# Towards a Contactless Vital Sign System

by

Xiaocong Ma

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements
for the M.A.Sc degree in
Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

# Abstract

Human vital signs are crucial parameters which reflect essential body functions and are often accessed by medical professionals at the first place during clinical diagnostics to provide immediate assistance in health status measurements. However, due to the recent COVID-19 pandemic, measurements made with direct body contact have become increasingly challenging and costly because of the spreading nature of this virus. Therefore, contactless vital sign measurements are highly desirable, and it motivates us to research and develop a new solution which is capable of performing real time heart rate (HR) detection, respiratory (RR) detection, and body temperature (BT) measurement together from a distant human subject under an ambient light environment. The thesis describes a new system framework, which utilizes the power of computer vision to collect remote video image data, processes them using signal processing and machine learning (ML) technologies simultaneously, and produces rapid updates on display. Furthermore, our validation analysis on the system has showed varied results based on different methodologies used, which enables us to apply the most suitable approach on each component for an optimized final integration.

At the time of completing this thesis, we have achieved a complete system integrated with remote HR, RR estimations and BT detection, which are all fully functional in both real-time and offline. To further refine the performance on HR estimation, we selected the extreme gradient boost model through a number of ML models we tested, as it not only gives the lowest root mean square error of 8.2 but also produces stable and robust output.

# Acknowledgments

Foremost, I would like to express my deepest appreciation to my respected academic supervisor, Professor Abdulmotaleb El Saddik, for his patient guidance, brilliant advice, immerse knowledge and tremendous support over the years. It is a great honor and joy to work under his supervision since day one. Aside the academic curriculum, Prof. Saddik is not only a supervisor, but also a trustworthy life mentor who provides enormous care to his students. Without his generous help and encouragement, especially during the period of COVID-19 pandemic hardship, this dissertation would not have been possible. His visionary leadership, extreme enthusiasm and unmatched dedication towards science and knowledge have inspired and invigorated me to move forward to above and beyond.

Secondly, I would like to express my sincere thanks to Mr. Izaldein Al-Zyoud for his massive support and seamless collaboration in this work, particularly in the part of machine learning. It is my great pleasure to have such a talented and hardworking partner to complete work together at full speed. I also would like to give my heartflet gratitude to Dr. Diana P. Tobón for her invaluable advice, assistance, feedback, and care provided throughout my research. It is you who helped me to realize my potential and kept me motivated in and out of the lab.

Next, I would like to thank Mr. Haopeng Wang who has made important effort in contribution of dataset collection, as well as all my friends and MCRlab members who participated and collaborated in this research.

Last, I am profoundly grateful to my parents and my wife for their endless love, understanding and support, both at home and overseas. You all make my life complete.

# Dedication

This thesis is dedicated to my father, the man who taught me to prepare for each moment to be a learning moment, and to never give up pursuing my dream.

# Table of Contents

# List of Figures

# List of Tables



# List of Equations

# Chapter 1. Introduction

## 1.1. Background, Problem Statement and Motivation

Human vital signs are essential information for both physical cardiac and psychological studies. Methods used to obtain vital signs, such as pulse taking and breath counting, have long become important procedures through out the human history of medicine. Ancient races like Arb-Islamic, Chinese, Egyptian, Indian, all have mastered ways of extracting vital signs from a patient through their unique practices of close body medical examination [1]. Traditional Chinese medicine, for example, applies four basic methods of diagnosing illness used: observation, auscultation/olfaction, inquiry, and palpation (See Figure 1). By doing so, doctors can utilize basic human perception and senses to determine the vital signs and thus identify the health status of the patients through accumulated knowledge and experience. Similar firsthand exams are also applied in the modern clinical diagnostics, but the close contact with body can result in high risks of virus transfer and infection in cases which contagious illnesses are presented.

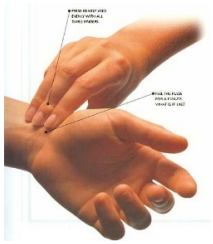


Figure 1 Traditional Chinese palpation [2]

In the modern domain of smart health and well-being, along with medical diagnostics, vital signs have given added values in terms of long-term health monitoring, athletes training, psychological state identifying and many more. By digitizing the analog vital sign information and merging between the physical and virtual world, modern technologies have achieved not only continuous monitoring on human body but also fast analyzing and accurate prediction like never before. One of the iconic works introduced by Dr. El Saddik [3] is the Digital Twin framework which coverages and synchronizes the multimedia digital world with the real-life entities in the physical world. In this new

framework, high speed communication links are created between the physical and digital worlds and necessary data are continuously collected, interchanged, optimized and feedback between the two for improved quality on both smart medicine and smart wellbeing applications.

To realize the Digital Twin architecture, it is essential to have efficient sensing devices for accurate data collection at the first place. Fortunately, under the rapid development of the e-health industry, commercial wearable smart devices, like as Fitbit™ and smart watches have been developed and improved over the recent years. While conventional hands-on devices can measure multiple vital signs with accurate results, their limitations do exist. One of them is that the user must physically wear the device and directly skin contact is needed prior to a measurement can be made. Others like devices are not easily transferable from one user to another, or bothersome, and perhaps irritating and distracting [4]. In addition to these problems, the 2020 coronavirus pandemic (COVID-19) has created a global crisis where social distancing and contactless economics are becoming the "new normal" [5]. Regular skin contact method for infection symptom detections have thus become increasingly difficult and risky due to the spreading nature of the virus, therefore, new remote monitoring system of human vital signs is being considered a critical technology solution in digital health to help tackling this urgent global need.

To respond to the pandemic and provide a more flexible, cost-efficient, and user-friendly solution, we take a different approach by adapting computer vision-based technologies, a.k.a. Remote Photoplethysmography (rPPG) [6]. The rPPG term is originally derived from the term Photoplethysmography, or PPG which was firstly introduced by Hertzman in 1938 to describe a non-invasive optical technique capable of detecting blood volume pulse (BVP) changes in the blood vessels using backscattered optical radiation (Hertzman(1938)) [7]. The strength of optical radiation is partially absorbed by human skin tissue, but due to heart cardiac activity there is a variation in terms of blood flow, and this variation can be modulated and thus detected by video camera as weak light pulsations that does not appear visible to human eye [8]. The technique of PPG requires a source of incoherent light and a photoreceiver for simply and efficient implementation (See Figure 2), which led to a quick development to pulse oximeter, a device that measures the arterial blood oxygen saturation level on the skin tissue. The use

of pulse oximeter is widely prevalent in medical care that it is often regarded as a fifth vital sign [9].

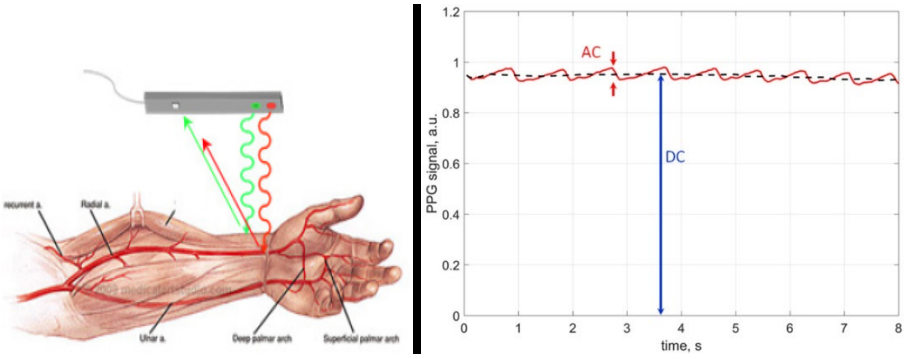



Figure 2 PPG detection of a blood pulse change [10]

rPPG, sometimes also referred as Image Photoplethysmography or iPPG, has remarkably similar implementation compared to the PPG as they use the same fundamental methodology. However, instead of using a close-to-skin pulse oximeter which consists small beams of LED light as its source and a photodetector as its receiver, the rPPG uses ambient light or even natural light as its source and a digital video camera as its receiver (See Figure 3).

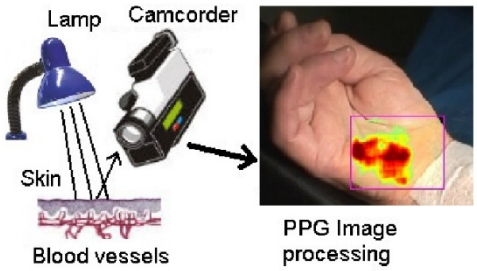



Figure 3 rPPG detection of a blood pulse change [8]

Unlike the PPG signals that can be measured right off the sensor, because the rPPG relies on video images, it requires further image processing to be done before the desired vital sign data can be extracted. As the sensitivity of modern camera sensors develops, it has made possible for the rPPG to obtain reasonably accurate results under ideal conditions compared to PGG's, although its accuracy is considerably inferior to the PPG due to noise, light change and motion artifact. Originally took place as a unique part in the sensor module of the Digital Twin framework, our research is aimed to combat these challenges in the

rPPG data and to provide multimodal vital signal information without a direct contact to the test subject. By examining the intensity change of a reflective light on a person's facial skin and applying sophisticated image and signal based analytic techniques, our integrated system is able to accurately determine the heart rate, respiratory rate, and body temperature of a human subject using optically collected human facial video from a low-cost camera system, within a designed range.

## 1.2. Design Development

Our objective of this research is to create a proof of concept on a remote estimation system that is integrated to sense multimodal vital signals in the real-time. The system's performance can be further refined by adapting ML model designed on time-series data produced by the system at an early stage. We gradually proceed towards this main objective by splitting the tasks in a modulated way:

1. **To obtain image sequences on desired facial region**: To monitor vital signs, the system needs to be capable of identifying a human at first place. The image acquisition module powered by open source computer vision models can provide such a feature and serves as a critical role to extract images that may contain the most useful clinical information to be processed further. Since our research is based on the rPPG principle, it is essential that we are able to obtain a consistent continuous sequence of the facial skin images. The images are the main input source of our system, and because the signals we are seeking are relatively weak outside the visualizable quantization levels, the quality of the images need to be maintained as high as possible.
2. **To calculate desired time-series vital sign data based on the images obtained**: When transforming the large size of 2D-images sequences into 1-D time series data array, despite the reduced size of data, the requirement of computational expense is greatly lowered, while important information on biometric features can still be retained. This is a crucial factor for our design to be running in the real time,

especially with constraint hardware. In addition, the selection of color channels for either heart rate or respiratory rate data, which gives a more robust quality of data production, is also performed in this module.

3. **To perform heart rate estimation based on signal processing techniques**: The initial heart rate estimation module applies signal conditioning, filtering and space transform techniques to determine the highest spectrum band which possibly represents the heart rate inside a selected range of frequencies.
4. **To perform a respiratory rate estimation module based on signal processing techniques**: This module is almost identical to the development of the heart estimation module. The differences we introduced here are different color channel / space and frequency range selections.
5. **To improve heart rate and/or respiration rate estimation modules based on multiple machine learning algorithms**: To further improve the performance and accuracy of both heart rate and respiratory rate modules, our research attempted to utilize the power of supervised learning to produce estimated end-to-end results directly. Further experiments on constructing additional ML models are made not only to replace numbers of signal processing stages, but also to create flexibility which enables simple system prototyping and deployment for real-time experimentation.
6. **To compare and selected the HR/RR model with the best performance:** To demonstrate the performance of each algorithm applied in the previous stage, the results from different models as well as from the signal processing are then validated and benchmarked against each other for a best model selection.
7. **To integrate a body temperature sensor that can fetch body temperature from facial skin**: The temperature measurement module serves as an add-on feature of the entire system which can be put on or removed depending on the need. A Serial Bus interface is implemented for the module to run in the real time.
8. **To finish a multimodal vital signs estimation system as a whole**: This is the final objective reached in this research. All modules work simultaneously to provide rapid updates on estimated results displayed beside the face of the test subject on screen.

## 1.3. Thesis Contribution

The major contributions of this research are listed as the following:

- Design and development of an initial system featuring real-time heart rate estimation using signal processing techniques
- Design and development a complete integration that consists multiple algorithms including facial feature detection, data extraction, signal processing, ML model design and Serial data communication all together as a whole
- Design and development a high accuracy light weight ML model to replace the most signal processing steps by providing an end-to-end HR estimation solution for easy deployment

## 1.4. Practical Application

The output of this research can be further developed and refined for real world applications under various scenarios. Ideally, such a system can be applied at hospital emergency room for pre-screening of the COVID-19 virus. Other applications such as fatigue detection in vehicles, lie detection at border security, or newborn monitoring in the nursing room. Last but not the least, multiple vital signs can serve as input for more complex analysis and study, such as emotion detection, psychological care, and paroxysmal prediction on chronic diseases.

## 1.5. Thesis Organization

The content of this thesis is organized as the follows:

- In Chapter 1, we provide an introduction that explains the background, motivation, objective, and contribution of this research.

- In Chapter 2, we conduct a literature review on the previous works that are related within the scope of our research.

- In Chapter 3, we present a short overview on the overall system architecture and design configuration. We also introduce the methodologies used for system pre-setup and image preprocessing.

- Chapter 4 presents the first system of our research, which is a heart rate estimation system based on signal processing techniques.

- Chapter 5 focuses on the development of machine learning models which bring an upgrade to the system of our research. We also integrate the respiratory rate estimation feature and provide offline testing results.

- Chapter 6 finalizes the system integration of our research. Here we include system optimization, the integration of body temperature detection feature, real time tests, and our findings.

- A conclusion and a final discussion on the potential of our work are included on Chapter 7.

## 1.6. Scholastic Output

1. Already published: Remote Photoplethysmography (RPPG) for Contactless Heart Rate Monitoring Using a Single Monochrome and Color Camera, *Lecture Notes in Computer Science Smart Multimedia*, 2020, pp. 248–262., doi:10.1007/978-3-030-54407-2_21.

# Chapter 2. Related Work

It is not a new concept for computer vision-based and rPPG methodologies to be adapted for contactless vital sign measurements in recent academic researches. By examining the previous works, we have set a solid foundation for our research based on their findings. Since our research consists multimodal vital signs, the related works are categorized according to the estimation on each vital sign as the follows.

## 2.1. Work on Heart Rate Estimation

Before we start the evaluation on any heart rate estimation methodology using rPPG images, it is noteworthy to mention some similar approaches and color channel selections.

From one of the recent surveys, C.Wang in his paper [11] summarized HR detection using facial videos into two different main categories: color intensity based methods which use rPPG signal that related to BVP changes [12], and motion-based methods which use subtle upright head motions in the vertical direction caused by pulse activities [13]. Because our work is mainly based on the rPPG signal, we do not plan to include any motion-based methods in the related works introduced in this research. Additionally, Wang has concluded the rPPG intensity-based methods are generally more effective since a stable reference object is usually needed in the background for correct movement detection.

Second, in paper [13] [14], their authors have compared green channel's signal strength in the rPPG signal against the strength other channels in RGB and other color spaces and summarized that the green channel gives the best Signal to Noise Ratio (SNR) according to their results. This agrees with findings from other serval research groups [15]. These researches also show a largest AC component of PPG waveform at frequency range between 520 to 580 nm, which is a clear indicator showing that spectrum of green light region has a strong absorption by erythrocytes [7], and thus can be the most robust and reliable part of the video images contributing to PPG and heart rate detection.

### 2.1.1. HR estimation using signal processing

Many previous works are focused on tailoring and applying the signal processing technique at any chosen stages of the entire process, including pre-processing, signal extraction, data conditioning and post-processing [11]. The pre-processing stage, the location of ROI selection on the facial images for example, can be somehow varied due to different situations. While the whole face ROI selection can contain undesired facial objects such as hair, glasses, and hat that do not affected by BVP change and lead to lowered accuracy, it is more resistant to nonrigid motion since the calculation on the entire face cancels the facial movement artifact to some degree [16]. In data extraction stage, both Rahman [17] [18] and Demirezen [19] tried to reduce the motion artifacts by using Independent Component Analysis (ICA) and nonlinear mode decomposition methods respectively, and Demirezen's has done better results according to his report. In the post-processing stage, we found time domain filtering and frequency domain peak-detection methods are the most applied methods in signal processing category.

### 2.1.2. HR estimation using machine / deep learning

We conclude that most of the research done using advanced machine/deep learning can be fit into two categories: either non-Convolution Neural Network (non-CNN) or CNN modeling.

In the non-CNN category, 2D image sequence is usually transferred into 1D time series data at first before any additional machine learning model is applied, therefore, models are usually adapted for creating dedicated functionalities within the process. In the data extraction stage, A. Osman et al. [20] conditioned cleaner heart rate data signals by modeled a Support Vector Machine which periodically shapes the time series waveform extracted from an offline video, while M. Bian [21] did the similar filtering by deploying a two-layer LSTM model instead. In [22] H. Monkaresi used multiple KNN-based ML model specially designed for each specific test subject. These models replaced the function of the common power spectrum component selection in the data extraction stage of an

offline detection process and gave a sound reduction to the final error percentage on the output. Based on this framework, Wang in [11] replaced the KNN model with a random forest K-star model that is capable to adaptively obtain a best channel out of the selections and produced more accurate results. Using similar strategy, Ghandian in [23] achieved accurate results above 85% by creating an adaptively selection model which is capable to select best channel in RGB color space from an offline video. In the final processing stage, Y.C. Hsu et al. [24] prototyped a regression learning model which replaced the function of peak spectrum component selection and claimed a more stable output compared to conventional signal processing.

The CNN modeling is in its own unique category because of the character of CNN – which is a deep learning class that is most commonly applied to visual imagery to extract key features. However, despite of the effectiveness of this specialized deep learning, the heart rate vital sign still appears to be a weak feature contained within the images that is hardly visualizable, and thus it is quite common that more processing stages, such as Eulerian Video Magnification (EVM) [25] are added to amplify the signal prior to the stage where CNN model can be utilized. One of the research achievements designed by Y.Qiu [16] from MCRLab utilized the power of ImageNet architecture and built a modification CNN model based on the original framework. With spatial decomposition and temporal filters added on 3D image data stack which obtained sequential images extracted from selected ROIs regions of a face video, her model is able to generate optimal prediction of heart rate by an accuracy of 74% within the design specification in the real time. Comparably, Y.Y.Tsou et al. proposed a Siamese-rPPG 3D CNN network [26] which models the spatial and temporal characteristics from two selected facial ROIs and claimed a top-notch performance in their researched results. Apart from using 3D image stack as input, E.Brophy in his work [27] took an unique approach and trained a CNN model which makes use of 2D time-series plots as input features. While this gives him highly competitive results on estimating heart rate from offline videos, X.Niu and his team have already taken a step further [28] by adding a combined spatial-temporal attention map to model training and improved the robustness of the network when dealing with various environment changes.

## 2.2. Work on Respiratory Estimation

Using video image data to extract and determine respiratory estimation can be quite similar to the heart rate estimation process. Respiration movement causes pressure variations in the blood vessels which can also lead to BVP changes not only in the chest but also on the facial area. Additionally, respiratory sinus arrhythmia [29], which is a term used to describe the effect of increasing and decreasing of HR at the time of breathing in and breathing out, can be another lead to indirectly calculate the RR given heart rate variation (HRV) is already known. Therefore, even though the rPPG signal is heart rate related, it is still possible to determine the respiration rate based on carefully selected ROI regions on the face.

It is again noteworthy that the majority of works have been done to determine RR rate can be based on either monitoring the motion of human body related to the person's breathing activity [30], or using rPPG methods and extracting RR data indirectly from other signals such as blood oxygen saturation levels (SpO2) and heart rate variation (HRV).

In [31], Braun et al. constructed a simple motion-based algorithm which splits input images into blocks and adaptively selects the block with highest likelihood of true RR activity as the ROI. In [32] Lin et al. used upper body as a reference object and detected harmonic movements of RR by modeling vertical movement of the face. Similarly, in [33] Tran et al. modeled their RR estimation algorithm by referencing the chest region movement to the upper body, and in [34] Shao et al. proposed a differential detection and tracking algorithm for RR estimation using shoulder edge movement. Furthermore, in [35] Wu et al. amplified the breath related body movements using EVM, and in [36] Yang et al. utilized vertical and horizontal nostril movement for their offline RR calculation.

To our best knowledge, many motion-based methods suffers when:

- The test subject is relatively too far from the camera
- The shutter speed is slow to cause motion blur under dark light condition
- A reference location needs to be carefully selected and tracked to detect the movement
- Body or facial tracking algorithms can cancel or cause wrong motion detection
- Unrelated movements can be hard to differentiate from RR-induced movement [37]

On the other hand, using the intensity-based methods can achieve similar accuracy compared to motion-based methods given that an effective skin region is selected [38], while not to worry about above challenges. In [39] Prathosh and his team modeled their algorithm to detect RR by investigating the intensity change of body reflective light caused by respiratory movement. In [40] Mirmohamadsadeghi and her team conducted a real-time RR detection experiment by detecting RR rate using inter-beat variations HRV data derived from detection of the HR values. Guazzi et al. in [41] used a commercial grade RGB camera and successfully constructed a novel algorithm to determine oxygen saturation change in visible light, since the change of $SPO_2$ can lead to a direct determination of RR rate. Likewise, Rosa et al. in [42] achieved the same goal by adapting the EVM technique to amplify the color change caused by the changes of $SPO_2$. Interestingly, taking another approach by analyzing the effect of various color spaces on determining RR rate from rPPG image data, Shourjya and his team in [43] reported that the Hue channel in HSV color space can give a significant advantage in terms of SNR over the regular RGB color space. Additional ML modelings are also added to the solution of intensity-based RR detection. In [44], Palaniappan and his colleagues compared diagnosis results of respiratory pathologies using either Support Vector Machine (SVM) or K-NN machine learning algorithms and concluded that the K-NN classifier can be better than the SVM for better pulmonary acoustic signals identification. In [37] Ghodratigohar introduced an ML model to adaptively select the best signal decomposition results for RR data representation and achieved an offline detection accuracy over 82%.

## 2.3. Work on Body Temperature Estimation

The body temperature detection is added to our research scope as this is one of the crucial parameters used clinically for symptoms detection during the COVID-19 pandemic. Our initial research has showed approaches can be made for contactless target temperature monitoring in two possible ways: Adding a low-cost thermography image-based camera, or an infrared thermometry sensor. While thermography infrared camera can produce images with multiple pixels, the thermometry sensor usually only produces one value based on the average temperature of the area it covers [45]. However, results of most studies

made on temperature detection are not quantizable into accurate measurable range. Using a thermography camera at a close distance, Both Formenti et al. in [46] and Tanda in [47] have found clear evidences of increasing and decreasing skin temperature in response to localized exercise in their test subjects. Soerensen in [45] reviewed the thermometry sensor and summarized it is a promising technology can be used towards unhealthy pig detection. In [48] Gade did a comprehensive survey and concluded that although thermography cameras can provide additional information on the temperature of the subject, its high cost and relatively low accuracy on pixelized temperature measurement still prevent it to become a popular choice when precise quantization on the measurement is needed.

## 2.4. Work on Multimodal Vital Sign Detections

Because our research is an integration of multimodal vital sign detection system, it is fair to review works made on similar integrations on the entire system level. In [49] Y.Sun demonstrated his study in extracting both HR and RR vital signs using rPPG signals extracted from offline videos under various exercising conditions. In [50] G.Sun et al. successfully extracted multiple vital signs on both thermal and regular images using an advanced infrared thermography system and made promising results on infectious diseases detection. Wei also worked together with his team in [51] to propose a blind source separation-based method that can synchronously measure both RR and HR in a dynamic way with sliding window data truncation. Impressively, Chen from MIT media lab teamed up with Mcduff from Microsoft in [52], together they developed a CNN based end-to-end deep learning network model called DeepPhys, which is capable to provide spatial-temporal visualization of physiological distributions and to extracting both HR and RR signals simultaneously.

## 2.5. Additional Thoughts

Through all the works we have examined in this research, we found the majority parts are focused on offline/non-real time vital sign estimations as they can achieve higher accuracy due to the fact that the performance and computational expense are not priorities in the design. In addition, many researches adapted ML models mainly to apply as a functionality replacement of a single stage within the whole processing, which can have limited reusability and portability on the development of a real-world application. Since our research aims to create a system that needs to give accurate multiple vital signs detection in the real time, it is very important for us to try out different methodologies and combinations, while keeping a reasonable balance between the system's overall accuracy and performance.

# Chapter 3. System Overview and Image Pre-processing

As we gradually proceed the system integration and testing simultaneously, it is important for us to split the tasks into individual modules so that to our progress and the complicity of the system are always in check.

## 3.1. System Architecture Overview

Our system design is completed based on the integration of two main methodologies: The HR and RR estimation is based on intensity-based rPPG process, whereas the body temperature estimation is based on thermometry sensor detection process. The overall system architecture consists 5 main parts that integrated and interconnected as shown below (See Figure 4).

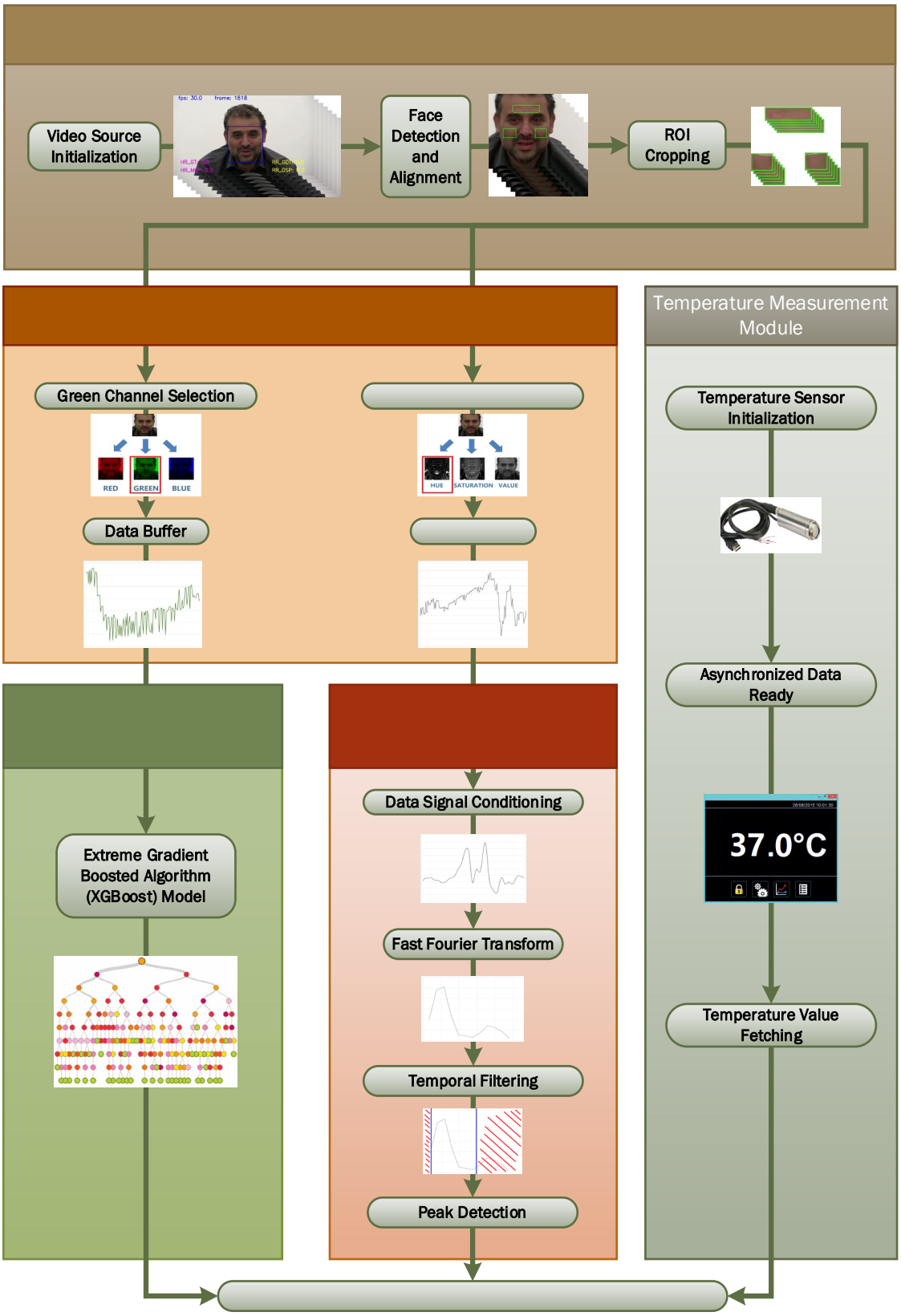



Figure 4 General System Architecture Flow Chart

The functionalities of each module are described as the following:

1. Video Frame and ROI Acquisition Module: This is the first module in rPPG vital sign processes for both HR and RR estimations. It takes a video sequence input and extract a group of ROI images on facial area if a face presented. Once a video source is initialized, a facial detection engine is applied to check if a face is presented on the received image. If a face is detected, a face image is then cropped, aligned, and resized into designed size, and 3 ROI images are extracted from the face image.

2. Vital Sign Data Extraction Module: This module performs color space transform and data averaging process for both HR and RR estimations. Once the cropped ROI images are available, the green channel image in RGB color space is selected in HR estimation, and the Hue channel image in HSV color space is selected in RR estimation. Both green and Hue images are then transformed into 1D time-series data and shaped with a format of 150 data samples per data buffer for further processing.

3. Heart Rate Estimation Module: Once the shaped time-series HR data is ready, it is fed into a pre-trained extreme gradient boosted model (XGBoost) to obtain estimate HR results. The XGBoost model is trained by using data generated and engineered from several datasets obtained from different research institutes, and it is deployed as an end-to-end solution for providing direct and accurate result.

4. Respiratory Rate Estimation Module: Multiple signal processing stages are utilized to obtain RR estimation results in this module. Once the shaped time-series RR data is ready, a data signal conditioning stage is followed to amplify the signal to noise ratio. Next, Fast Fourier Transform is applied to change data into frequency domain where unwanted frequency bands are filtered out. Finally,

peak detection technique is used to select the possible value of RR from the remaining of the frequency band.

5. Temperature Measurement Module: The thermometry process works as a stand-alone module for body temperature detection, where direct temperature measurement is taken place. Fast data readings obtained from the sensor is sent through a Serial communication buffer where slower asynchronized data fetching is taken place. The output from this module is then updated with HR and RR output at the same speed on screen.

## 3.2. System Integration Steps

The first step of our implementation consists only the heart rate estimation function using signal processing techniques which has been documented in the published work [53]. This is because we want to build a proof-of-concept project that can be simple enough for demonstrating early results prior to proceed further in this research.

In the second step, we try to build multiple machine learning models for both HR and RR detection, and we also compare the performance of each model.

In the last step we include the body temperature part and complete the entire system integration by selecting the best approach we found for each detection part in the second step. The real time integrated performance is also verified in this step.

## 3.3. Video Frame and ROI Acquisition Module

This module serves as a common initial step to obtain rPPG related image data, and all further modules in the HR and RR detection processes rely on it. The main function for this module is to extract raw region of interest (ROI) images from the input video source.

### 3.3.1. Face detection

Once an input video frame is received from either the webcam or video file, first it is passed into a face detection stage. The OpenCV Haar Feature-based Cascade Classifiers library is utilized for fast human face detection. If a face is identified, a blue rectangular box is then placed on the detected face and a face frame is cropped and resized to a size of 256 by 256 pixels to equalize the size of input data. The face is also tracked on every successive frame until the face is no longer detected or the video source input ends.

#### 3.3.1.1. The OpenCV cascade classifier

The Haar cascade classifier is well known for its effectiveness in object detection in computer vision. This machine learning method is originally proposed by Paul Viola and Michael Jones in [54] back in 2001 and has been developed by countless open source community since then. In theory, the classifier needs to be trained based on large number of images that has already been tagged for positive and negative presence of the target objects. Focusing on real-time applications, the OpenCV face classifier [55] which is developed on Haar cascade ML modeling, is a highly optimized open source deep learning face detector that can be obtained as a Python library.

### 3.3.2. Facial landmark prediction, face alignment, ROIs cropping

Once a cropped facial image is obtained, it is put through a pre-trained facial landmark predictor to estimate the location of 68 (x,y)-coordinates that map to facial structures on the face. Particular locations of some coordinates are then utilized in performing facial alignment and ROI locations identification (See Figure 5). Raw images on the identified three ROIs (two on cheeks and one on forehead, marked by green rectangles) are then cropped out again for further processing.


Figure 5 Real-time face detection, facial frame cropping, face alignment, and ROI identification

#### 3.3.2.1. The Dlib facial landmark predictor

Dlib 68-point facial landmark predictor is originally proposed by Kazemi and Sullivan back in 2014 [56]. This predictor model is trained on manually labeled data with specific (x,y)-coordinates of regions surrounding each facial structure [57]. It uses regression trees technique to estimate the facial landmarks based on pixel intensities instead of facial features which resulted in high quality predictions in the real-time. The specific library version we adapted in our research is the 68-points predictor, where the coordinates and selected ROIs are presented as below (See Figure 6).

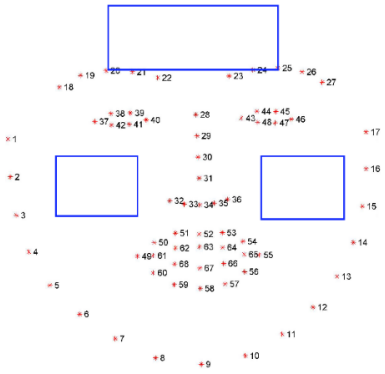


Figure 6 ROIs selection using Dlib C++ 68-points facial landmarks detection model

### 3.3.2.2. Facial alignment

In many cases, the detected face is not properly positioned in the face frame. To combat face rotation and to crop the correct regions of the ROIs, a facial alignment using Python Imutils library is applied to straight up the face pose (See Figure 7). Once landmark point #40 and #43 are located, their vertical y coordinate values are compared. If they are not the same, a rotate angle is calculated using a referenced horizontal line and the entire face frame is then rotated accordingly.

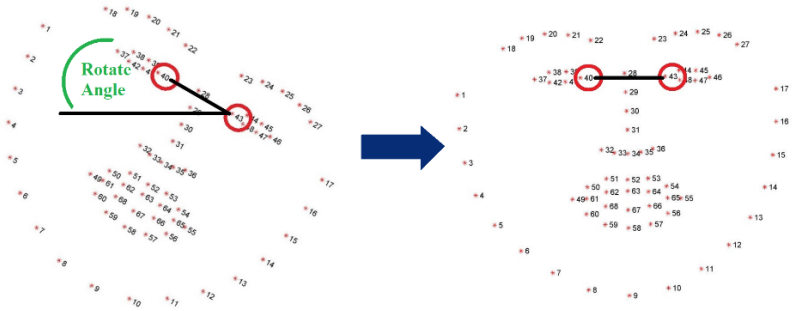



Figure 7 Facial alignment

### 3.3.2.3. ROI selections

While it is possible to use the entire face frame region to recover rPPG signals, the process may be strongly affected by non-rPPG related objects added to the image. Facial movements such as eye blinking, facial objects such as glasses and hat, and facial hair such as bangs and mustache, all would have negative contribution on the facial image as their color intensity are not affected by BVP change. To eliminate these obsoletes from the facial frame, partial regions, it is essential for us to select and track the most effective ROIs on the face where the presence of skin and blood vessels are at maximum. In [58], the author tested the rPPG signal strength on the face by splitting a face frame into 260 small blocks and evaluating the BVP strength on each block. He concluded that the forehead and cheeks are the most effective regions for HR estimation. In addition, some of the works made in the previous research have created complex algorithms to adaptively detect regions that have the best signal strength in BVP as ROIs [59], however, the added computational expense may slow the overall performance of our integrated system in the real-time. Thus, we choose to use simple rectangular box made on the Dlib 68-points mapping coordinates, where three individual ROIs are selected in this research: one on the forehead and two cheeks (See Figure 6). Specifically, the following coordinates are used on the ROI selections:

- Left cheek rectangular ROI: (X of #55, Y of #30) to (X of #13, Y of #34)
- Right cheek rectangular ROI: (X of #4, Y of #30) to (X of #49, Y of #34)
- Forehead rectangular ROI: (X of #20, (Y of #20)-20) to (X of #25, Y of #20)

The size of the ROI is varied from person to person depending on his/her actual facial landmark composition, but below are the approximate pixel sizes of cropped ROI images based on an average adult.

- Left cheek: 40x30
- Right cheek: 40x30
- Forehead: 80x20

## 3.4. Vital Signs Data Pre-processing and Extraction

The data extraction module serves as an intermediate data pre-processing stage in the rPPG (HR and RR) estimation process. Inside this module, ROI images produced by the previous module are split into specific color channels and converted into 1-D time series data values and filled into buffers.

### 3.4.1. Color space and color channel selection

When each raw ROI image is produced, it is split into three channels: Red, Green, and Blue based on the RGB color space. Most of our previous researches stated in Chapter 2.1 have shown that the green channel usually contains the highest signal to noise ratio (SNR) data in terms of BVP changes, thus it becomes our picked choice for HR estimation. Data on Red and Blue channels are then discarded and total amount of data in the process is reduced by 2/3.

### 3.4.2. Data averaging and slicing buffer length

Next, three 1-D data buffer arrays are created, and color intensity values from all ROI images in green channels is averaged and stored temporarily in their corresponding data buffers. Each buffer is shaped to accept a maximum of 150 data points, which can only be fully filled when 150 continuous ROI image frames are successfully received and processed. For a camera or video source which is recorded under a frame rate of 30 frames per second (FPS), the data buffers usually take up to 5 seconds to be filled once they are initialized or reset. The reason to set the slicing buffer length to 150-point long is that, by doing so, the amount of data it contains are large enough to retain high accuracy, while still fast enough to be filled and processed in nearly real-time.

## 3.5. Conclusion

In this chapter, we have presented a general overview of the proposed system and described how we planned to achieve the solution as our research develops. In addition, we have introduced the related methodologies used in system pre-setup, facial image detection, ROI selection, as well as data pre-processing. The engineered time-series data which implies the essence on multi-vital signs, has an essential role to play when it comes to both signal processing and machine learning processes need to be adapted in the following stages of our research.

# Chapter 4. Heart Rate Estimation Using Signal Processing

In this chapter, we focus on creating a proof-of-concept system for initial feasible assessment. These results obtained in this chapter are also used as references for performance validation purpose as our experiments proceed further.

## 4.1. General Signal Processing Procedures

Several key signal processing techniques are applied on the Green channel time-series data buffers in the heart rate estimation module, one step at a time (See Figure 8).

Firstly, data are spatially conditioned to filter out undesired noise and to enlarge the AC component's SNR. The spatial data conditioning stage includes the following processes:

- Data outlier filter: data with a threshold no larger than 1.5 standard deviation are retained
- Data Detrending: to eliminate DC offset changes during the process
- Data Interpolation: double the sampled data length by factor of 2; needed for retaining frequency domain accuracy and resolution after Fast Fourier Transform (FFT)
- Hamming filter: smooth the signal to be more periodic
- Normalization: remove DC offset
- Signal amplification: Enlarge AC component by a factor of 10, which is a pre-designed gain from EVM [35]
- Gaussian filter: Sigma = 1, to further smooth the data prior to FFT

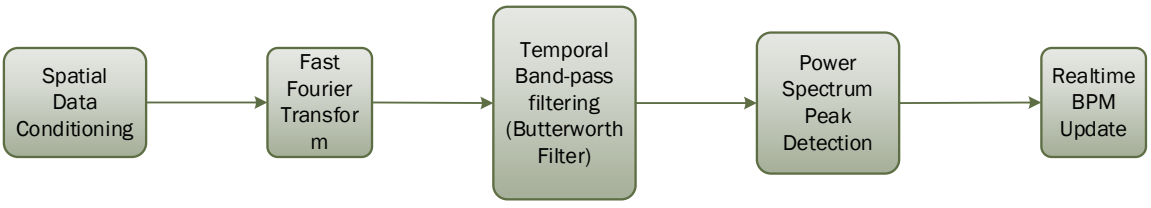



Figure 8 Signal processing flow chart in Heart Rate Estimation Module

Next, the FFT is applied to transform the conditioned time-series data into frequency domain. According to the Mayo Clinic [60], the normal human heart rate falls into the range from 60 to 100 beats per minute (BPM). Thus, we apply a $6^{th}$ order temporal Butterworth filter here to filter out any unwanted frequencies outside possible heart rate range (50-180 rate per minute or 0.833 to 3Hz respectfully). Last, peak detection block is used to extract the strongest signal band inside the selected range to determine the end heart rate results.

Since each ROI has its individual heart rate process, the final heart rate obtained from each process can be slightly varied due to the light, facial object, and other factors. The final BPM results is an averaged value out of all three, which is frequently updated using a 50-points sliding window at 1/3 size of the buffer, or every 2-3 seconds depending on the actual frame rate produced by the entire system.

### 4.1.1. Nyquist limit

To use signal processing technique and produce accurate heart rate results, the data sampling rate in the real-time system needs to be fast enough to be able to detect the maximum heart rate that possibly occurs. In this research we have set this value to 180 BPM or 3Hz. The Nyquist limit theory states that the highest frequency component that can be accurately represented should be equal or less than half of the sampling rate [61], that is

$$f_{max} \leq \frac{f_{sampling}}{2} \qquad \text{Equation 1}$$

Our sampling rate in this research is determined by the video image framerate as well as the performance of the all modules in the entire process. Thus, as long as a received frame rate is no less than 6FPS, we would not encounter a problem in dealing with the Nyquist limit.

## 4.2. Results and Evaluation

We exam the performance of our signal processing HR estimation in two steps:

1. Real time intermediate outputs validation
2. Volunteer group results evaluation

### 4.2.1. Intermediate outputs

Four stages of intermediate outputs in a 30-Seconds window are recorded to CSV sheets during a real time test.

#### 4.2.1.1. Raw input data

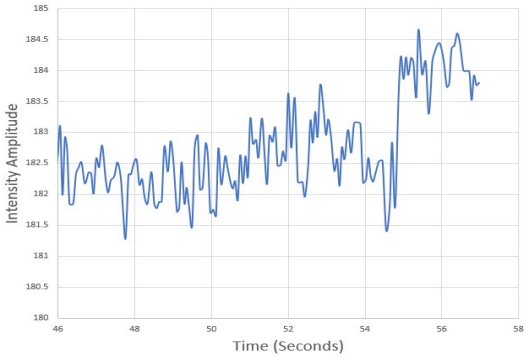



Figure 9 One buffer slice of raw data extracted from Vital Sign Data Extraction Module

#### 4.2.1.2. Conditioned data

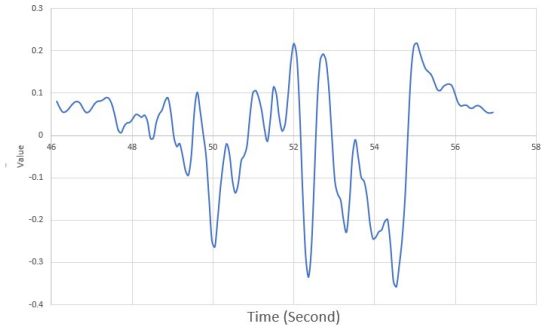



Figure 10 One buffer slice of conditioned data

#### 4.2.1.3. HR frequency power density spectrum

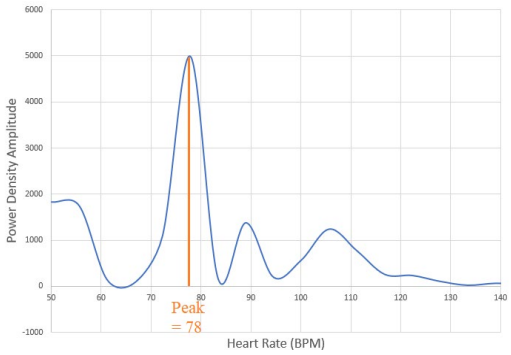



Figure 11 Power density spectrum of one buffer slice

#### 4.2.1.4. A sample of 30-seconds HR output


Figure 12 Output HR results over a 30-seconds window

In Figure 9, one buffer slice with 150 filled data points is plotted. In Figure 10, the same buffer has been conditioned and a clear ECG-like signal is presented. In Figure 11, same conditioned data is transformed into frequency domain, and an obvious peak of the spectrum happens around 78 BMP, which is then picked as the determined HR based on that buffer slice. In Figure 12, a 30-seconds window of final results is presented, where the determined average HR is about 78.

We compared the final results with readings manually retrieved from a BTChoice™ smart band (See Figure 13) which we used as our

ground truth reference in this chapter. The smart band is a commercial grade vital sign detector that provides more accurate HR results based on PPG data obtained from contacted wrist skin. Since the reading in the initial test is 82, we conclude our rPPG results are close to the ground truth.

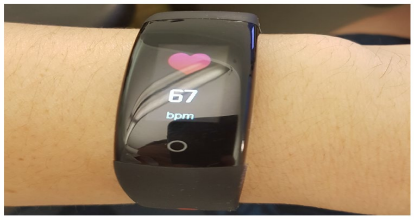



Figure 13 Ground truth reference: BTChoice$^{TM}$ smart wrist band [59]

### 4.2.2. Volunteer group results evaluation

For obtaining a general idea about how our intensity-based signal processing method performs, a group of volunteers with mixed gender and races aged from 18-50 are tested under the real-time mode. Approximate HR readings are taken when the output results are seemed to be stable, while the smart wrist band is also worn at the same time for taking ground truth measurements.

Two different tasks are performed for all test subjects:

- Sitting still pose, where the person is relaxed and breathing normally.

- Light exercise, where the person performs preferred light exercise to give a slightly pump up on HR. The readings are taken immediately after the exercise finishes.

Table 1 HR results from rPPG and PPG measurements [59]

| Test Subjects | Webcam Sit Still | Webcam Exercise | Smart Band Sit Still | Smart Band Exercise |
|---|---|---|---|---|
| Subject #1 | 62 | 88 | 64 | 91 |
| Subject #2 | 87(unstable) | 105 | 90 | 74 |
| Subject #3 | 70 | 92 | 74 | 96 |
| Subject #4 | 70 | 93 | 79 | no reading |
| Subject #5 | 67(unstable) | 55(unstable) | 68 | 86 |
| Subject #6 | 71 | 90 | 71 | 88 |
| Subject #7 | 72 | 96 | 71 | 95 |
| Subject #8 | 75 | 102 | 74 | 96 |
| Subject #9 | 70 | 80 | 71 | 90 |
| Subject #10 | 62 | 76 | 62 | 110 |
| Subject #11 | 72 | 85 | 80 | 87 |
| Subject #12 | 68 | 65(unstable) | 71 | 90 |
| Subject #13 | 87 | 99 | 84 | 95 |
| Subject #14 | 96 | 85(unstable) | 93 | 100 |
| Subject #15 | 74 | 102 | 75 | 95 |
| Subject #16 | 76 | 85 | 75 | 95 |
| Subject #17 | 68 | 101 | 64 | 95 |
| Subject #18 | 65 | 82 | 70 | 83 |
| Subject #19 | 80 (unstable) | 100(unstable) | 74 | 82 |
| Subject #20 | 80 | 88 | 73 | 86 |

#### 4.2.2.1. Measurement errors

We also notice some subjects experience difficulties when taking HR measurement with the webcam, and their final readings are highly unstable (See Table 1). This is probably due to heavy facial artifacts such as makeups, sunscreen and long hair that blocks the effect of BVP change on ROI skin, and thus, we treat these readings as data outliers.

#### 4.2.2.2. Percentage error and correlation values

The calculated percentage error values with respect to our ground-truth data with rPPG intensity-based method is given by [59]:

$$Percentage\ Error = \left|\frac{Camera\ Reading - Ground\ Truth\ Reading}{Ground\ Truth\ Reading}\right| \times 100\% \qquad \text{Equation 2}$$

Based on the results from 20 test subjects sampling pool we have calculated the overall averaged percentage error of our webcam measurement: 4.2% in sitting still, and 12.4% after exercise.

The Pearson correlation coefficients, which is a common statistic values used to measure the relation strength between two variables, in our case, the webcam measurement reading and the ground truth [62]. The coefficient lies within the range of +1 to -1 with a 0 indicate no association between two variables. Our calculations show a value of 0.59 using all participant's data, and a 0.81 with unstable participant's data excluded (See Figure 14).

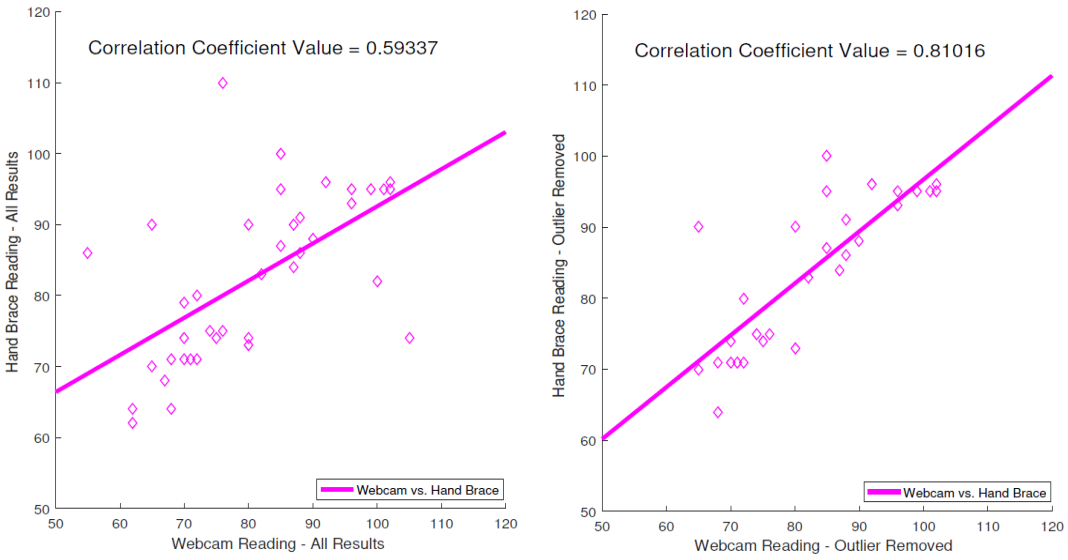



Figure 14 Correlation plots (Left – all data, Right – unstable data excluded) [59]

## 4.3. Conclusion

In chapter of our research, we have successfully implemented a real-time rPPG HR estimation system using the signal process techniques. We have obtained sufficient evidences that the system takes relatively accurate measurements to PPG measurements which set as our ground truth. However, since our results are taken by manual visual inspection, we also notice the various factors such as facial objects, light conditions and motion artifacts that may causing unstable results, and thus further refinements on the performance of the system are needed. Explained in the following chapter, our next goal is to engineer a ML model that can replace multiple signal processing steps and give an end-to-end solution by estimating vital sign directly from extracted raw time-series data.

# Chapter 5. Machine Learning and Respiratory Rate Estimation

In this chapter, we obtain a number of biosignal datasets, adapt various ML algorithms to create different models, validate the training process, and conduct offline model testing to evaluate their performance. Particularly, the evaluation we conduct in this chapter has enabled us to choose the most desirable algorithms later for a final integration.

## 5.1. Difference from signal processing

The integration of ML models in this research is partially built on top of the HR signal processing architecture, and the Video Frame and ROI Acquisition Module are kept the same. In the Vital Sign Data Extraction Module, we keep the same HR route from the previous architecture but create another dedicated route for the RR route. Furthermore, in both HR and RR estimation modules, we put the designed ML models side-by-side against the signal processing methods for a direct performance comparison. Also, the color channel and color space have been switched from Green and RGB to HUE and HSV for better SNR strength during RR signal data extraction.

## 5.2. ML Methodologies and Experimentation

Our research to estimate the real value of HR and RR is a regression predictive task in terms of ML modeling. To produce estimated HR and RR results directly from raw data, all ML models in our design utilize the power of supervised learning based on ground truth physical data provided from multiple datasets. In this paper, we train and validate every model using offline videos and physical ground truth data collected from three different datasets including one which we collected on our own.

### 5.2.1. Dataset Collections

When training ML models by supervised learning methods, it is crucial to have a set of accurate biosignal data we can set as the reference in the first place. To generate the ground truth data for this research, a total of four biometric different datasets come with synchronized offline videos and physical vital sign data have been obtained and examined before feature engineering takes place.

#### 5.2.1.1. The COHFACE[1] dataset[1]

Created by Idiap Research Institute in Switzerland, the COHFACE dataset is a relatively small dataset which has the following specifications:

- Videos are recorded by Logitech HD C525 webcam
- 640x480 pixels resolution with frame rate of 20FPS
- 40 adult participants including 12 females and 28 males, with mixed races from Caucasian, Black, and West Asian. Averaged age is at 35.6 years
- Total of 160 one-minute long RGB videos, 4 videos per participant
- Two light conditions: 400W halogen spotlight and natural light
- All participants are performing sit still poses at the time of recording
- Available biosignal data: time, raw ECG-like BVP recordings, raw ECG-like respiration recordings. Data are collected using chest BVP and respiration sensor belt

Issues found:

1. Our initial inspection on the dataset shows that instead of providing physical HR and RR data, this dataset only comes with raw BVP and respiration

[1] "Future research in this project might use the VIPL-HR database collected by Institute of Computing Technology Chinese Academy of Sciences, and the COHFACE Dataset made available by the Idiap Research Institute, Martigny, Switzerland

recordings. Thus, we convert all raw BVP and respiration data into time-series HR and RR data using Python BioSppy biosignal processing library [63]. However, this introduced a slight delay in video physical data synchronization as every processed HR/RR value requires a set of raw inputs for making calculations.

2. Half of the videos recorded under the natural light condition have strong black/white contrast on the participant's face (See Figure 15), which create failures on the face detection/facial landmark prediction we adapted early. This problem has reduced the data we are able to use from this dataset by almost 50%, as we often could not extract ROI images properly from the videos provided.

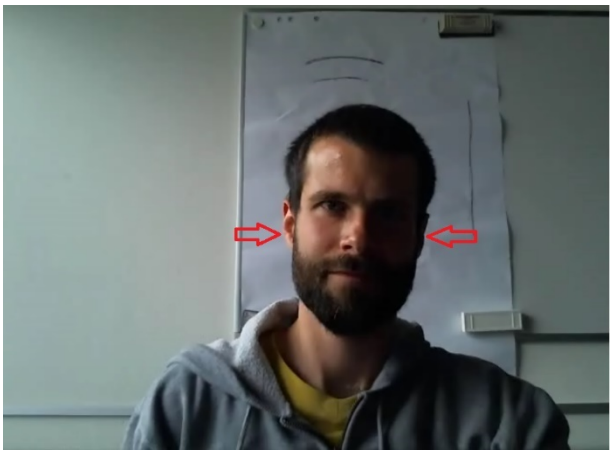


Figure 15 Strong brightness contrast on facial images from COHFACE dataset

3. Because this is a relatively old dataset, all the video recordings are highly compressed and noisy. The effectiveness on extracted rPPG data from these videos may be lowered, since the weak vital sign we are seeking from the recorded images may already be lost during the video compression process at the time the dataset was produced.

#### 5.2.1.2. The VIPL dataset [28]

Created by Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences in China, the VIPL-HR datasets is a large-scale multi-modal HR database with the following specifications:

- Videos are recorded by 4 different devices with various image qualities and speeds: Logitech C310 webcam at 25FPS, HUAWEI P9 phone front camera at 30FPS, RealSense F200 integrated color camera at 30FPS, RealSense F200 NIR infrared camera at 30FPS

- 2,378 visible light videos and 752 near-infrared videos (Only the visible light videos and their corresponding physical data are used in this research)

- Total of 107 adult participants, all are Chinese, gender ratio is about 30% to 70% female to male

- 9 different tasks of scenarios performed by each participant: sit still, head moving, talking, dark light, bright light, long distant, exercising of rope skipping, sit still while holding phone camera, head moving while holding phone camera

- Available biosignal data: time, raw ECG-like BVP recordings, HR, blood oxygen level SpO2. Data are collected using CONTEC CMS60C chest BVP sensor

Issues found:

1. This is a dataset with large numbers of participants and recordings. However, all the participants are from one race: Chinese, thus, the data effectiveness on training models for detecting vital signs from other races is unknown.
2. Many video recordings have their lengths mismatched to their paired physical recordings, and this can be up to a length of 5 seconds. This makes us feel that the recordings are not 100% fully synchronized properly.
3. Not all participants performed all 9 tasks of scenarios, thus the total number of recordings for each participant can be varied.
4. The dataset does not contain RR data, which is a major downside since we cannot utilize it to train a RR model.

#### 5.2.1.3. The MCR dataset

Collected and engineered as a middle-sized dataset by the Multimedia Communication Lab at the University of Ottawa, as a part of our own multi-dimension biosignal data collection, it has the following specifications:

- Videos are recorded by Canon iVIS HF G20 Camcorder (See Figure 16) using in MTS 1080P High Definition format, then compressed to 856x480 videos in MP4 format to save size.

- A total of 192 videos with length varied from 1 to 12 minutes are collected.

- Total of 18 adult participants with mixed races from Caucasian, Middle East, Asian, and Hispanic contributed for the data collection, and 4 out of 18 are female.

- Most recordings are performed when participants sit still. In addition, 6 recordings of intensive stationary bike riding exercise are performed by different participants.

- Available biosignal data: time, ECG, HR, RR. Data are collected using Zephyr Bioharness [64] medical professional grade thoracic expansion sensor planted on a chest belt (See Figure 16).

- It is noteworthy that our physical data collection has a more extended heart rate range compared to others (some of the heart rates were recorded at a speed above 120 BPM).


Figure 16 Canon Camcorder and Zephyr Bioharness chest belt sensor

Issues found:

1. Although high quality videos and biosignal data are available from our own collection, it has the smallest sample pool of participants.
2. The lengths of each video recording are not consistent.
3. Due to the COVID-19 pandemic, our lab is temporarily closed, and thus additional data collection cannot be performed at this point of time.

#### 5.2.1.4. The MMSE-HR dataset [65]

Created as a small sized dataset by the State University of New York at Binghamton, the MMSE-HR dataset contains high resolution image sequences collected from 40 different adult participants with 50-50% male and female in gender ratio. Unfortunately, due to many possible flaws we find in the HR data recordings of the MMSE-HR dataset (For example, heart rate varied from 40-120 during one recording where the video shows the participant is obviously in a sit still pose), we decide to no longer include it in this research.

### 5.2.2. Time-series data pre-generating

The next question is, how do we create a data structure to ensure effective and efficient feature engineering prior to the ML model training stage. Our approach is to build and choose the most optimal ML model within this research scope that primarily covers many of signal processing stages previously. Based on the image preprocessing output done in the previous chapter, we are able to create a ML model which takes 1D time-series data based on any intermediate output we choose, and by doing so, we believe this fits the best of our interest to produce a ML model which can take raw time-series input and replace all signal processing stages as an end-to-end solution. Compared to 3D stack of image sequences, the less amount of data not only makes it relatively easy to train and validate during the modeling process, but also gives a light impact on the system's performance as the overall integration evolves. As a result, instead of working on 3D data stacks, we choose to go with time-series ML modeling.

In fact, both the Video Frame and ROI Acquisition Module and the Vital Sign Data Extraction module can be combined as a pre-data-generating process for producing 150-points raw data using the offline video recordings and biosignal physical data. When a 150-points sliced buffer is produced using corresponding ROI image from a video recording, it is tagged with a value that is averaged from the all ground truth data within the same synchronized time window (See Figure 17). Next, a fail-detection has been implemented to ensure that a 150-points data is only produced when face detection and landmark

prediction are all successfully performed on the corresponding 150 continuous image frames. That means, if one frame fails to perform the ROI cropping, the whole buffer is reset, and time-series data generation is skipped for that time window. Last, we have utilized every video recording and biosignal data from all three datasets to generate the maximum number of possible time-series data.

| 1 | Buffer Counter Index | Participant Number | Ground Truth BPM | Ground Truth RPM | Last Frame Position | FPS | Data#_0 | Data#_1 | Data#_2 | Data#_3 | Data#_4 | Data#_5 | Data#_6 | Data#_7 | Data#_8 | Data#_9 | Data#_10 | Data# |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | p001 | 76 | 19.98 | 151 | 30 | 0.594414 | 0.735851 | 0.70578 | 0.538084 | 0.279372 | 0.560394 | -0.5993 | -0.17752 | -0.33164 | 0.265313 | -0.42522 | -1.07 |
| 3 | 1 | p001 | 79.4 | 19.98 | 302 | 30 | -3.81963 | -4.00547 | -3.75561 | -3.73997 | -3.00684 | -3.0705 | -2.66932 | -2.46122 | -2.42605 | -1.96874 | -1.27373 | -1.13 |
| 4 | 2 | p001 | 75.6 | 19.82 | 453 | 30 | 2.299267 | 2.34618 | 2.058305 | 2.353294 | 2.223642 | 2.103464 | 2.103595 | 1.566565 | 1.545403 | 0.724606 | 0.640815 | 0.825 |
| 5 | 3 | p001 | 88 | 19.9 | 604 | 30 | -0.25911 | -0.29232 | -0.53246 | -0.52651 | -0.00472 | -0.00412 | 0.24569 | 0.24177 | 0.309644 | -0.24031 | -0.3262 | -0.24 |
| 6 | 4 | p001 | 82 | 19.36 | 755 | 30 | -0.27291 | 0.01985 | 0.381935 | 0.286379 | -0.14083 | 0.173552 | 0.306112 | 0.014224 | 0.346415 | 0.403012 | -0.10161 | 0.227 |
| 7 | 5 | p001 | 82.4 | 18.04 | 906 | 30 | 0.255393 | -0.57304 | 0.336107 | -0.12396 | -0.11232 | -0.04719 | 0.396072 | -0.10303 | -0.12251 | -0.04532 | 0.232347 | 0.437 |
| 8 | 6 | p001 | 79.6 | 16.52 | 1057 | 30 | -0.0595 | -0.61586 | -0.53817 | -0.54039 | -0.84706 | -0.24414 | -0.26439 | -0.32772 | -0.2064 | -0.39311 | -0.69558 | -0.41 |
| 9 | 7 | p001 | 78.8 | 14.84 | 1208 | 30 | 0.428045 | -0.14874 | -0.16771 | -0.13162 | -0.38673 | -0.279 | -0.38859 | -0.433 | -0.4486 | -0.48454 | -0.85565 | -0.92 |
| 10 | 8 | p001 | 78.6 | 13.64 | 1359 | 30 | -0.07547 | -0.21786 | 0.395496 | -0.23617 | -0.25152 | -0.17012 | -0.22188 | -0.42771 | -0.31316 | -0.39053 | -0.33508 | -0.36 |

Figure 17 Partially generated output, with tagged HR and RR ground truth data inside the red rectangle

### 5.2.2.1. Sliced buffer length (windows size)

This length needs to be designed in the way that the total amount of data it contains should not be too much data for optimized real-time performance, but still be enough for retaining good accuracy and resolution for the ML model to receive or to be trained on as input. After a few trials, we decided to set the window size at a data length of 150-points, or 5 seconds time slice under a camera frame rate speed at 30FPS. There is 0% percentage stepping between each pair of adjacent windows, which means every sliced buffer will have a unique set of values without overlapping between each other.

### 5.2.2.2. Data engineering

The data engineering steps are as the following:

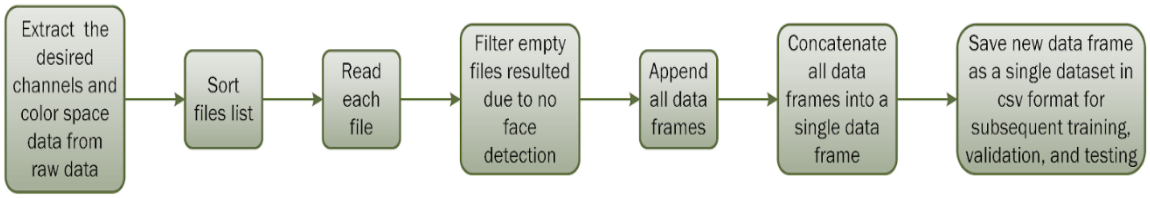



Figure 18 Data engineering steps

First, we extract data from within the raw data produced by the pre-data-generating stage. For example, Green channel in RGB color space for the HR data, and HUE channel in HSV color space for RR data. Next, extracted data files based on individual ROI are combined and sorted together. Files from each participant are then read in the order of one after the next, and empty files due to failed face/landmark detection are filtered out. After data reading is done, we append them all together. Finally, after all participants' data are read, they are concatenated into a single large data frame, and saved as a single dataset in CSV format. At this point, the data is ready to be used for training, validation and testing performed in the next stage.

### 5.2.2.3. Data visualization and analysis

The following histogram in Figure 19 is a summary of total distribution of extracted green channel data based on all engineered datasets, which its statistical summary is provided in Table 2. The figure gives a clear visualization that indicates the majority of collected HR data are generated within a clinical normal range at 60-100 BPM. This also sets the effective and accurate boundaries for validating and testing our models, as they will most likely be trained based on the data inside this range.

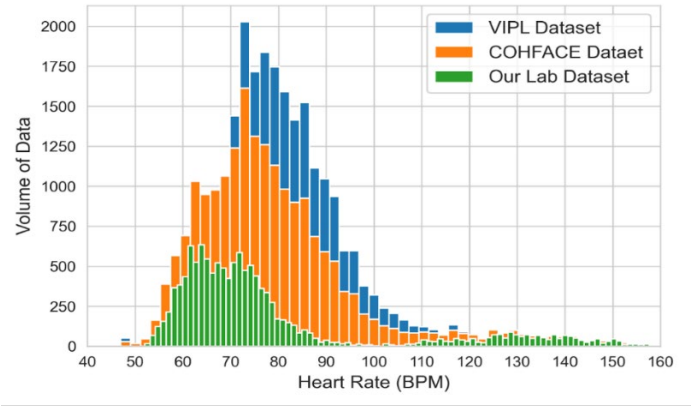



Figure 19 Histogram of each engineered dataset

Table 2 Statistical summary of extracted green channel engineered datasets

| Metric (BPM) | Our Lab | COHFACE | VIPL |
|---|---|---|---|
| **Mean** | 78.4 | 79.6 | 80.8 |
| **Std** | 23.9 | 18.2 | 14.8 |
| **Min** | 52 | 47 | 47 |
| **25%** | 63.8 | 68 | 72 |
| **50%** | 70.8 | 76 | 79 |
| **75%** | 78.8 | 86 | 87.5 |
| **Max** | 165.6 | 255 | 255 |
| **Count (per ROI)** | 12,438 | 19,944 | 24,662 |

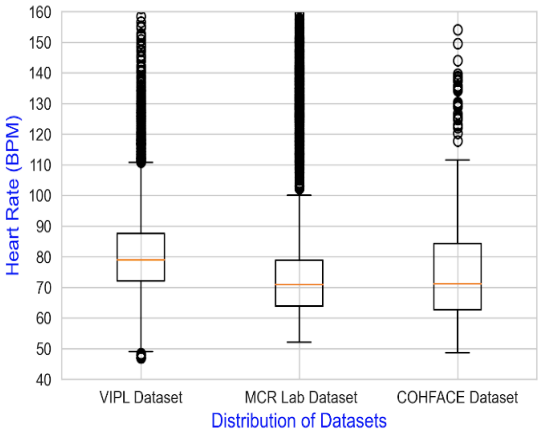



Figure 20 Heart rate data distribution

### 5.2.3. ML algorithms and design

Estimating the real value of HR is considered as a regression predictive modelling problem. The HR estimation module utilizes the power of supervised machine learning and deep supervised learning for producing estimated HR value results directly. This is achieved through designing and experimenting the following machine learning and deep learning algorithms, which are trained, validated, and tested offline using three datasets we have introduced above in 5.2.1.

#### 5.2.3.1. Supporting Vector Regression (SVR)

The SVR is a generalized regression version of Support Vector Machines (SVM) which is one of the most popular and practical algorithms used in supervised machine learning. SVM/SVR demonstrated solid performance in solving classification/regression problems which are characterized by multiple kernels (linear, polynomial, and radial). In this chapter, the SVR machine learning algorithm is considered as the starting point to explore the solution space in HR estimation due to the following major advantages; (i) superior generalization capability, (ii) high estimation accuracy, (iii) simple tuning and (iv) independent computational complexity from the dimensionality of the input space [66].

The SVR model is implemented using python open source library; scikit-learn [67] and designed using the following parameters: kernel='rbf' (Radial Basis Function kernel), C=10 (regularization parameter) , gamma=0.1 (support vectors inverse radius parameter), epsilon=0.1 (learning rate).

#### 5.2.3.2. Multi-Layer Perceptron (MLP) neural network

MLP is considered as the most useful type of deep neural networks that are capable of learning the relationship between input and output data. Our MLP neural network architecture consists of (i) One input layer with 150 input

neurons that will consume all features values for each ground truth as defined by the problem, (ii) Multiple hidden layers with different sizes. The input and hidden layers will use a rectified linear activation function in each node. (iii) One output layer with one neuron and linear activation function to predict (estimate) the real value of HR.

The designed MLP model architecture consists of the following:

- Sequential model with 150 input neurons in the input layer.
- 5 hidden layers with 125, 75, 50, and 25 in each layer respectively with rectified linear activation function in each neuron.
- One hidden layer with one neuron and linear activation function
- Loss function = 'mean_squared_error (mse)'
- Optimizer = 'adam',
- Validation metrics = 'mean absolute error (mae)'

Three regularization techniques are applied to the proposed deep neural networks model to achieve faster training, less overfitting, better generalization and improved prediction output, which are Batch Normalization, dropout, and early stopping techniques. 40% dropout applied after every hidden layer, while the early stopping technique applied to stop the model training when the validation loss performance metric stops improving.

### 5.2.3.3. Long Short-Term Memory (LSTM)

The next deep learning algorithm we explored in estimating HR is based on a well-known successful algorithm used recently in forecasting time-series problems, is called LSTM. LSTM is a special kind of Recurrent Neural Networks (RNN) developed in 1997 [68] to overcome the challenge of short-term memory in RNN; short-term memory caused by the vanishing gradient during backpropagation process. The architect of LSTM consists of a sequential model with one LSTM layer and multiple dense layers and has the following parameters:

- LSTM layer:
  Activation = 'tanh'
  Dropout=0.2
  Recurrent_dropout = 0.2
  Bias_initializer = 'glorot_normal'
  Recurrent_initializer = 'orthogonal'
  Kernel_initializer = 'glorot_uniform'
  Recurrent_activation = 'hard_sigmoid'

- Dense layer:
  Three layers with 64, 32, and 16 neurons in each layer respectively with LeakyReLU (alpha = 0.05) activation function in each neuron and 0.2 dropout
  One hidden layer with one neuron and linear activation function

- Loss function = 'mean_squared_error (mse)'
- Optimizer = 'adam'
- Validation metrics = 'mean absolute error (mae)'

Three regularization techniques applied to the proposed deep LSTM model to achieve faster training, less overfitting, better generalization and improved prediction output, which are Batch Normalization, dropout, and early stopping techniques. 20% dropout applied after LSTM and dense layers, while the early stopping technique applied to stop the model training when the validation loss performance metric stops improving.

#### 5.2.3.4. eXtreme Gradient Boosting (XGBoost)

The last machine learning technique considered for exploration and experimentation is called eXtreme Gradient Boosting (XGBoost), which was developed originally by Chen Tianqi in 2016 [69]. XGBoost is based on a gradient boosting. From its name, boosting refers to the traditional ensemble technique where additive models (decision trees) sequentially added until minimized loss by gradient descent algorithm stops improving when solving classification or regression problems. XGBoost demonstrated exceptional performance in speed and accuracy especially in winning machine learning competitions.

The XGBoost model is developed using an open source library [https://github.com/dmlc/xgboost]. The XGBoost model parameters are:

1. XGB_model: XGBRegressor (XGBoost Regression model)
2. Learning_rate: 0.01
3. N_estimators: 300 (number of trees)
4. Max_depth: 6 (depth of trees)
5. Random_state: 99 (fixed random number to enable reproducible results)
6. Early_stopping_rounds: 10 (number of iterations control early stopping during training).
7. Eval_metric: mae (mean absolute error as a validation metric)

Note: the model performance optimized after tuning hyperparameters no. 2, 3 and 4.

## 5.2.4. Results analysis

The previous machine learning and deep learning models experimented based on the three datasets as shown in Table 3. Which shows that participant's data used during model testing are separated without any overlap with the training dataset.

The following procedure are used to perform model training and testing:

1. Read data frame from engineered dataset (referenced in 5.2.2.2, last stage)
2. Create input feature (150 data points) and target variable (ground truth)
3. Normalize the input features in feature range between 0 and 1 calculated using the following equation:

$$X_{scaled} = X - Min(X) \times Max(X) - Min(X) \quad \text{Equation 3}$$

Where the X represents each feature data point

4. Build custom subsets for training, validation, and testing
5. Build ML/DL model using applicable open source library
6. Fit model to training and validating datasets
7. Test model performance using predict function (regression)
8. Plot model training and validation loss performance during training
9. Print model evaluation metrics
10. Save model parameters in tuple format using pickle open source library

Table 3 Dataset experimentation setup

| | Training Dataset | Test Dataset |
|---|---|---|
| Model1 | MCR lab (70%) | MCR lab (30%) |
| Model2 | MCR lab (70%) + COHFACE (70%) | MCR lab (30%) |
| Model3 | MCR lab (70%) + COHFACE (70%) + VIPL (70%) | MCR lab (30%) |

## 5.2.4.1. Model performance evaluation metrics

Three metrics adopted to evaluate the ML/DL models performance and benchmark it against the DSP one. These metrics are:

1. Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=0}^{N} |y_i^G - y_i^p|}{N} \quad \text{Equation 4}$$

2. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{\sum_{i=0}^{N} \frac{|y_i^G - y_i^p|}{y_i^G} \times 100}{N} \quad \text{Equation 5}$$

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N} (y_i^G - y_i^p)^2}{N}} \quad \text{Equation 6}$$

Where $y_i^G$ stands for the ground truth HR values, and $y_i^P$ for predicted values, respectively.

### 5.2.4.2. Models training, validation and testing

As mentioned previously, two ML and two DL models were trained, validated and tested to investigate and select the best that meet the following developed criteria:

1. Compare the performance of each model during training and testing using metrics Table 3.
2. Plot each learning curve to visualize the model health status during learning.
3. Plot each model HR signal to visualize model sensitivity and accuracy of the estimated HR signal in comparison to the ground truth signal.

Figure 21 shows how the real time performance comparisons are made particularly between different ML algorithms as well as different model configurations. Figure 22-24 show the learning curve for models trained on different combination of datasets using different ML algorithm (except SVR). In Figure 25 a 200-second sample output comparison among signal processing results, ML model results, and ground truth.

Our overall analysis on the performance of all models reported in Table 4 and Figures 22, 23, 24 and 25 showed clearly that XGBoost model #1 which is trained based on only MCR lab dataset provided the best performance even though its evaluation metrics are slightly behind model #3. The justification behind this decision is the following:

- Originally, we designed and trained the model based on all three detests with a training/validation split of (70/30) for Dataset 1 and 3 and kept the remaining 30% of our lab dataset for testing. The model we trained was showing a good performance based on the evaluation metrics but was not sensitive to HR signal changes, so we decided to explore adding each dataset and understand its impact on the model performance.

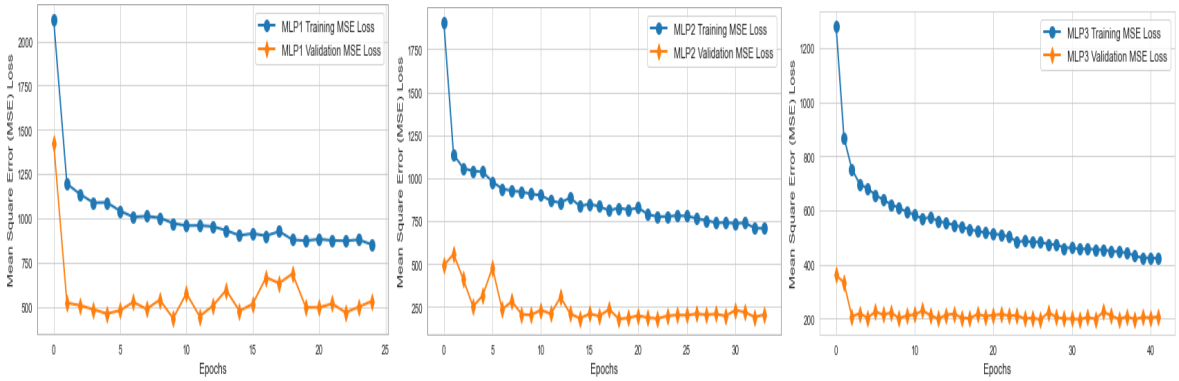
- The sensitivity of model 1 showed best performance in comparison to model 2 and 3 where model 2 suffered high margin of error due to large variance

between estimated signal and ground truth, and model 3 suffered less sensitivity (estimated HR fluctuating around the 75 BPM all the time). The additional added datasets did not contribute to the expected better machine learning accuracy, and we believe the performance reduction was caused by the noise and high video compression, since our MCR dataset is cleaner with better quality of video recorded.


Figure 21 Models benchmarking using two offline videos in progress

Table 4 Summary of evaluation metrics resulted from ML/DL models experimentation

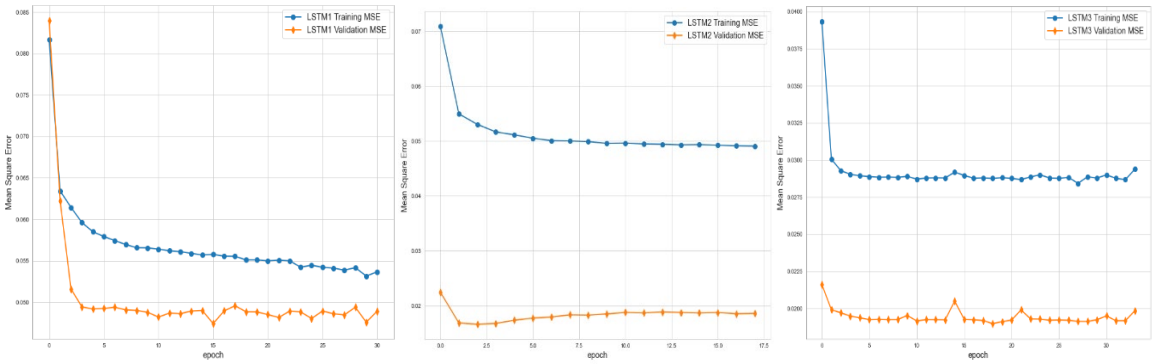
| | SVR | | | MLP | | | LSTM | | | XGBoost | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE |
| Model1 | 7.6 | 9.80% | 9.3 | 7.5 | 10.10% | 9.4 | 8.1 | 10.30% | 9.8 | 6.6 | 8.70% | 8.2 |
| Model2 | 7.3 | 9.40% | 8.9 | 7.8 | 10.40% | 9.6 | 8.6 | 10.90% | 10.2 | 8.3 | 10.60% | 10.1 |
| Model3 | 6.3 | 8.60% | 7.9 | 6 | 8.20% | 7.7 | 8.6 | 10.90% | 10.3 | 5.6 | 7.50% | 7 |
| DSP | 18.3 | 18.50% | 27.5 | 18.3 | 18.50% | 27.5 | 18.3 | 18.50% | 27.5 | 18.3 | 18.50% | 27.5 |




(a) (b) (c)

Figure 22 (a), (b), and (c) The learning curve of MLP model 1, 2, and 3 respectively.

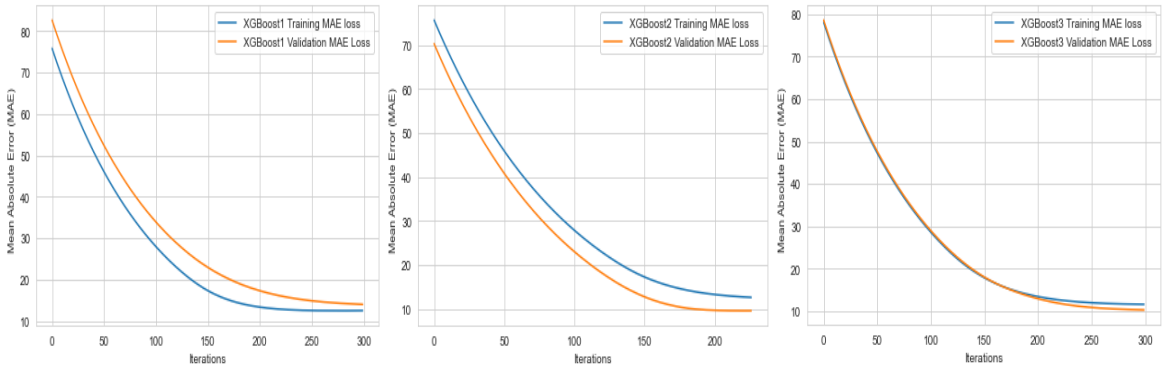
The Figure 22 shows model validation curves of three MLP models trained based on datasets configurations: MCR, MCR+VIPL, MCR+VIPL+COHFACE




(a) (b) (c)

Figure 23 (a), (b), and (c) The learning curve of LSTM model 1, 2, and 3 respectively.

The Figure 23 shows model validation curves of three LSTM models trained based on datasets configurations: MCR, MCR+VIPL, MCR+VIPL+COHFACE




(a) (b) (c)

Figure 24 (a), (b), and (c) The learning curve of XGBoost model 1, 2, and 3 respectively.

The Figure 24 shows model validation curves of three XGBoost models trained based on datasets configurations: MCR, MCR+VIPL, MCR+VIPL+COHFACE

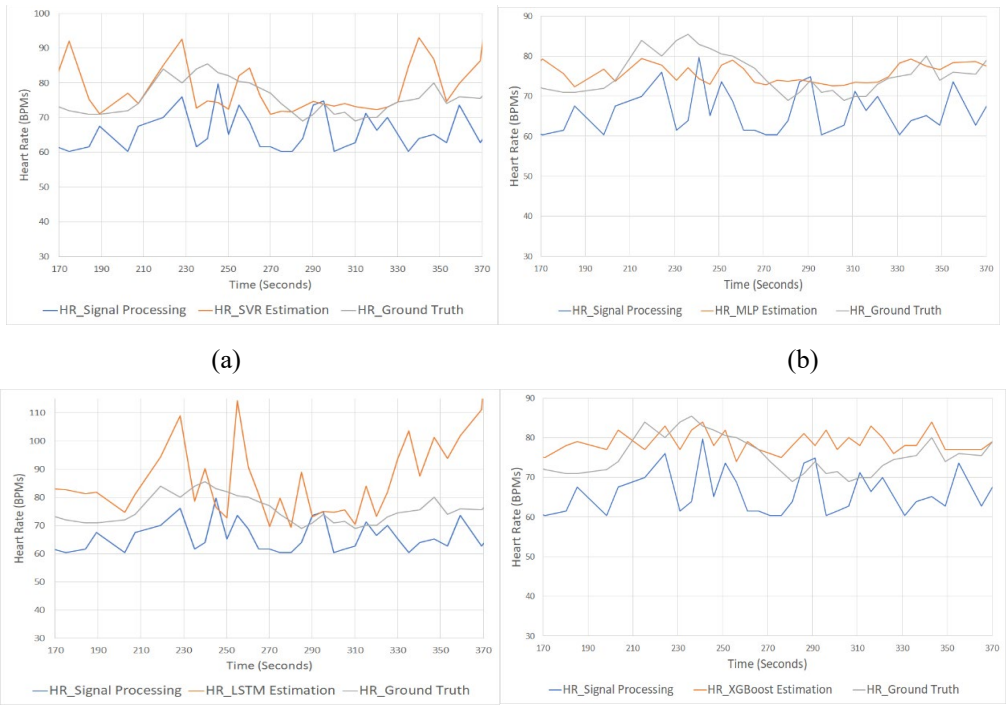



Figure 25 (a), (b), (c) and (d): 200 seconds of sample output comparison among signal processing results, ML model results, and ground truth.

The Figure 25 shows sample testing output of four best models selected from SVR, MLP, LSTM and XGBoost training algorithms

## 5.2.5. Respiratory rate ML modeling

The RR ML design process is very similar to the HR design process we presented above, with an exception that the large VIPL dataset does not come with RR ground truth data and thus needs to be excluded in the model training.

From the HR design we have concluded that based on the validation and testing results, the XGBoost model gives a robust performance and comparably good accuracy. Thus, we only perform RR design using this algorithm in this session.

#### 5.2.5.1. XGBoost model comparison results

We compared the XGBoost models against the signal processing, however, we find the results from XGBoost models are not satisfying since its output is not sensitive enough to give accurate estimation results, especially at the time when the test subject's ground truth has significant trend and fluctuations. We suspect this result can be caused by the following number of reasons:

- The frequency of RR is slower compared to HR. A length of 150-data points sample maybe optimal for training an HR model, but if we want to reach the same accuracy for RR while maintaining the same size of sample, we need to go for lowered sampling rate and longer sampling time. However, the longer sampling time for each buffer slice would reduce our system's performance in the real-time, as the output is then taking longer to be updated.

- The effect of rPPG signal reflected on image intensity change is mainly caused by the respiratory sinus arrhythmia described in [29], where the DC component of BVP signal can be varied at the time of breath-in and breath-out. This is an even weaker signal compared to BVP itself, and thus has a worse impact on the model training process.

- The VIPL dataset does not come with any respiratory data, thus the data we can utilize towards the model training is limited.

In addition, instead of using the same Green channel from RGB color space in HR data extraction, we found the HUE channel from the HSV color space presented by [43] gives better SNR strength when it comes to respiratory signal detection, and thus we switched to the HUE channel and kept signal processing method in estimating the RR value in the end.

Table 5 Summary of evaluation metrics resulted from ML/DL models experimentation for BR estimation

| | XGBoost - BR Estimation | | |
|---|---|---|---|
| | MAE | MAPE | RMSE |
| Model1 | 2.5 | 16.20% | 2.9 |
| Model2 | 4.2 | 23.90% | 4.9 |
| DSP | 4.9 | 26.70% | 5.6 |

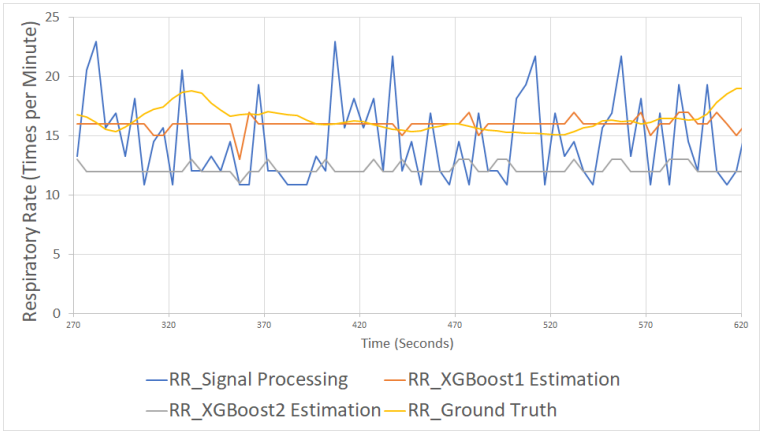



Figure 26 Sensitivity and accuracy of the estimated BR signal resulting from XGBoost models.

The ground truth curve (orange color line) in Figure 26 gives a clear bump at the time around 320 second. However, even though the output data from the XGBoost model (red and gray lines) look more stable compared to output using signal processing, they are not sensitive enough to pick up such a trend.

## 5.3. Limitations and Constraint

Serval constraints and limitation are applied in our research listed as the following:

- The majority of data we used for ML training lies within the range of 65-85 BPM. Thus, the sensitivity of our model for estimating values outside this range is reduced due to the limited amount of data provided.

- Video images from COHFACE dataset are highly compressed and noisy. Therefore, it is hard for the model to recognize the color intensity change between frames, and its contribution to our model training and refinement is debatable.

- Video recordings from all datasets are performed under indoor ambient light conditions, thus the model performance under natural or darkened light conditions are reduced.

- We lack participants from Black and Caucasus races in all datasets, thus the performance of the model under test subjects from these races are not yet validated.

- The COVID-19 pandemic has prevented us from accessing our laboratory facilities. Because of that, additional data collection and model performance testing have not resumed yet.

## 5.4. Final Thoughts and Conclusions

In this chapter, we have created ML models that are intended to replace the signal processing methods for better performance and accuracy. Firstly, we obtained four datasets, which three of them are utilized later for engineering data for ML design purposes. Based on a different combination of datasets, we then built our ML models using four different ML algorithms: SVM, MLP, LSTM and XGBoost. Last we obtained model validation and testing results and made a side-by-side comparison among all models as well as the signal processing method. We had our constraint during the special pandemic period, and we believe the models can have large room for further improvement once the social distancing ban is lifted.

# Chapter 6. Body Temperature Detection and System Integration / Testing

Inside this chapter, we select the best algorithms in HR and RR estimation based on the results obtained from the previous chapter. In addition, an infrared temperature sensor module is synchronized to the system for providing body temperature detection feature in the temperature module. Once the entire system integration is finalized, we conduct more tests and draw further conclusion based on the outcomes received.

## 6.1. Optimized Algorithm for the Final Integration

Base on the testing and evaluation results from the previous chapter, we conclude that the XGBoost model is so far the best approach for estimating the heart rate, given its advantages in sensitivity and accuracy compared to other models as well as the original signal processing method. On the other hand, due to the data and time constraint, our trained ML models for estimating respiratory rate do not perform up to our expectation, and thus we choose to keep the signal processing method in the final integration at this point.

## 6.2. Temperature Detection Module

Two different methods have been explored in our research for obtaining body temperature: Thermography and thermometry measurements.

### 6.2.1. Trial on thermography camera

Originally, we want to build our temperature detection module based on thermography methodology. However, due to the COVID-19 crisis, we faced obstacles such as lack of item availability and delay in parts ordering, and we are only able to receive a FLIR Lepton™ 2.0 thermal camera module (See Figure 27) at a cost within our budget.

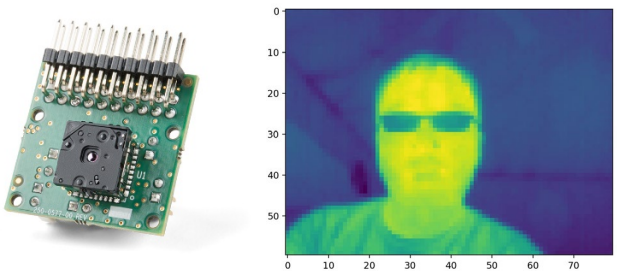



Figure 27 FLIR Lepton™ 2.0 thermal camera module and thermal image output

However, further experiments show that this camera module lacks two important parameters which are crucial for our body temperature detection:

1. No radiometry features

   The radiometry feature is a critical need in the thermography for image associated temperature measurement. Basically, it provides pixel-wise or frame-wise temperature information on top of generated thermal images. Lacking this feature will not only make temperature information hard to extract, but also create difficulties on referencing a base temperature since every temperature measurement made on different background needs to be calibrated with a 35°C blackbody for obtaining correct readings.

2. Accuracy

   At the price point of a low-cost thermal camera, it is hard to obtain a high resolution on the Lepton™ camera as its specification states its accuracy only gives ±5°C @ 25°C. We cannot afford to have a measurement off by an error rate up to 20%, which would potentially bring a significant increase in false-positive rate of fever detection.

In addition to above issues, both of our original face detection and landmark detection modules only works on regular RGB and Black/White images. We need to search for another way to track the face and extract desired ROIs on the thermography images.

As a result, based on our previous research in Chapter 2.3 and the challenges we faced using the Lepton$^{TM}$ 2.0 thermal camera module, in order to get the best balance between the cost and accuracy, we give our favor to infrared thermometry sensors over thermography image cameras.

### 6.2.2. Infrared thermometry sensor

Compared to most of thermography image cameras in the current market, infrared thermometry sensor products are easy to implement, simple to setup, and have the best accuracy per cost ratio. These sensors usually sense electromagnetic waves ranged from 700nm to 14000nm and equipped with photodetectors which can pick up infrared energy emitted by the targeted object [70]. Within the limited availability on the current market, we ordered the OMEGATM OS-MINIUSB miniature infrared senor as our choice of selection (See Figure 28), with a cost nearly the same as the FLIR Lepton$^{TM}$ 2.0 thermal camera.

Figure 28 OMEGATM OS-MINIUSB 20:1 miniature infrared temperature sensor

The sensor comes the following specifications:

- Temperature sensible range: -20 to 1000 $^{O}$C
- Fast response time in 125ms
- USB 2.0 compatible communication interface
- Measurement accuracy with in 1$^{O}$C
- Operation temperature range: 0 to 75 $^{O}$C
- Open Modbus Protocol

### 6.2.3. Sensor placement

To maximize the accuracy of the measurement, an effective measurement distance and sensible area according to its datasheet is show below in Figure 29:

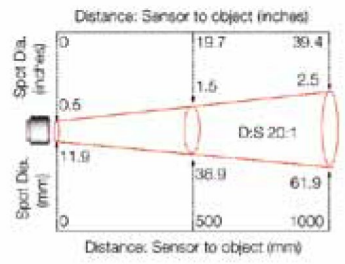



Figure 29 Effective measuring distance for OMEGATM OS-MINIUSB infrared senor [71]

The sensor has a sensible circular plane with a diameter of 36.9mm with subject placed at a distance of 0.5m. This is a match position as we can place the sensor on the top of the rPPG webcam while taking effective data simultaneously. According to [72], the forehead can be an ideal region for the most accurate skin temperature measurement on human facial area (See Figure 30), and thus the sensor needs to be pointed to this region during the test for obtaining the best result.


Figure 30 Forehead is the best thermal effective region for temperature measurement [72]

### 6.2.4. Sensor communication

Our integration of the infrared senor is mainly achieved by the implementing of a Python Serial Communication interface based on Open Modbus protocol. Figure 31 below shows the steps to get a temperature measurement on a distant subject.

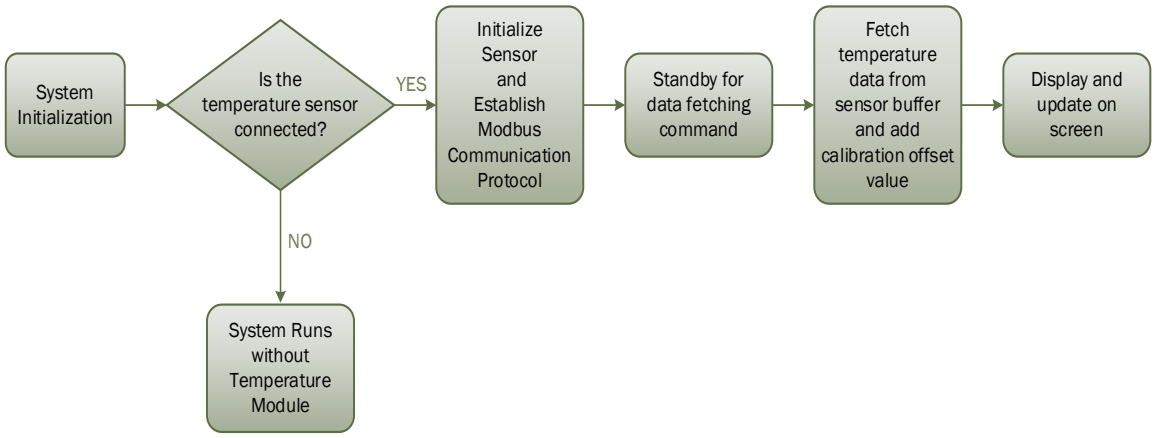



Figure 31 Temperature sensor communication flow chart

As described about, the temperature module is designed to be a flexible attachment to the whole system, and thus the system is still capable to perform the HR and RR estimation using obtained rPPG data through the webcam without a connected temperature sensor.

### 6.2.5. Sensor calibration

We use a clinical thermometer (See Figure 32) that makes contact with the skin of test subject as our ground truth reference for temperature measurement test. At the beginning we discovered the infrared temperature sensor, particularly the one we ordered, is little off calibrated as the detected temperature is lowered for approximately 2 $^{O}$C compared to the reading from the thermometer. We also confirmed our findings by taking measurement at both ice water (ideally 0 $^{O}$C) and boiling water (ideally 100 $^{O}$C) using

factory measurement software provided by OMEGA. Thus, an offset of +2 $^{O}$C is added to the measured values in our implementation for more accurate output.


Figure 32 Thermometer - skin contact type

## 6.3. Integrated System Real Time Performance Evaluation

At this point, as the temperature sensor is added to the system, we have completed our integrated system which is fully functional (See Figure 33)

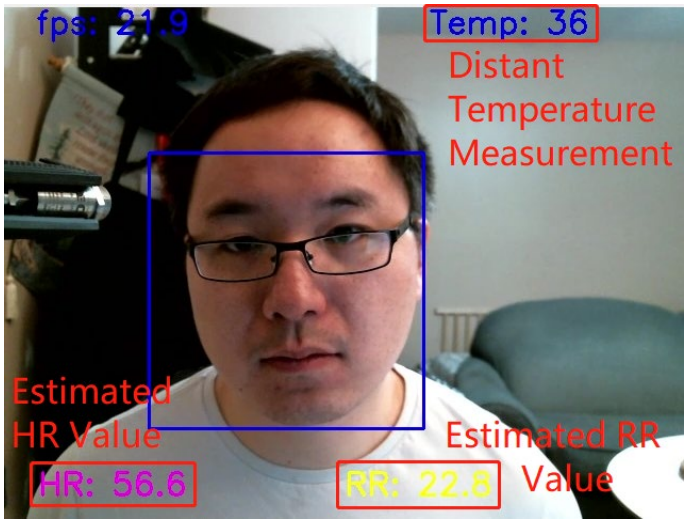



Figure 33 Results of multimodal vital signs are output and updated on display

It is unfortunately and challenging to perform real-time test during the period of COVID pandemic because social distancing prevents people to participant in our experiments. Thus, instead evaluating results from people to people, we decide to take another approach and exam the results obtained under various of scenarios. As test

references, we used a FitBit™ Inspire HR Heart Rate & Fitness Tracker smart wrist band for HR ground truth, self breath-counting for RR, and a commercial grade Braun™ forehead baby thermometer for temperature measurement (See Figure 34).

Figure 34 FitBit™ HR smart band and Braun™ thermometer

Our finding regarding to the performance of the system are listed as the following:

- The system gives the best accuracy under bright ambient light condition. Both outdoor nature light and dark light conditions lower the accuracy for HR, which make sense to us because the XGBoost model used in HR process is trained based on data collected from same indoor ambient light condition.

- Large body movements increase the estimated HR and RR results. This is probably due to the slow shutter speed of the webcam which creates a blurred image. Blurred ROI images can contain a group of pixels with lighter intensity values, which lead to noisy and incorrect change on the AC component of the rPPG data.

- Performance can be varied from machine to machine. Even though the ML model for HR is light weight design, the face detection/landmark prediction modules can take a quite significant amount of computational power to process. Constraint machines with less than ideal hardware will find themselves struggling in running our implementation, thus further performance optimization is still required to speed up the integrated system.

- HR estimation accuracy is at its best when tested under normal sit still conditions. Error becomes large when HR estimation is performed under intensified physical condition as HR of the test subject go up. Nevertheless, this is still under our expectation as the HR model is trained under ground truth data mostly distributed within the range of 60-90 BPMs.

- Wearing makeup and sunscreen still consistently present a major challenge for determine the rPPG signals, because any skin area covered by these layered materials reflects will have the strength of reflective light lowered, and thus their rPPG signals are either absorbed or weakened.

- Sitting distance can also be tricky when real time testing performed. Sitting too far, the body temperature sensor will lose its focus area as the size of sensible plane increase, and rPPG signals also become weaker for detections. Sitting too close, facial movement will be enlarged thus reduce accuracy on the HR and RR.

# Chapter 7. Conclusion and Future Works

## 7.1. Conclusion

To support the worldwide efforts in fighting the current and future waves of the COVID-19 pandemic, we have proposed a complete working prototype for multimodal vital sign estimation in this research. The system is capable of detection the heart rate, respiratory rate and body temperature based on both rPPG signals extracted from facial images sequences and infrared data read from facial skin. Specifically, the initial research is focused on extracting time-series data slice from color-intensity data calculated from tracked offline video image sources. To improve the overall performance of the system, we have adapted both signal processing techniques and several ML models algorithms in both HR and RR estimation processes. All ML models are trained by supervised learning using physical ground truth data from multiple datasets to provide more robust performance in the real time. We also conduct a validation analysis on the performance of the system by applying head-to-head comparison among all methodologies used in this research. Our finding showed that the ML model has made a significant improvement over a traditional signal processing algorithm, particularly by improving the stability of the output on HR estimation. Thus, by deploying the ML model in HR estimation, while keeping signal processing algorithm in RR estimation, we conclude this is the best configuration to be used in the final system integration. Additionally, despite of the facing obstacles due to the COVID crisis during development of the system, we are able to obtain necessary software datasets and hardware components and complete the entire integration of the prototype gradually one step at a time.

The significances of this thesis research are listed as follows:

- A fully functional multimodal vital signs estimation system featuring both real-time and offline contactless HR, RR estimations and body temperature detection

- A complete system integration that consists multiple algorithms including facial feature detection, data extraction, signal processing, ML model design and Serial data communication all together as a whole

- A high accuracy light weight ML model that can replace the majority of the computational signal processing steps and be deployed as a simply end-to-end HR estimation solution

- A performance evaluation based on the comparison among various ML model

## 7.2. Future Works

To extend the potential of our system, further advancement can be made on top of this research:

- As far as supervised learning goes, it is always good news to obtain additional data, either by exploring additional datasets or collecting by our own.
    - The current datasets we obtained do not have sufficient ground truth data evenly distributed across reasonable low to high frequency spectrums, thus data obtained under different conditions such as intensive exercising is highly desirable.
    - Since the VIPL dataset does not contain any RR data, more datasets with RR ground truth data are highly desirable.
    - Additional video data with high quality and less compression is more ideal.
    - Our current datasets do not have large variation on race of their participant, thus data contributed by Black and White people are also desirable.
- Further performance optimization can possibly be achieved in several ways:
    - Adapting more efficient face detection / landmark prediction models can reduce the overall processing time, and thus increase system performance and camera's sampling framerate.
    - Utilizing more powerful hardware with better quality of camera can contribute to both performance and accuracy of the system.
    - Adding motion-based data extraction can serve as an additional feature in ML model design and potentially contribute for accuracy improvement.
- We have not yet explored CNN type of ML designs, and this approach may have deep potential as CNN models is capable of taking advantage of local spatial coherence in multi-dimension images and extracting relevant information at low computational cost.
- Additional vital sign parameters such as blood oxygen saturation level can be added to the system.
- The results of this research can serve as a foundation towards more complex analysis such as emotional detection, lie detection and fatigue detection.

# References


[1] R. Hajar, “The Pulse in Ancient Medicine Part 1,” *Heart Views Off. J. Gulf Heart Assoc.*, vol. 19, no. 1, pp. 36–43, Mar. 2018, doi: 10.4103/HEARTVIEWS.HEARTVIEWS_23_18.

[2] A. C. Y. Tang, “Review of Traditional Chinese Medicine Pulse Diagnosis Quantification,” *Complement. Ther. Contemp. Healthc.*, Oct. 2012, doi: 10.5772/50442.

[3] A. El Saddik, “Digital Twins: The Convergence of Multimedia Technologies,” *IEEE Multimed.*, vol. 25, no. 2, pp. 87–92, Apr. 2018, doi: 10.1109/MMUL.2018.023121167.

[4] “What is RPPG? | Noldus,” *What is RPPG? | Noldus*. https://www.noldus.com/blog/what-is-rppg (accessed Jun. 18, 2020).

[5] Y. Funabashi, “COVID-19 and the advent of a ‘contactless economic system,’” *The Japan Times*, May 10, 2020. https://www.japantimes.co.jp/opinion/2020/05/10/commentary/world-commentary/covid-19-advent-contactless-economic-system/ (accessed Jun. 18, 2020).

[6] R. Sinhal, K. Singh, and M. M. Raghuwanshi, “An Overview of Remote Photoplethysmography Methods for Vital Sign Monitoring,” *Computer Vision and Machine Intelligence in Medical Image Analysis Advances in Intelligent Systems and Computing* 2019, pp. 21–31., doi:10.1007/978-981-13-8798-2_3

[7] A. A. Kamshilin and N. B. Margaryants, “Origin of Photoplethysmographic Waveform at Green Light,” *Phys. Procedia*, vol. 86, pp. 72–80, Jan. 2017, doi: 10.1016/j.phpro.2017.01.024.

[8] U. Rubins, A. Miscuks, O. Rubenis, R. Erts, and A. Grabovskis, “The analysis of blood flow changes under local anesthetic input using non-contact technique,” Nov. 2010, vol. 2, pp. 601–604, doi: 10.1109/BMEI.2010.5640023.

[9] T. A. Neff, “Routine Oximetry,” *Chest*, vol. 94, no. 2, p. 227, Aug. 1988, doi: 10.1378/chest.94.2.227a.

[10] U. F. O. Themes, “Engineering and Clinical Aspects of Photoplethysmography,” *Thoracic Key*, Mar. 04, 2017. https://thoracickey.com/engineering-and-clinical-aspects-of-photoplethysmography/ (accessed Jun. 18, 2020).

[11] C. Wang, T. Pun, and G. Chanel, “A Comparative Survey of Methods for Remote Heart Rate Detection From Frontal Face Videos,” *Front. Bioeng. Biotechnol.*, vol. 6, 2018, doi: 10.3389/fbioe.2018.00033.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.,” *Opt. Express*, vol. 18, no. 10, pp. 10762–10774, May 2010, doi: 10.1364/OE.18.010762.

[13] D. Da He, E. S. Winokur, and C. G. Sodini, “A continuous, wearable, and wireless heart monitor using head ballistocardiogram (BCG) and head electrocardiogram (ECG),” *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2011, pp. 4729–4732, 2011, doi: 10.1109/IEMBS.2011.6091171.

[14] U. S. Freitas, “Remote Camera-based Pulse Oximetry,” *eTELEMED 2014*, 2014. /paper/Remote-Camera-based-Pulse-Oximetry-Freitas/c6bfcb672e1d9cbf4009564a4e55258081a54865 (accessed Jun. 18, 2020).
[15] U. Bal, “Non-contact estimation of heart rate and oxygen saturation using ambient light,” *Biomed. Opt. Express*, vol. 6, no. 1, pp. 86–97, Jan. 2015, doi: 10.1364/BOE.6.000086.
[16] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A. E. Saddik, “EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video,” *IEEE Trans. Multimed.*, vol. 21, no. 7, pp. 1778–1787, Jul. 2019, doi: 10.1109/TMM.2018.2883866.
[17] H. Rahman, M. Ahmed, S. Begum, and P. Funk, “Real Time Heart Rate Monitoring from Facial RGB Color Video Using Webcam,” *9th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS)*, Malmö, Sweden, May 2016.
[18] Q. Fan and K. Li, “Non-contact remote estimation of cardiovascular parameters,” *Biomed. Signal Process. Control*, vol. 40, pp. 192–203, Feb. 2018, doi: 10.1016/j.bspc.2017.09.022.
[19] H. Demirezen and C. E. Erdem, “Remote Photoplethysmography Using Nonlinear Mode Decomposition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 1060–1064, doi: 10.1109/ICASSP.2018.8462538.
[20] A. Osman, J. Turcot, and R. E. Kaliouby, “Supervised learning approach to remote heart rate estimation from facial videos,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015, vol. 1, pp. 1–6, doi: 10.1109/FG.2015.7163150.
[21] M. Bian, B. Peng, W. Wang, and J. Dong, “An Accurate LSTM Based Video Heart Rate Estimation Method,” in *Pattern Recognition and Computer Vision*, Cham, 2019, pp. 409–417, doi: 10.1007/978-3-030-31726-3_35.
[22] H. Monkaresi, R. A. Calvo, and H. Yan, “A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam,” *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1153–1160, Jul. 2014, doi: 10.1109/JBHI.2013.2291900.
[23] H. Ghanadian, M. Ghodratigohar, and H. Al Osman, “A Machine Learning Method to Improve Non-Contact Heart Rate Monitoring Using an RGB Camera,” *IEEE Access*, vol. 6, pp. 57085–57094, 2018, doi: 10.1109/ACCESS.2018.2872756.
[24] Y. Hsu, Y.-L. Lin, and W. Hsu, “Learning-based heart rate detection from remote photoplethysmography features,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4433–4437, doi: 10.1109/ICASSP.2014.6854440.
[25] L. Liu, L. Lu, J. Luo, J. Zhang, and X. Chen, “Enhanced Eulerian video magnification,” in *2014 7th International Congress on Image and Signal Processing*, Oct. 2014, pp. 50–54, doi: 10.1109/CISP.2014.7003748.
[26] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, and S.-H. Chang, “Siamese-rPPG network: remote photoplethysmography signal estimation from face videos,” in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, Brno Czech Republic, Mar. 2020, pp. 2066–2073, doi: 10.1145/3341105.3373905.

[27] E. Brophy, W. Muehlhausen, A. F. Smeaton, and T. E. Ward, “Optimised Convolutional Neural Networks for Heart Rate Estimation and Human Activity Recognition in Wrist Worn Sensing Applications,” *ArXiv200400505 Cs Eess Stat*, Mar. 2020, Accessed: Jun. 18, 2020. [Online]. Available: http://arxiv.org/abs/2004.00505.

[28] X. Niu, H. Han, S. Shan, and X. Chen, “VIPL-HR: A Multi-modal Database for Pulse Estimation from Less-constrained Face Video,” *ArXiv181004927 Cs*, Nov. 2018, Accessed: Jun. 18, 2020. [Online]. Available: http://arxiv.org/abs/1810.04927.

[29] M. Chen, Q. Zhu, H. Zhang, M. Wu, and Q. Wang, “Respiratory Rate Estimation from Face Videos,” *2019 IEEE EMBS Int. Conf. Biomed. Health Inform. BHI*, pp. 1–4, May 2019, doi: 10.1109/BHI.2019.8834499.

[30] R. Janssen, W. Wang, A. Moço, and G. de Haan, “Video-based respiration monitoring with automatic region of interest detection,” *Physiol. Meas.*, vol. 37, no. 1, pp. 100–114, Jan. 2016, doi: 10.1088/0967-3334/37/1/100.

[31] F. Braun, A. Lemkaddem, V. Moser, S. Dasen, O. Grossenbacher, and M. Bertschi, “Contactless Respiration Monitoring in Real-Time via a Video Camera,” in *EMBEC & NBC 2017*, Singapore, 2018, pp. 567–570, doi: 10.1007/978-981-10-5122-7_142.

[32] K.-Y. Lin, D.-Y. Chen, and W.-J. Tsai, “Image-Based Motion-Tolerant Remote Respiratory Rate Evaluation,” *IEEE Sens. J.*, vol. 16, no. 9, pp. 3263–3271, May 2016, doi: 10.1109/JSEN.2016.2526627.

[33] Q.-V. Tran, S.-F. Su, and V.-T. Nguyen, “Pyramidal Lucas—Kanade-Based Noncontact Breath Motion Detection,” *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 50, no. 7, pp. 2659–2670, Jul. 2020, doi: 10.1109/TSMC.2018.2825458.

[34] D. Shao, Y. Yang, C. Liu, F. Tsow, H. Yu, and N. Tao, “Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 11, pp. 2760–2767, Nov. 2014, doi: 10.1109/TBME.2014.2327024.

[35] Wu, Hao-Yu, *et al.*, “Eulerian Video Magnification for Revealing Subtle Changes in the World.” *ACM Transactions on Graphics*, vol. 31, no. 4, 2012, pp. 1–8., doi:10.1145/2185520.2185561.

[36] X. Yang and T. Bourlai, “Video-Based Human Respiratory Wavelet Extraction and Identity Recognition,” in *Surveillance in Action: Technologies for Civilian, Military and Cyber Surveillance*, P. Karampelas and T. Bourlai, Eds. Cham: Springer International Publishing, 2018, pp. 51–75, doi: 10.1007/978-3-319-68533-5_3.

[37] H. Ghanadian *et al.*, “A Machine Learning Method to Improve Non-Contact Heart Rate Monitoring Using an RGB Camera.” *IEEE Access*, vol. 6, 2018, pp. 57085–57094., doi:10.1109/access.2018.2872756.

[38] D. M. Tveit, K. Engan, I. Austvoll, and Ø. Meinich-Bache, “Motion based detection of respiration rate in infants using video,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 1225–1229, doi: 10.1109/ICIP.2016.7532553.

[39] A. P. Prathosh, P. Praveena, L. K. Mestha, and S. Bharadwaj, “Estimation of Respiratory Pattern From Video Using Selective Ensemble Aggregation,” *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2902–2916, Jun. 2017, doi: 10.1109/TSP.2017.2664048.

[40] L. Mirmohamadsadeghi, S. Fallet, V. Moser, F. Braun, and J.-M. Vesin, “Real-time respiratory rate estimation using imaging photoplethysmography inter-beat intervals,” in *2016 Computing in Cardiology Conference (CinC)*, Sep. 2016, pp. 861–864, doi:10.22489/cinc.2016.249-283.
[41] A. Guazzi *et al.*, “Non-contact measurement of oxygen saturation with an RGB camera,” *Biomed. Opt. Express*, vol. 6, pp. 3320–38, Sep. 2015, doi: 10.1364/BOE.6.003320.
[42] A. de Fátima Galvão Rosa and R. C. Betini, “Noncontact SpO2 Measurement Using Eulerian Video Magnification,” *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2120–2130, May 2020, doi: 10.1109/TIM.2019.2920183.
[43] S. Sanyal and K. K. Nundy, “Algorithms for Monitoring Heart Rate and Respiratory Rate From the Video of a User’s Face,” *IEEE J. Transl. Eng. Health Med.*, vol. 6, pp. 1–11, 2018, doi: 10.1109/JTEHM.2018.2818687.
[44] R. Palaniappan, K. Sundaraj, and S. Sundaraj, “A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals,” *BMC Bioinformatics*, vol. 15, p. 223, Jun. 2014, doi: 10.1186/1471-2105-15-223.
[45] D. D. Soerensen and L. J. Pedersen, “Infrared skin temperature measurements for monitoring health in pigs: a review,” *Acta Vet. Scand.*, vol. 57, no. 1, Feb. 2015, doi: 10.1186/s13028-015-0094-2.
[46] D. Formenti *et al.*, “Thermal imaging of exercise-associated skin temperature changes in trained and untrained female subjects,” *Ann. Biomed. Eng.*, vol. 41, no. 4, pp. 863–871, Apr. 2013, doi: 10.1007/s10439-012-0718-x.
[47] G. Tanda, “The use of infrared thermography to detect the skin temperature response to physical activity,” *J. Phys. Conf. Ser.*, vol. 655, p. 012062, Nov. 2015, doi: 10.1088/1742-6596/655/1/012062.
[48] R. Gade and T. B. Moeslund, “Thermal cameras and applications: a survey,” *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, Jan. 2014, doi: 10.1007/s00138-013-0570-5.
[49] S. Yu, S. Hu, V. Azorin-Peris, J. A. Chambers, Y. Zhu, and S. E. Greenwald, “Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise,” *J. Biomed. Opt.*, vol. 16, no. 7, p. 077010, Jul. 2011, doi: 10.1117/1.3602852.
[50] G. Sun *et al.*, “Remote sensing of multiple vital signs using a CMOS camera-equipped infrared thermography system and its clinical application in rapidly screening patients with suspected infectious diseases,” *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.*, vol. 55, pp. 113–117, Feb. 2017, doi: 10.1016/j.ijid.2017.01.007.
[51] B. Wei, X. He, C. Zhang, and X. Wu, “Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions,” *Biomed. Eng. Online*, vol. 16, no. 1, p. 17, Jan. 2017, doi: 10.1186/s12938-016-0300-0.
[52] W. Chen and D. McDuff, “DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks,” May 2018, Accessed: Jun. 18, 2020. [Online]. Available: https://arxiv.org/abs/1805.07888v2.

[53] X. Ma, *et al.*, “Remote Photoplethysmography (RPPG) for Contactless Heart Rate Monitoring Using a Single Monochrome and Color Camera.” *Lecture Notes in Computer Science Smart Multimedia*, 2020, pp. 248–262., doi:10.1007/978-3-030-54407-2_21.
[54] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, vol. 1, p. I-511-I–518, doi: 10.1109/CVPR.2001.990517.
[55] “OpenCV.org.” https://opencv.org/about/ (accessed Jun. 18, 2020).
[56] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, Jun. 2014, pp. 1867–1874, doi: 10.1109/CVPR.2014.241.
[57] “Facial landmarks with dlib, OpenCV, and Python,” *PyImageSearch*, Apr. 03, 2017. https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/ (accessed Jun. 18, 2020).
[58] S. Fallet, V. Moser, F. Braun, and J.-M. Vesin, “Imaging photoplethysmography: What are the best locations on the face to estimate heart rate?,” in *2016 Computing in Cardiology Conference (CinC)*, Sep. 2016, pp. 341–344, doi:10.22489/cinc.2016.098-236.
[59] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, and K.-W. Cheung, “Block-based adaptive ROI for remote photoplethysmography,” *Multimed. Tools Appl.*, vol. 77, no. 6, pp. 6503–6529, Mar. 2018, doi: 10.1007/s11042-017-4563-7.
[60] “2 easy, accurate ways to measure your heart rate,” *Mayo Clinic*. https://www.mayoclinic.org/heart-rate/expert-answers/faq-20057979 (accessed Jun. 18, 2020).
[61] L. Levesque, “Nyquist sampling theorem: Understanding the illusion of a spinning wheel captured with a video camera,” *Phys. Educ.*, vol. 49, p. 697, Nov. 2014, doi: 10.1088/0031-9120/49/6/697.
[62] “Data Analysis - Pearson’s Correlation Coefficient.” http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442 (accessed Jun. 18, 2020).
[63] “Welcome to BioSPPy — BioSPPy 0.6.1 documentation.” https://biosppy.readthedocs.io/en/stable/index.html (accessed Jun. 18, 2020).
[64] “The System | Zephyr™ Performance Systems.” https://www.zephyranywhere.com/system (accessed Jun. 18, 2020).
[65] “3D Facial Expression Database - Binghamton University.” http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html (accessed Jun. 18, 2020).
[66] M. Awad and R. Khanna, “Support Vector Regression,” in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, M. Awad and R. Khanna, Eds. Berkeley, CA: Apress, 2015, pp. 67–80, doi: 10.1007/978-1-4302-5990-9.
[67] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn.* Res. 12, 2825–2830, Accessed: Jun. 18, 2020. [Online]. Available: https://arxiv.org/abs/1201.0490.

[68] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[69] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[70] “Infrared Temperature Sensors - Sure Controls.” https://www.surecontrols.com/infrared-temperature-sensors/ (accessed Jun. 18, 2020).
[71] “USB Infrared Temperature Sensor for Benchtop, Laboratory and Education.” https://www.omega.ca/en/sensors-and-sensing-equipment/temperature/sensors/infrared-sensors/os-miniusb/p/OS-MINIUSB-SN201?gclid=Cj0KCQjw6sHzBRCbARIsAF8FMpWVmvvpoBz4cPJIMssWCQkziEMOZcxUP4doM5zf8rxwxwxXlAHxXjgaAs7jEALw_wcB&gclsrc=aw.ds (accessed Jun. 18, 2020).
[72] E. Y. K. Ng, G. J. L. Kawb, and W. M. Chang, “Analysis of IR thermal imager for mass blind fever screening,” *Microvasc. Res.*, vol. 68, no. 2, pp. 104–109, Sep. 2004, doi: 10.1016/j.mvr.2004.05.003.

# Appendix

## Software Choice and Coding Environment

The entire software of our research is implemented in Python version 3.5 under Microsoft Windows 10 version 2004. Below are the details of the choices we pick for our coding environment:

- PyCharm IDE: 2020.1.2
- Anaconda3: 4.8.3
- Tensorflow: 2.1.0 GPU
- Nvidia CUDA Development: 10.2
- OpenCV: 4.2.0
- Scikit-Learn: 0.22.1

## Hardware Setup

Our implementation has been tested on several PCs come with different hardware configurations, mainly on two of them with the following specifications:

Setup 1: Asus Laptop

- CPU: Intel Core I5 8440U 1.7Ghz
- GPU: Nvidia MX150
- RAM: 16GB
- Hard drive: 512GB SSD

Setup 2: Desktop PC

- CPU: Intel Core i7 8700K 3.7Ghz
- GPU: Nvidia Geforce 1070
- RAM: 32GB
- Hard drive: 1TB SSD

In addition to the above hardware, we have used BTChoic$^{TM}$ and wrist contact smart bands as our vital sign ground truth comparison reference when real time test is performed.

# Project Setup

The research implementation is compatible on laptop or desktop computer equipped with a regular webcam, a USB port (for temperature sensor connection), and with Python 3.5 and related libraries installed. A fast CPU and large size of RAM are preferred for speeding up the frame rate of the camera. For an ideal vital sign estimation, the test person should have a clean facial skin exposed without make ups or sunscreen, and the light on the face should be equally bright. They need to breathe normally, and maintain a stable pose sitting or standing in front of the camera and sensor within a distance of 0.5 to 1.5 meters (See Figure 35).

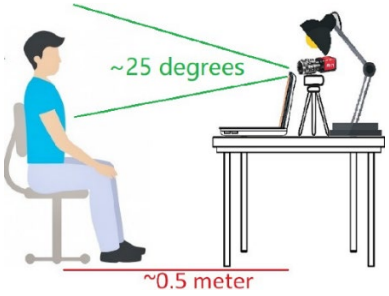



Figure 35 Real time vital signs estimation setup