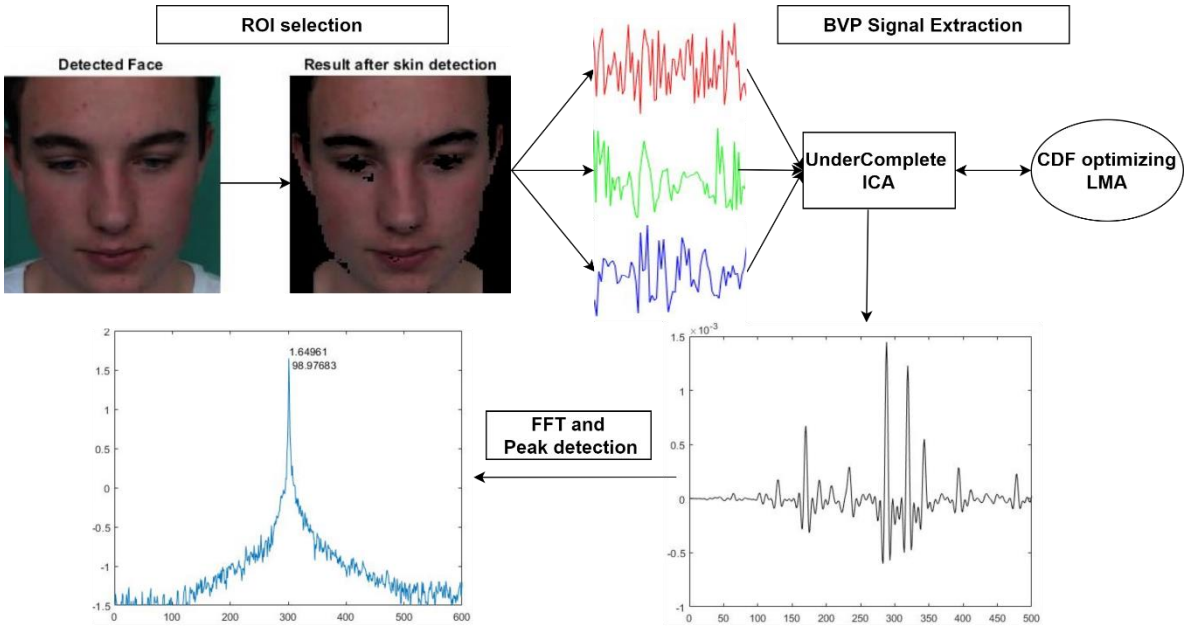

Fig. 4. Workflow for ICA Based Method for HR estimation under ambient light conditions

a) ROI Selection

ROI selection deals with identifying the face using the Viola-Jones face detector, followed by skin segmentation. The skin would be segmented using Cb and Cr components of the YCbCr color model. Subsequently, a spatial averaging operation for each channel was performed on each image frame of the video. A detrending process would also be applied to remove slow non-stationary drifts in the signal. Finally, an overlapping moving window operation would be used for each channel for constructing raw signals.

b) BVP Signal Extraction and HR estimation

The method would translate the BVP extraction as an undercomplete problem that would take three mixture signals and extracts a single IC, consisting of BVP information from all three channels. This problem would be solved using the newly proposed method, which will use a cumulative distribution function (CDF) of the raw signals, followed by its optimization. This optimization algorithm will allow the solution to converge to the desired values of the unmixing matrix in both conditions. Finally, using frequency domain methods, HR would be estimated by extracting the cardiac frequency and converting it to beats per minute (bpm).

3.3.2. Testing the feasibility of the ICA based method under darker conditions

Estimation methods developed under normal light conditions would require accurate ROI extraction, which is not feasible in darker or dim light environments. For instance, it could be possible that the conventional face detectors might not work due to poor illuminance. This would result in PPG information loss due to inaccurate tracking of ROI. To solve this problem, the RGB image frames of the video should be enhanced using state-of-the-art image processing methods. For example, gamma correction and brightness correction may improve the ROI tracking, thereby extracting PPG information and physiological parameters estimations. Therefore, the methods developed under normal light conditions will be enhanced using conventional image enhancement methods for accurately tracking the different skin tones for ROI selection. Alternatively, image enhancements methods would be developed to improve the quality of videos acquired without using an external light source.

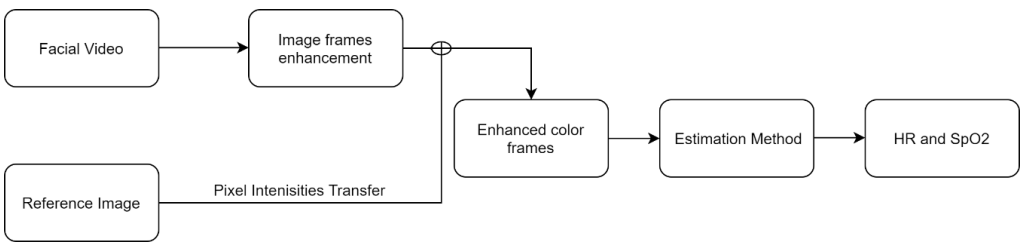

Fig. 5. Framework for HR estimation in a darker environment.

For this purpose, the data collection step also includes collecting colored face images in normal light conditions. The tentative approach would use pixel intensities of the image acquired during normal light conditions to enhance the image frames of the video

collected in the darker environment. Fig. 5. depicts the workflow of the proposed approach.

3.3.3. Deep Learning based non-contact estimation method for HR and SpO2 estimation

Conventional non-contact estimation methods used a well-defined mathematical model with certain assumptions for HR and SpO2 estimation. These assumptions may not hold in all scenarios. Furthermore, the conventional estimation methods did not work well with all skin tones. Additionally, these methods did not possess any mechanism to deal with the redundancy present in the video, which enhances their complexity due to extra information. In other words, conventional methods lack generalizability and reliability in terms of estimation accuracy and are also sophisticated. Hence, there is a need to develop methods, which can reduce these assumptions to a minimum, with acceptable generalizability and reliability. Considering these limitations and abilities of deep learning in solving computer vision problems, this study will propose deep learning methods for non-contact HR and SpO2 estimations. Moreover, the deep learning methods possess self-learning ability, which contributes to its good generalization capability. However, deep learning frameworks cannot be directly applied to physiological parameters estimations tasks. The reason is that changes in pixel intensities in the image frames consist of several other components such as motion due to breathing, periodic movements due to continuous flow of blood. Therefore, applying deep learning frameworks by feeding the videos to output the physiological parameters is not a good approach. The probable solution to this problem is to devise certain representations encompassing the spatial and temporal information present in the video, allowing accurate PPG extraction, thereby estimations.

This part of the study aims to develop spatial-temporal representations from the facial videos, which can extract the information of interest for parameters estimation. Furthermore, deep learning's attention mechanisms will be used to extract only the PPG information, rather than redundant and irrelevant information. This way, spatial-temporal representations with attention mechanisms will ensure the method's reliability for accurate estimations. Additionally, techniques like transfer learning will also be used to develop deep learning methods for estimating HR and SpO2 under normal and dark conditions.

3.4. Performance metrics

Based on the literature review, it is found that the root mean square value, Pearson correlation, and Bland-Altman plot analysis have been predominantly used by estimation studies. Furthermore, few studies have also reported mean and standard deviation errors to justify their performance. Clinical relevance was also justified in a few studies using an accuracy measure, which is the percentage of samples with clinically acceptable error limits between estimated and ground truth values. For HR, the clinically accepted error limit is ± 5bpm, whereas, for SpO_2, it is ±3 %. The performance analysis of all the methods developed during this study will be measured by all the parameters discussed here, i.e., mean error, error standard deviation, root mean square value, Pearson correlation, Bland-Altman analysis, and accuracy. Furthermore, the performance analysis

ICA and deep learning based methods developed during this study will also be performed using these metrics

4. Timeline

Table 5. Timeline for the proposed study.

Activity	**Months**					
	1-6	**7-12**	**13-18**	**19-24**	**25-30**	**31-36**
***Literature Review and identification of crucial factors for estimation studies.[1]**						
Document submission to ethical and data protection committee and approval.						
Facial images, videos, and ground truth collection.						
***Independent component analysis based method for HR estimation under normal light conditions.[2]**						
Independent component analysis based method for HR and SpO2 estimation under darker environment.						
Identifying the best deep learning method for HR estimation (No deep learning method proposed for SpO2).						
Deep Learning method for HR and SpO2 estimation under normal light.						
Fine tuning and development of Deep Learning method for HR and SpO2 estimation for darker conditions.						
Thesis writing and compilation.						

Work completed | Pending Work

Note: * indicates the work submitted for publications to the respective journal.

1. The corresponding publication was submitted on October 8, 2021.
2. The corresponding publication was submitted on July 3, 2021 whose first revision was sent on November 14, 2021.

5. Results

This section presents the results of already completed work during the course of the study. As mentioned in the timeline, ethical permission followed by data collection has been completed. Furthermore, a systematic review of currently existing HR and SpO_2 non-contact methods have already been carried out and submitted for publication in Computer Methods and Programs in Biomedicine journal. The findings of this review were further used to develop a non-contact estimation study based on the commonly used method ICA and was submitted to IEEE Journal of Biomedical and Health Informatics. Further subsections present the details of work carried out untill the proposal submission date.

5.1. Ethical Approval and Data Collection

Data collection from the volunteers was carried out after the ethical permission by the University of Madeira (UMA). The approved consent forms by UMA's data protection and ethical committee are presented in Appendix. An image and a video synchronized with the pulse oximeter were collected for each participant post consent form signing. This resulted in a database consisting of 40 participants (12 females and 28 males) with the age range between 19 and 61 years. The images were collected in ambient light conditions while the video samples synchronized with pulse oximeter readings were acquired in dark conditions.

5.2. Availability and Performance of Face based Non-Contact Methods for Heart Rate and Oxygen Saturation estimations: A systematic review

5.2.1. Study screening results

Out of 332 articles retrieved from the search strategy presented in table 1 using multiple databases, 32 articles were included, followed by data collection and analysis. While screening these articles, 18 more studies were included by thoroughly checking the references list of every included article. This technique is called *snowballing*. Fig. 6 depicts the PRISMA flow diagram illustrating the article's screening process. Consequently, 50 articles were included in the final review, out of which 38/50 (76%) studies estimated a single, 10/50 (20%) studies estimated two, and the remaining estimated three physiological parameters. It is important to note that no explicit search was performed for other parameters except HR and SpO2, although other parameters were also calculated as a part of HR estimation studies.

However, the study did not include non-contact methods using chest, arm, palm, or finger for HR estimations since this review is constrained to face based methods only. Furthermore, studies constituting fetal heart rate monitoring were not considered since HR estimation was performed using the lower abdominal area.

One article's full text [66] could not be retrieved for which the authors were contacted, but no responses were received.

5.2.2. Population characteristics

5.2.2.1. Age and gender

21/50 (42%) did not report the age, while the remaining 29/50 (58%) studies reported the age range. The minimum and maximum age range for all studies lie between 18 and 80. Furthermore, it is difficult to plot the distribution of age ranges due to the considerable heterogeneity.

Gender has been reported by 34/50 (68%) studies, while 16 studies did not provide any information. Except for 4/50 (8%) studies [10, 17, 40, 56], all the study samples were male dominant. However, these studies have collected data from a relatively lower number of participants.

5.2.2.2. Ethnicity and skin color

Numerous studies proved the importance of considering the ethnicity or skin color for HR and SpO2 estimation studies since darker skin tone poses more challenges for estimation tasks than white skin. 14/50 (28%) studies reported the subjects' ethnicity, skin color, or tone information for estimation studies. Furthermore, 8/14 (57.14%) [23, 24, 31, 50, 57, 58, 60] studies used the Fitzpatrick scale to define the subjects' skin tone. The study conducted by Haan et al. [22] included the subjects from i-vi (all scales), while two studies conducted by Wang et al. [23, 24] included subjects with the i-v scale. The remaining studies included subjects with a scale ranging from ii-iv. On the other hand, 6/14 (42.85%) studies instead mentioned the ethnicities of the subjects: 3 studies [11, 25, 53] considered two or more,2 [10, 41] studies considered subjects with Asian ethnicity, and one study conducted by Kumar et al. [54] has considered skin color.

5.2.3. Study design

5.2.3.1. Physiological parameters

Out of the total included studies, 42/50 (84%) studies belongs to HR estimations, while 5/50 (10%) to SpO2 estimations. Additionally, 3/50 (6%) studies estimated both physiological parameters simultaneously. Furthermore, out of 42 studies, HRV, breathing rate, eye blink and step counts were estimated in 4/50 (8%) [18, 53, 54, 56], 4/50 (8%) [13, 18, 51, 52], and 1/50 (2%) each studies, respectively.

5.2.3.2. Databases used

For HR studies, 34/45 (75.55%) studies used self-created databases with the number of participants ranging from 4-117 with 25.82 +/- 25.11 (mean ± std). In contrast, the remaining studies used benchmark databases to prove the efficacy of their respective HR estimation methods. Furthermore, 24/34 (35.29%) studies have only used their databases, while the rest have used self-created as well as publicly available databases. 3/11 (27.27%) studies [9, 17, 32] have used a single database, while 8/11 (72.72%) [16, 21, 30, 34-38] studies have employed more than one database for performance analysis of their HR estimation algorithms. All SpO2 studies created their databases with the number of participants ranging from 4-46 with 20.5 +/- 16.78.

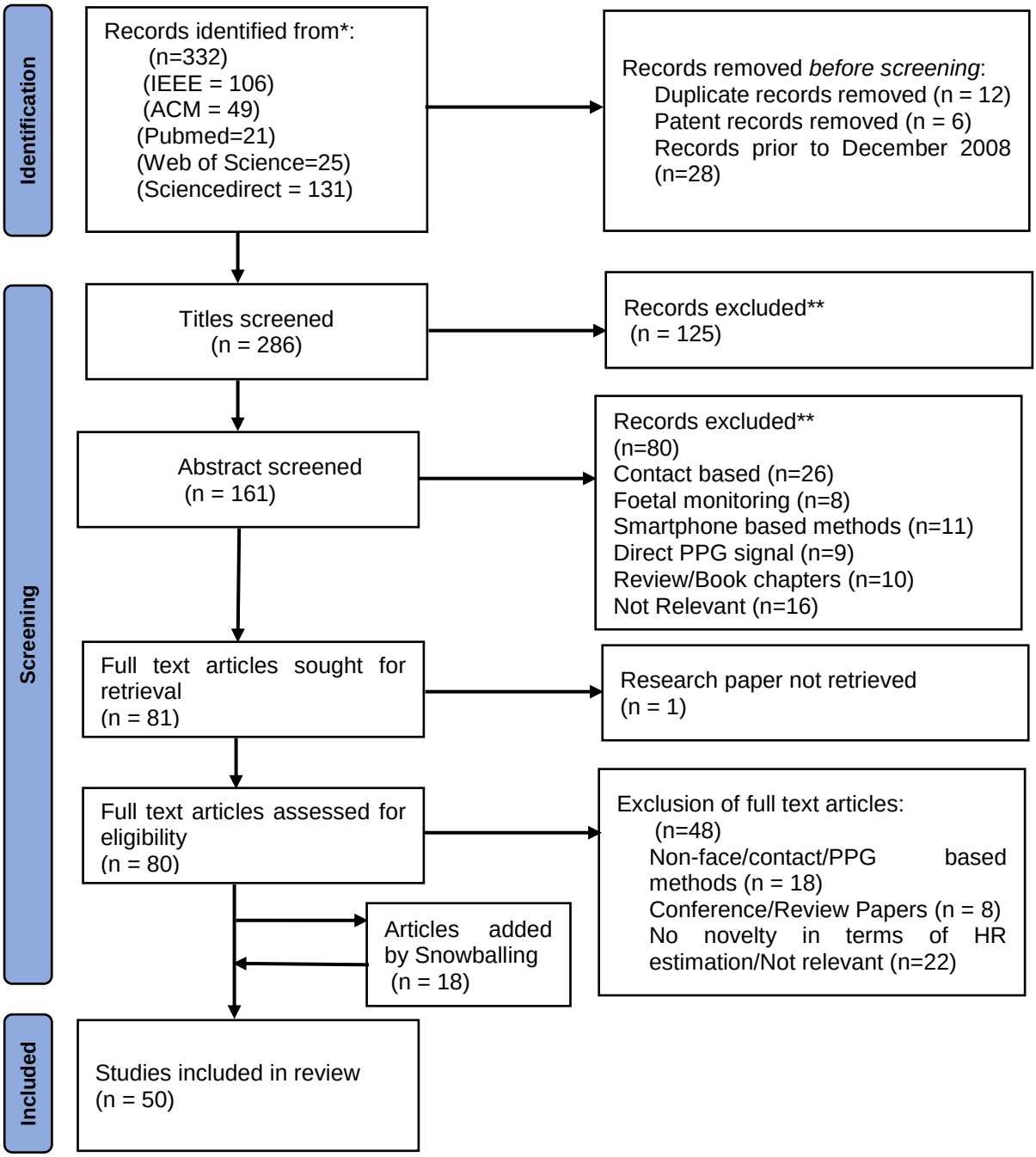

Fig. 6. Prisma Flow Diagram

5.2.3.3. Region of interest selection

Face based physiological parameters estimation needs a ROI, which will be used to extract the source signal. In total, all HR estimation studies used six ROIs, namely, face, cheeks, nose, forehead, areas near eyes, and mouth. 7/45 (15.5%) [12, 39, 42, 47, 50, 52, 53] studies have used two or more ROIs from the face region, while face and cheeks were used by 27/45 (60%) and 15/45 (33.33%) studies, respectively. On the other hand, nose and forehead have been used by five studies each while remaining used mouth and areas near eyes. The ROI distribution for HR and SpO2 studies is shown in Fig. 7.

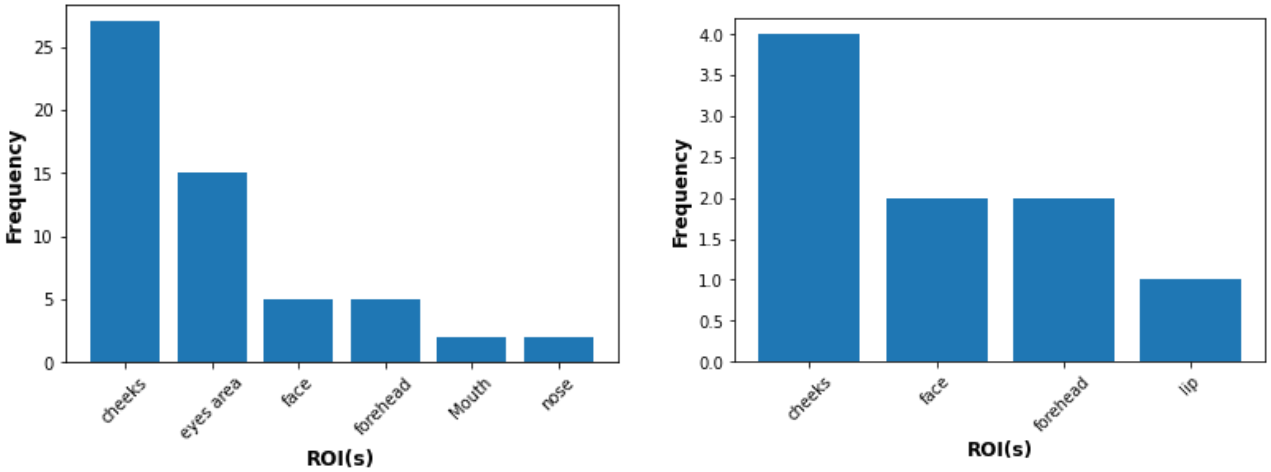

Fig. 7. ROI selection distribution for HR estimation studies (left) and SpO2 estimation studies (right).

5.2.3.4. Artifacts

Artifacts removal is important for BVP or PPG signal extraction using facial videos since the PPG component has a relatively weaker strength and amplitude than the artifacts. HR studies have addressed two types of artifacts, namely motion and illumination artifacts. 4/45 (8.88%) HR studies have addressed and alleviated the effect of illumination, while motion artifacts have been addressed and mitigated in 16/45 (35.55%) studies. On the other hand, 16/45 (35.55%) studies [13, 14, 16, 22, 23, 25, 29, 31, 33-35, 38, 40, 42, 54] have addressed and proposed strategies to lessen both artifact effect. 9/45 (20%) HR and all SpO2 estimation studies did not address any artifact.

5.2.3.5. Estimation methods

As mentioned before, conventional HR estimation methods extract the BVP/PPG signal from the RGB signal traces, calculate the highest frequency, and multiply it by 60 for frequency to bpm conversion. Among conventional HR estimation methods, 9/45 (20%) studies [11, 14, 16, 18, 19, 41, 48, 52, 53] used Independent Component Analysis (ICA), 6/45 (13.33%) studies [22, 23, 25, 36, 39, 56] used color subspace transformations (CHROM/POS), and 6/45 (13.33%) studies [7, 10, 26-29] used Ensemble Mode Decomposition (EMD) and its variants. However, with the advent of deep learning, several end-to-end HR estimation methods have also been proposed, which use different types of neural networks architectures for estimating the heart rate using a facial video.11/45 (24.44%) [13, 30-38, 52] studies utilized neural networks and their variants for HR estimation. Other PPG extraction methods used by HR estimation studies are wavelet transforms [40, 50, 62], filtering based methods [47], autoregressive models [51], Gaussian mixture models [46], eulerian video magnification [9], independent vector analysis [10, 17, 37], maximum ratio diversity [54], sub-band selection [40] and multiple linear regression [7]. Fig. 8. depicts the distribution of the estimation methods used in the literature. On the other hand, all SpO2 studies used the ROR method followed by

regression, except the study conducted by Gastel et al. [57], which used blood volume pulse signature for SpO2 extraction.

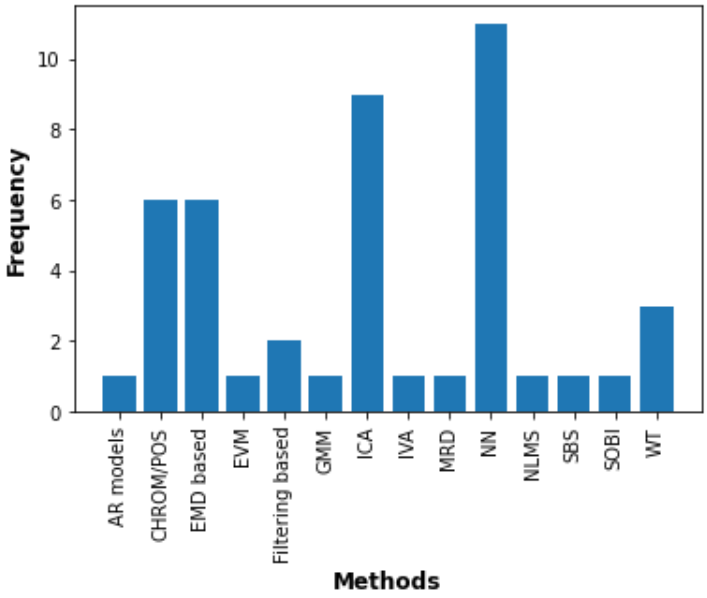

Fig. 8. Various estimation methods used for estimation of HR studies

5.2.4. Instruments used

5.2.4.1. Reference devices

Four reference devices have been used for comparing the estimated values with ground truth. We have only reported using reference devices for the self-created databases since reference devices information for the benchmark databases could be easily extracted from the respective articles. As mentioned before, 34 HR estimation studies have created their databases. 8/34 (23.52%) studies have used electrocardiogram as a reference device, 22/34 (64.70%) studies have used pulse oximeters as a reference device, while one study conducted by Wang et al. [23] used both reference devices. Few HR estimation studies used Arm Band HR monitor [31] and sphygmomanometer [7, 27]. On the other hand, SpO2 ground truth acquisition was carried out using a pulse oximeter only.

5.2.4.2. Camera characteristics

The distance between the subject's face and the camera is an essential parameter since a larger distance between face and camera deteriorates the strength of PPG signal information. Hence, it is necessary to identify a suitable shooting distance for a cleaner PPG signal. The shooting distance for HR and SpO2 estimations lies between 0.3 and 2 meters, respectively. While, the widely used shooting distances are 0.5 and 1.0 m, used by 11/50 (22%) and (13/50) (26 %), respectively. However, 4/50 (8%) studies used less than 0.5 m shooting distance, while 1.5 meters or greater for 10/50 (20%) studies. Furthermore, few studies have acquired the videos using more than one shooting distance. Specifically, the studies conducted by Song et al. [41] and Tran et al. [25] tested the effect of video shooting distances on HR estimation, whereas different shooting distances for

different activities were also used by Li et al. [42]. 12/50 (24%) HR estimation studies did not report camera shooting distances.

Camera resolutions also play an important role in accurate HR estimation by providing finer details from individual image frames, which are crucial to detect subtle color changes for extracting PPG information. Higher camera resolution provides more information but also needs intense computations. Hence, identifying a camera resolution with minimal information loss is a non-trivial task for accurate HR estimation. A diversified range of video camera resolutions has been employed to estimate accurate HR and SpO2 estimations. 19/45 (42.22%) HR estimation studies used cameras with a resolution of 640×480. In contrast, the twelve studies used 320×240, 1280×720, and 1920×1080, each employed by four HR studies.

A higher frame rate provides a larger number of contiguous images for a video, thereby providing more information to detect the blood volume pulse from raw RGB signal traces. Similar to resolution, a higher framerate has more computational requirements. Hence, it is necessary to use a framerate that ensures minimal loss and provides a noise-free PPG signal. Except for one, by Song et al. [36], all estimation studies have reported the frame rates for video acquisition with a range of 12-120 frames per second (fps). 29/50 (58%) estimation studies used 30 frames per second (fps) for video acquisition. Other frame rates used by estimation studies were 15 fps [11, 12, 18, 19, 57], 20 fps [22-24, 37, 38], 25 fps [22-24, 38, 59]. However, numerous studies have also gathered the video samples at a higher sampling rate, for instance, 50 fps [56], 60 fps [30], 100 fps [56].

5.2.5. Clinical studies

Although most estimation studies were conducted on healthy individuals, three clinical studies [51, 56, 62] have been included using the search strategy mentioned in Table 1. Most importantly, these studies have estimated two or more physiological parameters. Yu et al. [56] conducted a study on geriatric patients which aimed at estimating heart rate and heart rate variability, while the study undertaken by Tarassenko et al. [51] estimated heart rate, SpO2, and breathing rate of the patients undergoing dialysis. The study conducted by Bal [62] aimed at estimating the heart rate and SpO2 in the pediatric intensive care limit.

5.2.6. Performance metrics

5.2.6.1. HR estimation studies

The performance analysis for HR estimation studies utilized five metrics, namely, mean and Standard Deviation error (SD), Root Mean Square Error (RMSE), Mean of Error-Rate percentage (MER), signal to noise (SNR) ratio, and correlation. Among all of them, the majority of studies used RMSE and correlation. The mean and standard deviations of all metrics are given in table 6. Few studies have tested their estimation algorithms under different application scenarios or using multiple databases wherein average RMSE or correlation values were calculated for the analysis. Additionally, accuracy and Bland-Altman analysis were also included. The details of all metrics are found in the supplementary file (Table 6-Table 12). 25/45 (55.55 %) studies reported RMSE, out of

which 10/25 (40%) achieved the RMSE within ±5 bpm, whereas the remaining studies have the mean RMSE 2.73 bpm and standard deviation of 1.32 bpm. A box-whisker plot in Fig. 9 (left) depicts the error distribution of error metrics. Higher RMSE values for the two studies corresponds to testing the proposed methods under challenging conditions such as fitness exercise and cold pressure test (subjects were asked to put hands in very cold water), while collecting the video and the ground truth HR values.

Pearson correlation values have been reported in 26/45 (57.77%) of the HR estimation studies. Among them, 15/26 (57.69%) [7, 9, 11, 16, 18, 22, 27, 29, 30, 38, 40-42, 52, 62] studies achieved the correlation value of 0.90 or more, while the average correlation value for the remaining studies is 0.77 with the standard deviation of 0.093.

10/45 (22.22%) [7, 14, 17, 21, 25, 31, 40, 42, 45, 62] studies have reported accuracy, which is, calculated as the percentage of samples with the error difference ±5 bpm between ground truth and estimated HR values. This metric is used to justify the clinical relevance since the clinically accepted error between reference device measurement and estimation is ±5bpm [67]. The Pearson correlation values and accuracy distribution is shown in Fig. 9 (right).

Table 6. Performance Metrics Statistics

Metric	Number of Studies Used	Mean ± SD
Mean Error	9	0.57 ± 1.49
Standard Deviation	17	4.91 ± 2.41
RMSE	28	5.15 ± 3.24
MER	8	6.03 ± 1.26
Correlation	26	0.88 ± 0.11
Signal to Noise ratio	5	3.17±1.75

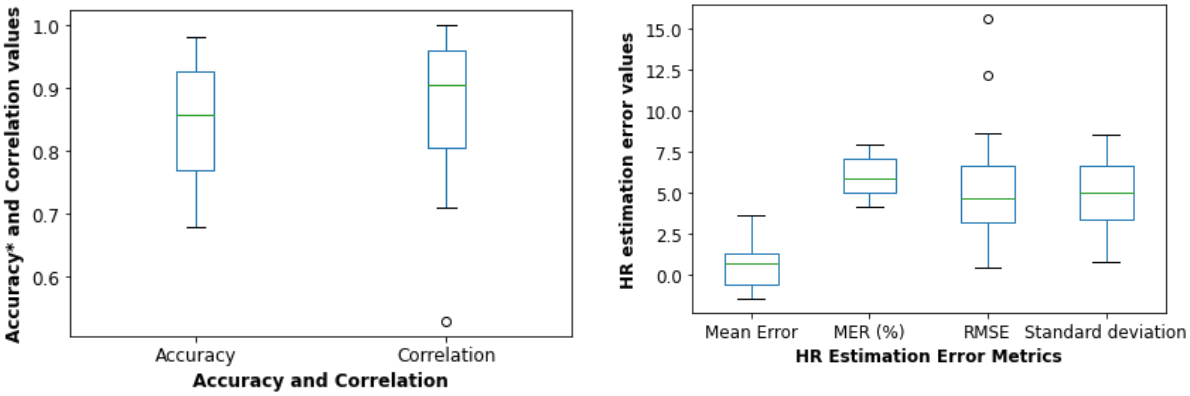

Fig. 9. Error metrics distribution of HR estimation studies.

23/45 (51.11%) studies have included B-A plots in their analysis. Additionally, one study by Lin [47] did not present the B-A plot, rather the level of agreements. Fig. 10. (left) depicts the mean bias and upper and lower level of agreement for HR estimation studies in chronological order. 8/23 (34.78%) studies achieved the mean difference within the clinically accepted range, while the rest may need significant improvements in the future.

5.2.6.2. SpO2 estimation studies

As mentioned before, 7/8 (87.5%) non-contact estimation studies used regression analysis for SpO2 (with range 80-100%) calculations, utilizing the ROR using two wavelength light intensities. 5/8 (62.5%) [12, 57-59, 61] studies have reported R^2 values. Furthermore, the root mean squared metric (A-rms %) was calculated for two studies [12, 60], while the same number of studies [59, 62] have used Pearson correlation value to test the algorithm's performance. 3/5 (60%) studies [57-59] have achieved an R^2 value of 0.8 or more, while the remaining two studies achieved relatively lower values, 0.65 and 0.58, respectively. Overall, the mean R^2 value is 0.78, with a standard deviation of 0.14. Furthermore, 5/8 (62.5%) studies have used B-A plots to showcase the performance of the proposed methods, as depicted in Fig. 10 (right).

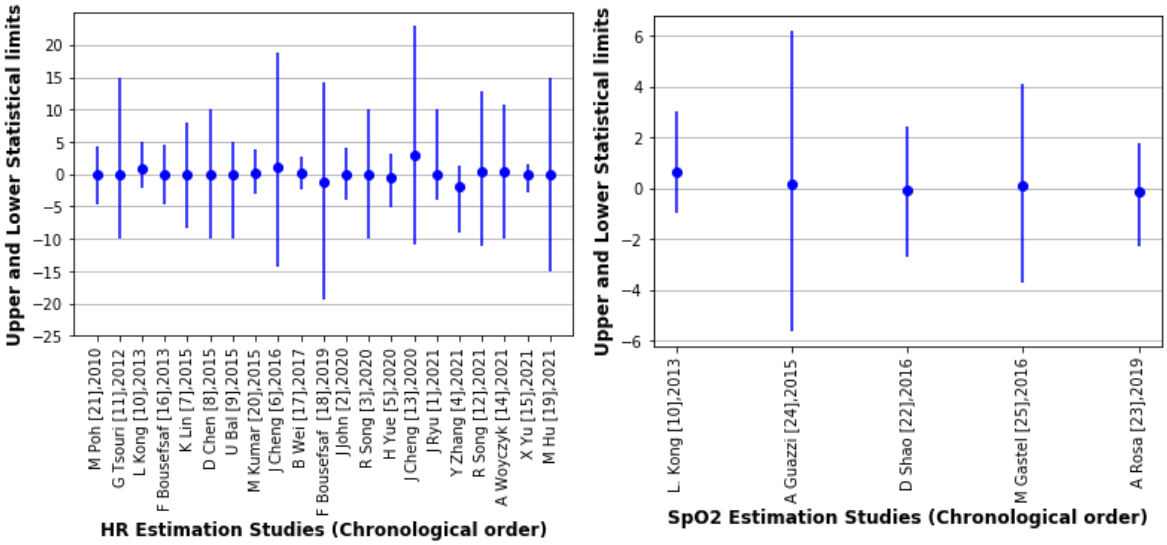

Fig. 10. A summary of Bland-Altman analysis for HR (left) and SpO2 (right) estimation studies.

5.2.7. Challenges

Non-contact methods for HR estimation studies deal with three types of noise: camera quantization, motion, and illumination noise. Almost all the studies have assumed a constant light illumination incident on every region of the face, which does not comply with the real-time situations. Secondly, the low strength of the PPG signal compared to the noise due to motion and illumination artifacts poses a significant challenge in extracting the cleaner PPG signal. Low resolution and camera shooting distance further degrade the quality of the acquired PPG signal.

5.2.8. Studies quality assessment Results

We have identified seven vital parameters for HR estimation studies, namely: camera characteristics (camera resolution and shooting distances), Bland-Altman analysis, results score performance metrics (RMSE and correlation), artifacts, accuracy (error < +/- 5 bpm), number of subjects used for the study, and inclusion/exclusion of ethnicity. On the other hand, SpO2 estimation studies quality was assessed using the following four parameters: camera characteristics, number of subjects, inclusion/exclusion of Bland-Altman analysis, and coefficient of determination (R^2). Other parameters similar to HR

studies such as artifacts, accuracy, RMSE, correlation, and ethnicity were not reported in most studies, hence not included in this analysis.

The studies were categorized into three categories: strong, fair, and weak, as depicted in Fig. 11. Based on the proposed protocol, 5 HR studies were identified as "strong" reporting maximum specified parameters, while the number of "fair" and "weak" studies were found to be 30 and 11, respectively. On the other hand, 3 SpO2 studies are categorized as weak, two as fair, and three as strong.

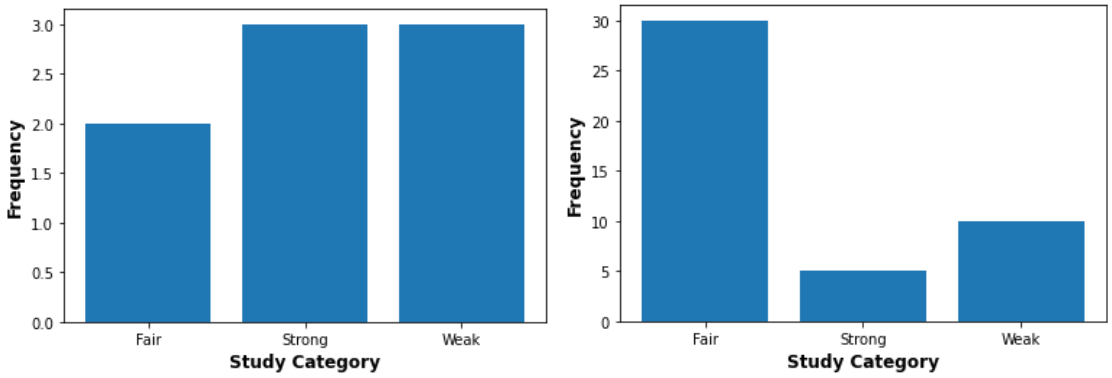

Fig. 11. Study categorization results a) HR and b)SpO

5.3. Motion and Illumination Resistant Facial Video based Heart Rate Estimation Method using Levenberg-Marquardt Algorithm Optimized Undercomplete Independent Component Analysis

This method was tested under constrained and natural conditions using three benchmark databases: VIPL-HR, UBFC-rPPG, and COHFACE. VIPL-HR database was used for performance validation under constrained conditions, UBFC-rPPG tested the method's performance for rigid and non-rigid motions, and illumination variations effect on the proposed method was tested using COHFACE database. It is necessary to analyze the performance under-constrained and unconstrained conditions since testing a method under constrained conditions gives insight into its steps and precision, whereas unconstrained conditions test its robustness. The detailed description of the databases used for this study is presented in the following subsections, with a summary in Table 7.

5.3.1. Databases

5.3.1.1. VIPL-HR

The database consists of 2378 videos with visible light spectra and 752 videos with Near Infrared (NIR) spectra from 107 subjects (79 males and 28 females), aged between 22 and 41 years. The database did not provide age-specific information. Nine variable scenarios were considered for sample collection. For each scenario, the samples were collected using digital cameras of different frame rates and NIR cameras. Each database sample comprises a 30 second (s) subject video, BVP signal, HR, and SpO2 values [65]. This study has used the subset of videos that corresponds to a frame rate of 30 frames per second (fps) with 1920 × 1080 pixels resolution, covering the HR range between 47 and

100 beats per minute (bpm). The ground truth heart rate was acquired using a CMS60C pulse oximeter synchronized with the subject's video. This resulted in 107 samples (one from each individual) that were analyzed for testing Levenberg-Marquardt Algorithm optimized Undercomplete ICA (U-LMA). One sample (p41) was left out due to insufficient video length (18 seconds) smaller than the processing window (25 seconds).

5.3.1.2. UBFC-rPPG

UBFC-rPPG is a publicly available database consisting of 50 video samples, synchronized with a CMS50E pulse oximeter (sampling rate 60 Hz). The videos are available in uncompressed form with a resolution of 640 × 480 pixels at 30 frames per second, covering the HR range between 63 and 112 bpm. Each video is 2 minutes long in which participants were asked to sit facing the camera and play a mathematical game that causes abrupt rise and fall in HR value promoting rigid and non-rigid movements [41]. The database did not provide age-specific information. All videos were used to test the performance of the proposed method [68].

5.3.1.3. COHFACE Database

COHFACE dataset is a collection of 160 videos with physiological recordings for HR and respiration rate from 40 healthy subjects with a mean age of 35.6 years. The dataset is composed of 60 seconds videos from 12 females and 28 males covering the HR range between 54 to 97 bpm. The videos were recorded with a resolution of 640 × 480 pixels at 20 fps with the synchronized BVP measurements using BVP model SA9308M, belt model SA9311M (sampling rate 256 Hz) [69]. The dataset offers constrained and challenging natural conditions, especially in terms of illumination variations over the facial region. Therefore, this study tests the performance of the proposed method using natural conditions video samples.

Table 7. Database Summary used for this study

Database Features	**VIPL-HR**	**UBFC-rPPG**	**COHFACE**
No. of subjects	107	50	40
Video Resolution	1920 X 1080	640 X 480	640 X 480
Frame rate	30	30	20
Video Duration	30 seconds	90 seconds	60 seconds
Ground truth sensor	CMS60C	CMS50E	SA9308, SA9311M
Shooting Distance	1 meter	1 meter	-
HR range	47-100 bpm	63-112 bpm	54-97 bpm
Considered Artifacts	Constrained*	Rigid and Non-rigid motion	Illumination

*Constrained: Conditions with minimum permissible motion and illumination.

5.3.2. ROI selection and Signal construction

Before this step, the RGB image frames of the video were preprocessed for adjusting the pixel intensities, using gamma correction. ROI selection deals with face detection followed by segmenting the skin in the YCbCr color space in which Y represents the luminance with pixel intensity ranges between 16 and 235, while for chrominance blue (Cb) and chrominance red (Cr) components, the pixel values lie between 16 and 240. The thresholds used for Cb and Cr components are in the range of 77 to 127 and 133 to 173, respectively, with no thresholding for the luminance component [72]. Finally, the ROI is selected as 70% height and 60% width of the segmented skin region. Figure 12 depicts the results of the face detection and skin segmentation process. The regularization parameter was set to an empirically defined value for de-trending the temporal RGB traces, i.e., 10. The raw signal was constructed using a moving window operation with a 96% overlap (1-sec increment) for each color channel.

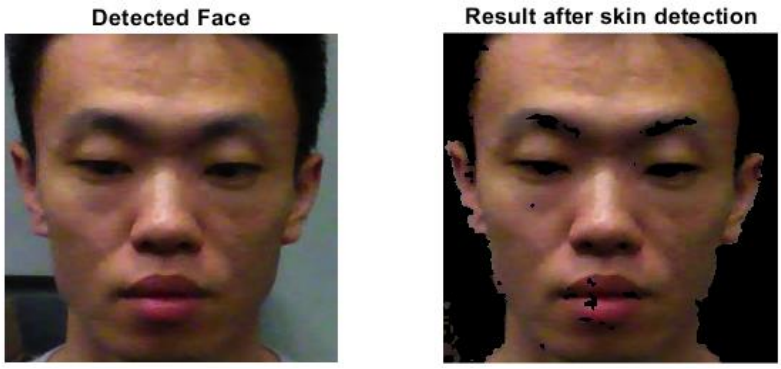

Fig 12. Face detection and skin segmentation

5.3.3. BVP Signal Extraction and HR estimation

BVP signal extraction was performed using U-LMA. The unmixing matrix W was first initialized randomly, and the values of damping parameter λ were set empirically as 5 and 2.5, respectively, as a part of standard LMA initialization. Subsequently, the customized LMA was employed to maximize the entropy of the proposed non-linear CDF optimization function using 1000 iterative steps because none of the video samples took this many iterations for convergence to global maxima. Finally, the optimized unmixing matrix W was used to extract the BVP signal (after bandpass filtering). Finally, Fast Fourier Transform (FFT) was applied to the resultant signal, then calculated the $\log_{10}$ value of peak maxima and multiplying it with 60 for mean heart rate estimation.

5.3.4. Performance Analysis

As mentioned before, the performance of the proposed U-LMA method is analyzed considering three scenarios: constrained, Rigid and non-rigid motions, and illumination variations. Table 7 specifies that the VIPL-HR database has been used for performance testing under the constrained or stable scenario, UBFC-rPPG for testing its robustness in rigid and non-rigid motions and, COHFACE in illumination variations scenarios. Bland-Altman and regression plots will be presented and analyzed for each scenario, considering the respective measured parameters for the plots.

5.3.5.1. Constrained Scenario

For the constrained (VIPL-HR database) scenario, the subjects were asked to sit in the still position at a distance of one meter away from the camera with the ceiling lamp switched on. The Bland-Altman and regression plot for the constrained scenario are shown in Fig 13. The mean bias for the proposed method is -0.35 bpm which is near to zero error difference between ground truth and estimated values. In other words, on average, the heart rate calculated by the algorithm measures 0.35 bpm less than the conventional BVP sensor used.

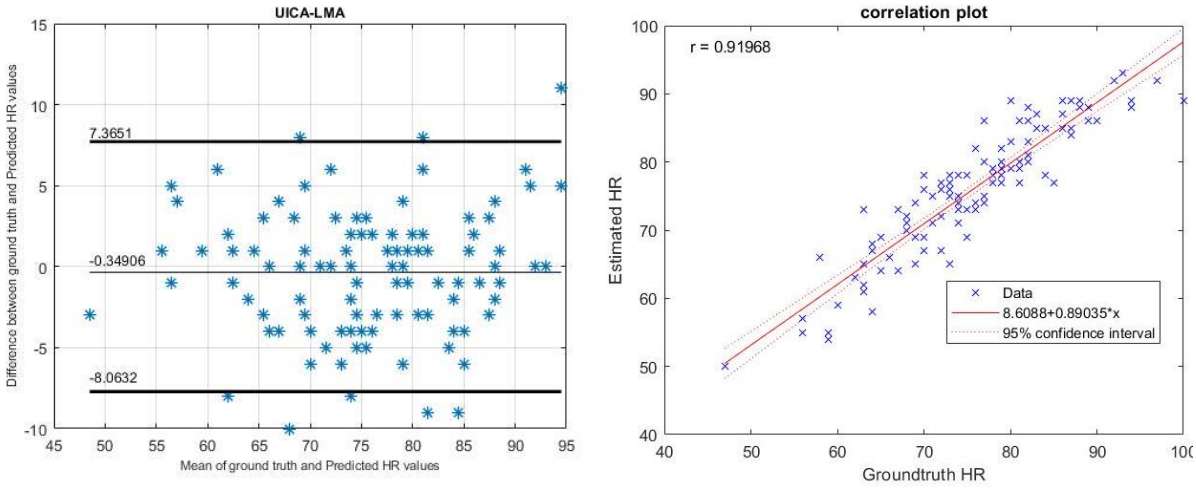

Fig. 13. Bland-Altman Plot and regression plots for the constrained scenario.

Furthermore, most of the differences lie within upper (7.3651) and lower (-8.0632) level statistical limits, which justifies the good performance of the method. Additionally, the Pearson correlation value denoted by *r* for this scenario is 0.92, confirming a higher correlation between ground truth and estimated values. Therefore, Bland-Altman's analysis and high correlation value justify the excellent performance of the proposed method under the constrained scenario.

5.3.5.2. Rigid and Non-rigid motions

Performance analysis under this scenario was performed using all video samples of the UBFC-rPPG database. The videos were collected while subjects were playing a time-sensitive mathematical game which causes an abrupt increase or decrease in HR values along with involuntary head movements due to the subject's action. The samples also have a certain amount of illumination variations since the video samples were collected considering natural conditions. Fig 14 depicts the Bland-Altman, and regression plots. As expected, the mean bias for this scenario is 1.84 bpm due to the presence of motion artifacts which means, the U-LMA predicts 1.84 bpm more than the traditional BVP. Consequently, the upper statistical limit of the Bland-Altman plot is slightly greater than 10.00 bpm; however, the method achieved a far lower statistical limit -6.6005. All of the data points lie between the upper and lower statistical limits. Interestingly, the ground truth and estimated HR values showed a very high correlation (0.94), despite having a higher overall mean difference. Hence, the Bland-Altman analysis and regression plot confirm the proposed method's effectiveness under challenging motion conditions and

handling abrupt rise and fall of HR values, considering the mean of the HR values during the interval.

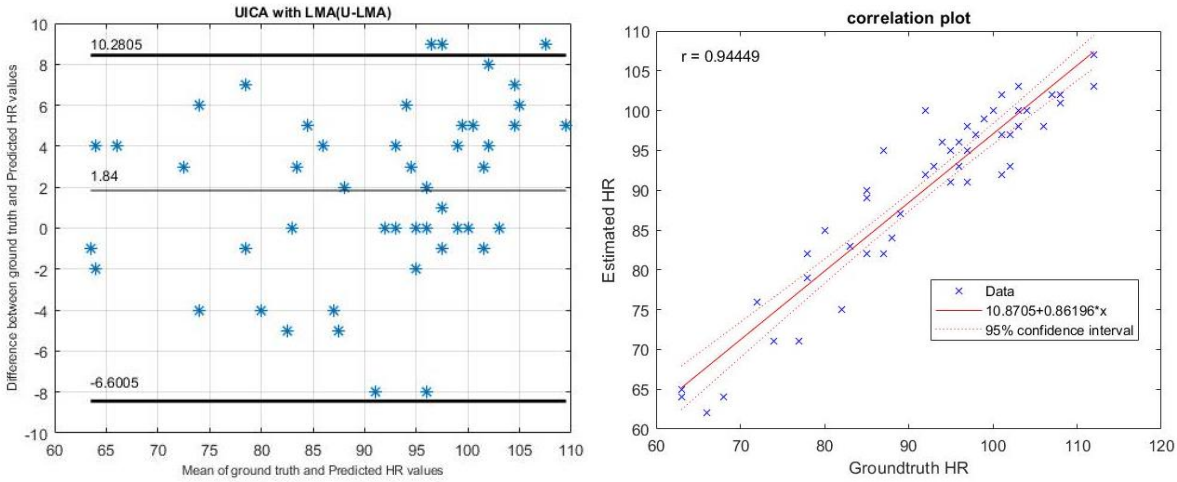

Fig 14. Bland-Altman Plot for Rigid and Non-rigid motion scenario.

5.3.5.3. Illumination Variations Scenario

COHFACE database is utilized to assess the ability of U-LMA under different illumination scenarios. It is worth noting that the samples for the database have motion artifacts too, but with the predominance of uneven illumination distribution over the face due to ambient light. The performance analysis using Bland-Altman and regression plot is presented in Fig 15. The mean bias achieved with the illumination scenario is -0.85 with lower and upper statistical limits -9.7546 and 8.0546, respectively. Similar to the other scenarios, almost all the differences lie within the statistical limits. Furthermore, the Pearson correlation achieved under this scenario is 0.92, confirming a good correlation between estimated and ground truth values. Both plots and their measured parameters prove the efficiency of U-LMA for HR estimation using illumination variant facial videos.

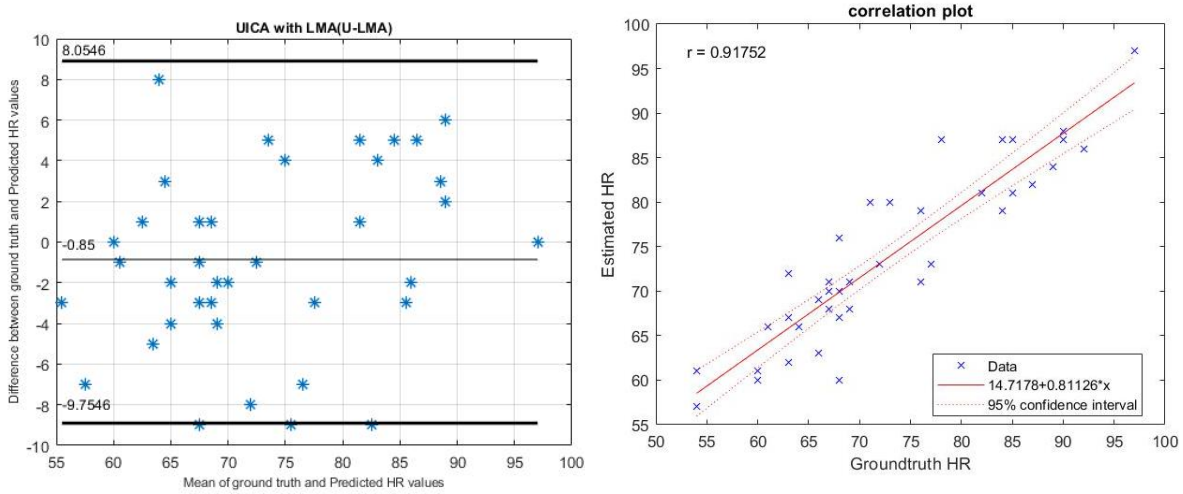

Fig 15. Bland-Altman and regression plot for Illumination scenario.

5.3.5. Comparative analysis

The available related conventional rPPG methods in the literature are based on single color channel selection, ICA, color subspace transformations, and Wavelet based methods. U-LMA's performance was compared with all, except Wavelet based methods, since these methods use the time-frequency domain and empirically set coefficients, unlike other PPG methods included in the study.

The single color channel selection method deals with utilizing a filtered signal extracted from a single color channel. Therefore, GREEN proposed by Verkrysse et al. [4] was included, which extracts the BVP signal from the green color channel of the RGB color space. Most ICA based rPPG methods use ICA-Poh (JADE) [2, 28, 37, 38] and FastICA [9, 12, 52, 53]; hence they were included in this analysis. JADE uses kurtosis, whereas FastICA uses a negentropy based optimization function for unmixing matrix estimation. Two color subspace transformations CHROM [10] and POS [14], were also included in the analysis due to their dependence on optimization procedures like ICA based methods. CHROM is a motion intolerant algorithm, while POS works better for uneven illumination variations. BCG method [54] was also included since it works on tracking the periodic movements of the head. As the effect of both types of motion on the proposed method is also tested in this study, it is worth including BCG as one of the state-of-art methods for performance comparison. Finally, the method presented by Song et al. [21], which is a combination of the color subspace transformation method and KDICA proposed by Ajou Chen [55], was also used to test the performance of the proposed method. The KDICA uses a Laplacian kernel for kernel density estimations which needs pulse and artifacts spectrum to be in antiphase.

Furthermore, to assess the performance of the proposed non-linear cumulative density function approximated by tanh and a customized LMA, another variant of the proposed undercomplete analysis (U-neg) is also introduced, which utilizes the differential entropy or negentropy as an objective function. This objective function is optimized using the standard ICA procedure given by Hyvärinen et al. [44].

GREEN, ICA-Poh, CHROM, POS, and BCG were implemented using the standard implementation included in the iPhys toolbox by Mc Duff et al. [56]. Furthermore, the MATLAB implemented versions of FastICA by Hyvärinen et al. [44] and KDICA by Ajoy Chen [55] were used to simulate the respective HR estimation methods as mentioned above, keeping other steps (ROI selection, bandpass filtering, and FFT) identical to the proposed U-LMA. All methods were tested under three scenarios, as explained in previous subsections. In other words, the video samples from all three databases were tested for all the methods used for comparative analysis by calculating RMSE, MAPE, mean error, standard deviation, accuracy, and Pearson correlation values under 0.01 significance level (alpha).

5.3.6.1. Constrained scenario

Similar to section 5.3.5.1, the videos from the VIPL-HR database were used for comparative analysis. The performance metrics for the comparative analysis are presented in table 8. Among all methods, BCG was the worst performing method for this scenario,

despite minimal motion and illumination variation artifacts. BCG is susceptible to poorly perform in the presence of involuntary head movements, which may lead to false identification of face tracking points for estimation [54]. CHROM, GREEN, ICA-Poh, and CHROM methods exhibited almost the same performance. Poor performance of GREEN is obvious due to inappropriate method formulations and performance validations [28]. Furthermore, the original study suggested that along with green, the red and blue channel also contains complimentary PPG information [4], which is also confirmed in this study. The poor performance of ICA-Poh is due to lower framerate as compared to ground truth sensor value since this leads to an inappropriate mapping of BVP peaks, hence inaccurate interbeat intervals for HR calculation [2]. CHROM's performance greatly depends on its alpha tuning procedure which works better for different magnitudes of specular distortions and pulse signals. The noise due to involuntary movements is inevitable, which might have degraded the performance [21]. Like CHROM, the POS method's accuracy significantly depends on its alpha tuning procedure which is suboptimal in the case of similar specular and pulse components magnitude. In this case, the specular variation components projected on the two axis may not be in absolute antiphase due to the presence of noise, leading to false estimations of alpha, hence poor BVP signal extraction. U-neg has performed relatively better than the above-mentioned state-of-the-art methods due to effective information gathering from red, green and blue color channels but failed to suppress the effect of inevitable noise.

Table 8. Performance metrics for the methods under Constrained Scenario.

Methods	RMSE (bpm)	MAPE (%)	Std. Dev. (bpm)	Mean (bpm)	Accuracy (%) (<=5 bpm)	r*
Green	21.48	24.74	12.32	17.63	1.89	0.23
ICA - Poh	19.15	22.37	11.70	15.21	2.83	0.24
CHROM	16.16	19.19	15.67	4.20	9.43	0.22
POS	18.80	22.30	11.94	14.56	0.94	0.31
BCG	24.87	27.79	12.82	21.35	7.55	-0.04
KernelICA	13.15	14.93	11.20	6.97	18.69	0.51
FastICA	13.10	15.37	10.38	8.06	19.62	0.60
UICA without LMA (U-neg)	17.03	17.21	13.84	10.01	23.58	0.47
UICA with LMA(U-LMA)	**3.85**	**4.07**	**3.86**	**-0.35**	**84.91**	**0.92**

*r-Spearman correlation is calculated at the 0.01 significance level.

On the other hand, Kernel ICA and FastICA have performed considerably well than other methods and U-neg. However, the performance of KernelICA was suffered due to the same reason as CHROM. However, FastICA has shown better performance than other methods except for U-LMA, which once again proved the effectiveness of negentropy based optimization function by ensuring statistical independence among independent components. On the other hand, the proposed U-LMA achieved the best results justifying its performance due to its ability to use higher order statistics for processing non-linear

signals and effective optimization procedure using LMA. Moreover, the highest accuracy with clinically accepted error difference was also achieved by U-LMA.

5.3.6.2. Motion Scenario

The video samples from the UBFC-rPPG database were used to assess the effect of rigid and non-rigid motions on HR estimation. Table 9 presents the performance metrics for all the compared methods. Overall, all methods performed well under motion scenario due to uncompressed videos. Similar to the constrained scenario, BCG performed worst for the motion scenario too. BCG method's performance was suboptimal due to the presence of rigid and non-periodic head movements [54]. Better performance of the GREEN method indicates that the method is effective for uncompressed videos and also data-driven. ICA-Poh performed relatively well due to the accurate selection of the BVP signal since there was no loss of information from the videos. Interestingly, the statistical independence among components suffered due to similarity of motion and pulse spectra under motion scenario, which led to the almost similar performance of ICA-Poh and FastICA. Furthermore, KernelICA and U-neg also showed similar performance for different reasons; KernelICA uses motion intolerant chrominance signals followed by KDICA, whereas U-neg uses a negentropy based function for unmixing matrix estimation using undercomplete ICA, combining PPG information from all color channels. Although RMSE, MAPE, mean error, and error standard deviation of U-neg was reduced, the accuracy was degraded in the motion scenario, as expected. CHROM and POS performed relatively better than all the methods except U-LMA. This is due to their ability to perform well under motion scenarios due to extraction motion resistant signals followed by the alpha tuning procedure. Nevertheless, the proposed U-LMA outperformed all the methods, reporting the minimum value of errors, highest accuracy, and Pearson correlation, justifying its best performance and clinical relevance.

Table 9. Performance metrics of the methods under rigid and non-rigid motion scenario

Methods	RMSE (bpm)	MAPE (%)	Std. Dev. (bpm)	Mean (bpm)	Accuracy (%) (<=5 bpm)	r*
Green	28.08	20.26	25.20	12.90	44	0.34
ICA – Poh	20.49	13.34	19.56	6.72	58	0.54
CHROM	14.08	9.00	13.16	-5.35	66	0.70
POS	14.27	9.37	13.51	-4.98	62	0.71
BCG	36.08	33.75	16.90	31.97	8	0.03
KernelICA	20.40	14.67	18.90	8.12	46	0.59
FastICA	20.36	14.92	19.63	6.10	46	0.56
UICA without LMA (U-neg)	14.97	12.86	11.27	9.98	22	0.59
UICA with LMA (U-LMA)	**4.57**	**4.00**	**4.22**	**1.84**	**78**	**0.94**

*r-Spearman correlation is calculated at the 0.01 significance level.

5.3.6.3. Illumination Variation Scenario

The effect of illumination variations on the methods used for this study was evaluated using the COHFACE database. GREEN method has shown a negative correlation for this scenario, indicating their susceptibility for uneven illumination distribution. GREEN method is susceptible to the illumination variation artifacts due to varying light intensity distribution [4]. ICA-Poh did not perform well due to the low frame rate of the videos, as explained in the study conducted by Poh et al. [2]. POS performance was suboptimal due to heterogeneous illumination conditions due to its assumption of independent intensity variations [14]. On the other hand, CHROM and BCG performed better than these three methods in terms of accuracy. However, these methods could not perform well due to the susceptibility of BCG for illumination variations [54] and considerable larger differences between actual and estimated specular distortions in the video for CHROM [10]. Furthermore, the other ICA based methods KernelICA and FastICA performed relatively better than the methods mentioned above. However, the degraded performance of the KernelICA is due to the same reason as the CHROM method, along with the inability of kernel density based ICA to perform under a higher degree of illumination distortion. Specifically, the KDICA used laplacian Kernel, which did not work due to illumination and pulse spectra overlapping. FastICA performed better than KernelICA due to the implication of a statistically better optimization function to separate specular and PPG information. On the other hand, U-neg achieved better results than other state-of-the-art methods depicting the significance of undercomplete ICA and negentropy. Furthermore, U-LMA achieved the lowest error values and highest accuracy and correlation values showing its superiority for HR estimation under illumination variations scenario. Table 10 presents the performance of methods under illumination variations scenario.

Table 10. Performance metrics of the methods under illumination variations scenario.

Methods	RMSE (bpm)	MAPE (%)	Std. Dev. (bpm)	Mean (bpm)	Accuracy (%) (<=5 bpm)	r*
Green	22.76	26.93	20.39	10.61	22.50	-0.07
ICA – Poh	24.88	28.52	17.84	17.56	35.00	0.06
CHROM	19.49	22.48	10.54	16.47	35.00	0.25
POS	26.28	30.71	12.63	23.14	22.50	0.05
BCG	25.53	27.00	15.72	20.27	2.50	0.19
KernelICA	14.50	12.61	11.60	8.90	50	0.36
FastICA	15.76	14.89	10.68	11.71	45	0.41
UICA without LMA(U-neg)	11.13	13.03	10.76	3.33	25	0.60
UICA with LMA(U-LMA)	**4.48**	**5.16**	**4.45**	**-0.85**	**80**	**0.92**

*r-Spearman correlation is calculated at the 0.01 significance level.

5.3.6.4. RMSE analysis

The RMSE for the HR estimation has been predominantly analyzed in most of the studies conducted so far. It is calculated as the square root of the averaged squared error differences among different samples, providing the overall distribution of errors. Figure 16 depicts the box and whisker plot of RMSE for analyzing the RMSE distribution among methods based on databases and vice-versa. These RMSE plots provide a deep insight into the performance of state-of-the-art and proposed methods for all the databases used in the study.

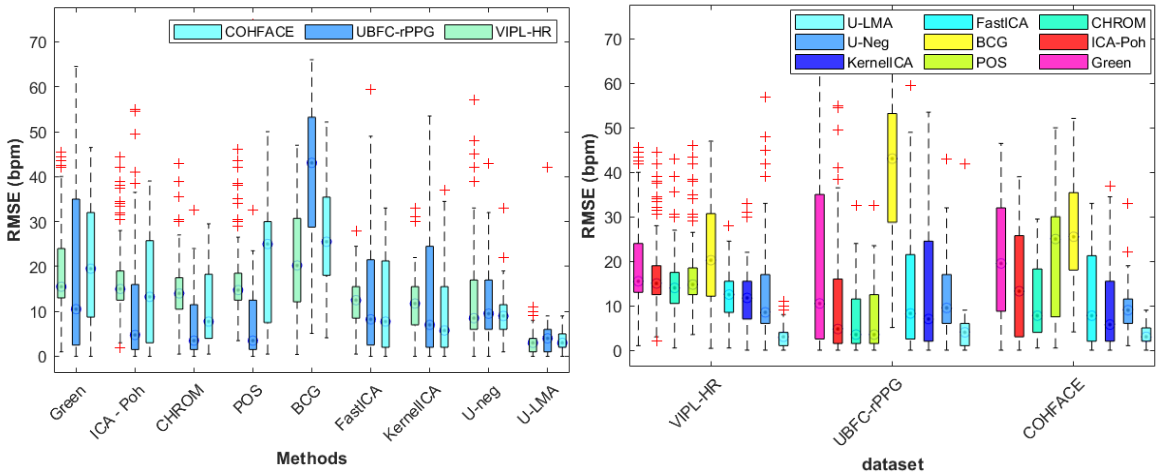

Fig 16. RMSE Box and whisker plot for the methods and databases used.

UBFC –rPPG database was challenging for all methods used in the study due to its realistic conditions considered during video acquisition, whereas the performance with VIPL-HR is the best due to constrained conditions. The COHFACE database is also challenging in terms of illumination variations throughout the samples. All methods performed poorly under the motion scenario. However, CHROM, POS, and U-neg performed better in this scenario, depicting their ability to deal with different types of motions. Among all, the worst performing method is BCG, as depicted in the RMSE plots in figure 9. BCG was unable to cope up with the inevitable color distortions due to involuntary motions and illuminations var as mentioned in the original study [54]. The RMSE of the GREEN method was also very high since the study did not use any formulation for BVP signal extraction. The performance of all three ICA based state-of-the-art methods was similar, despite different objective functions used for unmixing matrix estimations. All methods suffered from the permutation problem, which makes it challenging to choose the appropriate BVP component and discarding other components simultaneously. Like ICA based methods, color subspace transformation methods CHROM and POS also exhibited similar performance except for the illumination variations scenario in which the POS method failed to perform well. Since the videos were not recorded using an external source of light, causing serious illumination variations on different facial regions producing an effect similar to multiple light sources (e.g., entirely black on one side and bright on the other side). U-neg performed relatively better than any state-of-art method, except motion scenario, in which CHROM, POS, achieved better performance than U-neg. It is also because of the uncompressed version

of videos which ensured detailed subtle color change information required to extract BVP components. On the other hand, U-LMA performed far better than U-neg and other state-of-the-art methods either in terms of databases or state-of-art methods comparison. The proposed methods performed relatively better due to the proposed undercomplete ICA, which ensures better BVP information extraction from all three channels of RGB color space. Furthermore, U-neg used negentropy (differential entropy) for optimizing W with standard ICA implementation, but the experiments conducted during the study revealed that the entropy of CDF approximated by tanh yielded better statistical independence than negentropy. Additionally, the lowest RMSE ranges by U-LMA when compared to U-neg in all scenarios were due to better optimization of unmixing matrix W using customized LMA proposed in this work.

6. Conclusion

This study focuses on exploring the potential of the RGB image color model for physiological parameters estimations under dark environments. An extensive review of existing studies was conducted to identify the research challenges associated with physiological parameters estimations. It was found that among all the available image models, the RGB color model possesses immense potential due to the relatively more robust pulsatile strength of the extracted PPG signal. However, some challenges are associated with the utilization of the RGB model, i.e., susceptibility towards different types of motions and illumination variations. Furthermore, illumination variation artifacts can be more severe in dark environments. Existing methods may not be able to perform well in dark scenarios. Considering these challenges, the study intends to find the answers to four research questions resulting from the comprehensive review. Conclusively, these research questions explore non-contact conventional and deep learning methods for HR and SpO2 estimations in the dark environment using identified performance metrics.

To date, a database consisting of videos acquired under a dark environment has been created to test the developed methods under a dark environment. Following the findings through the critical review of existing state-of-the-art studies, an Undercomplete ICA algorithm optimized by the Levenberg-Marquardt algorithm was proposed under ambient light conditions. Future work includes testing the performance of this method under dark conditions followed by improving it to work under these conditions using a self-created database collected under a dark environment.

Furthermore, deep learning will be explored to estimate physiological parameters estimations under the dark scenario. Specifically, novel deep learning methods will be developed for HR and SpO2 estimations under ambient conditions and modified to estimate these parameters under a dark environment. Specifically the deep learning models will be trained using video samples collected under ambient light conditions and testing them under a self-created (dark light environment) database. Finally, a performance analysis of deep learning and conventional methods under dark environments will be performed to identify the best clinically relevant methods to be tested under these conditions.

7. References

1. Rideout, V. and J. Beneken, *Parameter estimation applied to physiological systems.* Mathematics and Computers in Simulation, 1975. **17**(1): p. 23-36.
2. Mok, W.Q., W. Wang, and S.Y. Liaw, *Vital signs monitoring to detect patient deterioration: An integrative literature review.* International journal of nursing practice, 2015. **21**: p. 91-98.
3. Fairchild, K.D., et al., *Vital signs and their cross-correlation in sepsis and NEC: a study of 1,065 very-low-birth-weight infants in two NICUs.* Pediatric research, 2017. **81**(2): p. 315-321.
4. Moss, T.J., et al., *Signatures of subacute potentially catastrophic illness in the intensive care unit: model development and validation.* Critical care medicine, 2016. **44**(9): p. 1639.
5. Hassan, M.A., et al., *Heart rate estimation using facial video: A review.* Biomedical Signal Processing and Control, 2017. **38**: p. 346-360.
6. Kranjec, J., et al., *Non-contact heart rate and heart rate variability measurements: A review.* Biomedical signal processing and control, 2014. **13**: p. 102-112.
7. Lin, K., D. Chen, and W. Tsai, *Face-Based Heart Rate Signal Decomposition and Evaluation Using Multiple Linear Regression.* IEEE Sensors Journal, 2016. **16**(5): p. 1351-1360.
8. van der Kooij, K.M. and M. Naber, *An open-source remote heart rate imaging method with practical apparatus and algorithms.* Behavior research methods, 2019. **51**(5): p. 2106-2119.
9. Qiu, Y., et al., *EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video.* IEEE Transactions on Multimedia, 2019. **21**(7): p. 1778-1787.
10. Cheng, J., et al., *Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition.* IEEE journal of biomedical and health informatics, 2016. **21**(5): p. 1422-1433.
11. Poh, M.-Z., D.J. McDuff, and R.W. Picard, *Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.* Optics express, 2010. **18**(10): p. 10762-10774.
12. Moço, A. and W. Verkruysse, *Pulse oximetry based on photoplethysmography imaging with red and green light.* Journal of clinical monitoring and computing, 2021. **35**(1): p. 123-133.
13. Chen, W. and D. McDuff. *Deepphys: Video-based physiological measurement using convolutional attention networks*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
14. Kado, S., et al., *Spatial-Spectral-Temporal Fusion for Remote Heart Rate Estimation.* IEEE Sensors Journal, 2020. **20**(19): p. 11688-11697.
15. Huang, Y.-T., Y.-T. Peng, and W.-H. Liao. *Enhancing object detection in the dark using U-Net based restoration module*. in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019. IEEE.
16. Macwan, R., Y. Benezeth, and A. Mansouri, *Heart rate estimation using remote photoplethysmography with multi-objective optimization.* Biomedical Signal Processing and Control, 2019. **49**: p. 24-33.
17. Qi, H., et al., *Video-based human heart rate measurement using joint blind source separation.* Biomedical Signal Processing and Control, 2017. **31**: p. 309-320.

18. Poh, M.-Z., D.J. McDuff, and R.W. Picard, *Advancements in noncontact, multiparameter physiological measurements using a webcam.* IEEE transactions on biomedical engineering, 2010. **58**(1): p. 7-11.
19. Tsouri, G.R., et al., *Constrained independent component analysis approach to nonobtrusive pulse rate measurements.* J Biomed Opt, 2012. **17**(7): p. 077011.
20. Pursche, T., J. Krajewski, and R. Moeller. *Video-based heart rate measurement from human faces*. in *2012 IEEE International Conference on Consumer Electronics (ICCE)*. 2012. IEEE.
21. Cheng, J., et al., *Remote Heart Rate Measurement From Near-Infrared Videos Based on Joint Blind Source Separation With Delay-Coordinate Transformation.* IEEE Transactions on Instrumentation and Measurement, 2021. **70**: p. 1-13.
22. Haan, G.d. and V. Jeanne, *Robust Pulse Rate From Chrominance-Based rPPG.* IEEE Transactions on Biomedical Engineering, 2013. **60**(10): p. 2878-2886.
23. Wang, W., et al., *Algorithmic principles of remote PPG.* IEEE Transactions on Biomedical Engineering, 2016. **64**(7): p. 1479-1491.
24. Wang, W., et al., *Robust heart rate from fitness videos.* Physiological measurement, 2017. **38**(6): p. 1023.
25. Tran, Q., et al., *Adaptive Pulsatile Plane for Robust Noncontact Heart Rate Monitoring.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019: p. 1-13.
26. Yue, H., et al., *Non-contact heart rate detection by combining empirical mode decomposition and permutation entropy under non-cooperative face shake.* Neurocomputing, 2020. **392**: p. 142-152.
27. Chen, D., et al., *Image Sensor-Based Heart Rate Evaluation From Face Reflectance Using Hilbert–Huang Transform.* IEEE Sensors Journal, 2015. **15**(1): p. 618-627.
28. Song, R., et al., *Remote Photoplethysmography with An EEMD-MCCA Method Robust Against Spatially Uneven Illuminations.* IEEE Sensors Journal, 2021.
29. Zhang, Y., et al., *Illumination variation-resistant video-based heart rate monitoring using LAB color space.* Optics and Lasers in Engineering, 2021. **136**: p. 106328.
30. Yu, Z., et al. *Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
31. Wu, B.F., et al., *Neural Network Based Luminance Variation Resistant Remote-Photoplethysmography for Driver's Heart Rate Monitoring.* IEEE Access, 2019. **7**: p. 57210-57225.
32. Bousefsaf, F., A. Pruski, and C. Maaoui, *3D convolutional neural networks for remote pulse rate measurement and mapping from facial video.* Applied Sciences, 2019. **9**(20): p. 4364.
33. Niu, X., et al., *RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation.* IEEE Transactions on Image Processing, 2020. **29**: p. 2409-2423.
34. Hsu, G.-S.J., et al., *A deep learning framework for heart rate estimation from facial videos.* Neurocomputing, 2020. **417**: p. 155-166.
35. Yu, Z., et al., *AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching.* IEEE Signal Processing Letters, 2020. **27**: p. 1245-1249.

36. Song, R., et al., *Heart Rate Estimation From Facial Videos Using a Spatiotemporal Representation With Convolutional Neural Networks.* IEEE Transactions on Instrumentation and Measurement, 2020. **69**(10): p. 7411-7421.
37. Hu, M., et al., *Robust Heart Rate Estimation with Spatial-Temporal Attention Network from Facial Videos.* IEEE Transactions on Cognitive and Developmental Systems, 2021: p. 1-1.
38. Hu, M., et al., *ETA-rPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement.* IEEE Transactions on Instrumentation and Measurement, 2021. **70**: p. 1-12.
39. Zhao, C., et al., *Visual heart rate estimation and negative feedback control for fitness exercise.* Biomedical Signal Processing and Control, 2020. **56**: p. 101680.
40. Ryu, J., et al., *A measurement of illumination variation-resistant noncontact heart rate based on the combination of singular spectrum analysis and sub-band method.* Computer Methods and Programs in Biomedicine, 2021. **200**: p. 105824.
41. Song, R., et al., *New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method.* Computers in Biology and Medicine, 2020. **116**: p. 103535.
42. Li, X., et al. *Remote heart rate measurement from face videos under realistic situations*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
43. Shi, J. *Good features to track*. in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. 1994. IEEE.
44. John, J., S. Krishna, and R.R. Galigekere, *Automatic and Adaptive Signal- and Background-ROIs With Analytic-Representation-Based Processing for Robust Webcam-Based Heart-Rate Estimation.* IEEE Access, 2020. **8**: p. 34728-34736.
45. Gupta, P., B. Bhowmick, and A. Pal, *MOMBAT: Heart rate monitoring from face video using pulse modeling and Bayesian tracking.* Computers in Biology and Medicine, 2020. **121**: p. 103813.
46. Woyczyk, A., V. Fleischhauer, and S. Zaunseder, *Adaptive Gaussian Mixture Model Driven Level Set Segmentation for Remote Pulse Rate Detection.* IEEE Journal of Biomedical and Health Informatics, 2021. **25**(5): p. 1361-1372.
47. Lin, Y. and Y. Lin, *Step Count and Pulse Rate Detection Based on the Contactless Image Measurement Method.* IEEE Transactions on Multimedia, 2018. **20**(8): p. 2223-2231.
48. Zhang, C., et al., *Simultaneous detection of blink and heart rate using multi-channel ICA from smart phone videos.* Biomedical Signal Processing and Control, 2017. **33**: p. 189-200.
49. Verkruysse, W., L.O. Svaasand, and J.S. Nelson, *Remote plethysmographic imaging using ambient light.* Optics express, 2008. **16**(26): p. 21434-21445.
50. Bousefsaf, F., C. Maaoui, and A. Pruski, *Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate.* Biomedical Signal Processing and Control, 2013. **8**(6): p. 568-574.
51. Tarassenko, L., et al., *Non-contact video-based vital sign monitoring using ambient light and auto-regressive models.* Physiological measurement, 2014. **35**(5): p. 807.
52. Wei, B., et al., *Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions.* Biomedical engineering online, 2017. **16**(1): p. 1-21.

53. Gupta, O., D. McDuff, and R. Raskar. *Real-time physiological measurement and visualization using a synchronized multi-camera system*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016.
54. Kumar, M., A. Veeraraghavan, and A. Sabharwal, *DistancePPG: Robust non-contact vital signs monitoring using a camera.* Biomedical optics express, 2015. **6**(5): p. 1565-1588.
55. Fischler, M.A. and R.C. Bolles, *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. Communications of the ACM, 1981. **24**(6): p. 381-395.
56. Yu, X., et al., *Noncontact Monitoring of Heart Rate and Heart Rate Variability in Geriatric Patients Using Photoplethysmography Imaging.* IEEE Journal of Biomedical and Health Informatics, 2021. **25**(5): p. 1781-1792.
57. Van Gastel, M., S. Stuijk, and G. De Haan, *New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring.* Scientific reports, 2016. **6**(1): p. 1-16.
58. Guazzi, A.R., et al., *Non-contact measurement of oxygen saturation with an RGB camera.* Biomedical optics express, 2015. **6**(9): p. 3320-3338.
59. Shao, D., et al., *Noncontact Monitoring of Blood Oxygen Saturation Using Camera and Dual-Wavelength Imaging System.* IEEE Transactions on Biomedical Engineering, 2016. **63**(6): p. 1091-1098.
60. Rosa, A.d.F.G. and R.C. Betini, *Noncontact SpO 2 Measurement Using Eulerian Video Magnification.* IEEE Transactions on Instrumentation and Measurement, 2019. **69**(5): p. 2120-2130.
61. Kong, L., et al., *Non-contact detection of oxygen saturation based on visible light imaging device using ambient light.* Optics express, 2013. **21**(15): p. 17464-17471.
62. Bal, U., *Non-contact estimation of heart rate and oxygen saturation using ambient light.* Biomed Opt Express, 2015. **6**(1): p. 86-97.
63. Blackford, E.B. and J.R. Estepp. *Effects of frame rate and image resolution on pulse rate measured using multiple camera imaging photoplethysmography*. in *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging*. 2015. International Society for Optics and Photonics.
64. Stricker, R., S. Müller, and H.-M. Gross. *Non-contact video-based pulse rate measurement on a mobile service robot*. in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 2014. IEEE.
65. Niu, X., et al. *VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video*. in *Asian Conference on Computer Vision*. 2018. Springer.
66. Zhang, Y., Z. Chen, and H.I. Hee, *Noninvasive measurement of heart rate and respiratory rate for perioperative infants.* Journal of Lightwave Technology, 2019. **37**(11): p. 2807-2814.
67. Instrumentation, A.f.t.A.o.M., *Cardiac monitors, heart rate meters, and alarms.* American National Standard (ANSI/AAMI EC13: 2002) Arlington, VA, 2002: p. 1-87.
68. Bobbia, S., et al., *Unsupervised skin tissue segmentation for remote photoplethysmography.* Pattern Recognition Letters, 2019. **124**: p. 82-90.
69. Heusch, G., A. Anjos, and S. Marcel, *A reproducible study on remote heart rate measurement.* arXiv preprint arXiv:1709.00962, 2017.
70. Giavarina, D., *Understanding bland altman analysis.* Biochemia medica: Biochemia medica, 2015. **25**(2): p. 141-151.

71. Artusi, R., P. Verderio, and E. Marubini, *Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval.* The International journal of biological markers, 2002. **17**(2): p. 148-151.
72. Mahmoud, T.M., *A new fast skin color detection technique.* World Academy of Science, Engineering and Technology, 2008. **43**: p. 501-505.
73. De Haan, G. and V. Jeanne, *Robust pulse rate from chrominance-based rPPG.* IEEE Transactions on Biomedical Engineering, 2013. **60**(10): p. 2878-2886.
74. Balakrishnan, G., F. Durand, and J. Guttag. *Detecting pulse from head motions in video*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.
75. Hyvärinen, A. and E. Oja, *Independent component analysis: algorithms and applications.* Neural networks, 2000. **13**(4-5): p. 411-430.
76. McDuff, D. and E. Blackford. *iphys: An open non-contact imaging-based physiological measurement toolbox*. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019. IEEE.

8. Publications

1. Gupta, A., Ravelo-García, A. G., & Morgado-Dias, F. (2021). Motion and Illumination Resistant Facial Video based Heart Rate Estimation Method using Levenberg-Marquardt Algorithm Optimized Undercomplete Independent Component Analysis. (submitted in IEEE Journal of Biomedical and Health Informatics).

   Authors Contributions

   Gupta, A., Ravelo-García, A. G., & Morgado-Dias, F. conceptualized the study.Gupta, A. was responsible for implementing the methodology, result simulation, and preparation of the initial draft. Finally, the initial draft was reviewed for corrections by Ravelo-García, A. G., & Morgado-Dias, F, followed by the draft's rectification by Gupta, A.

2. Gupta, A., Ravelo-García, A. G., & Morgado-Dias, F. (2021). Heart rate and Oxygen saturation estimation using facial videos: A systematic review. (submitted in Computer Methods and Programs in Biomedicine).

   Authors Contributions

   Gupta, A., was responsible for developing the search query for conducting multiple database articles search, which was reviewed by Ravelo-García, A. G., & Morgado-Dias, F. Furthermore, Gupta, A. conducted the title, abstract, and full-text screening for inclusion and exclusion of research articles for the review. Subsequently, Gupta, A. prepared the initial draft. Finally, it was reviewed for corrections by Ravelo-García, A. G., & Morgado-Dias, F, followed by draft rectification by Gupta, A.