INFLAMMATORY BIOMARKER ANALYSIS FROM WEARABLE SWEAT PATCHES VIA SMARTPHONE-BASED IMAGE PROCESSING

SHAHAK ROZENBLAT

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING & COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

FEBRUARY, 2026

© SHAHAK ROZENBLAT, 2026

Abstract

Remote and accessible medical screening is a critical component of modern healthcare, particularly for the early detection of systemic inflammation. Detection of inflammatory biomarkers has been shown to correlate with the presence and progression of a wide range of serious diseases and clinical conditions. Wearable sweat-based biosensors offer a promising, non-invasive alternative to traditional blood sampling. However, their deployment for quantitative point-of-care testing (POCT) remains challenging due to limitations in reliably translating visual signals into quantitative measurements. In this thesis, we present a computational pipeline for the quantitative colourimetric analysis of inflammatory biomarkers in wearable sweat-based biosensors. We demonstrate the feasibility of the pipeline to simultaneously measure three key inflammatory biomarkers, Interleukin-6 (IL-6), Interleukin-1beta (IL-1beta), and C-reactive protein (CRP), from pictures of a wearable sweat patch that could be captured using a smartphone. To address the challenges of smartphone-based imaging, such as variability in different lighting and camera sensors, we integrated a custom calibration layer featuring ArUco markers and 49 colour reference swatches into the wearable sweat patch. The proposed pipeline leverages these features to perform image alignment and then a two-stage normalization process that consists of luminance histogram matching and linear regression-based colour correction. A lightweight U-Net model is employed to segment the regions of the sweat patch, followed by a feature extraction algorithm operating in the CIELAB colour space to compute a colourimetric measurement (ΔE). Experimental evaluation demonstrates that our two-step normalization reduces measurement variability by approximately 70% across diverse lighting conditions. We further show that measurements remain consistent across different smartphone devices and capture distances. In a clinical validation study involving human participants ($N = 8$), the pipeline successfully differentiated between healthy controls and patients with elevated inflammation. Notably, the extracted ΔE measurement outperformed standard Enzyme-Linked Immunosorbent Assay (ELISA) measurements in binary classification for IL-6, achieving an area under the receiver operating characteristic curve of 1.0 for CRP and 0.9375 for IL-6. Overall, we show that our proposed pipeline is feasible, which marks a pivotal step towards a rapid, reliable, and accessible solution for non-invasive inflammatory screenings for POCT, including for sweat patches.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Salahandish, for her invaluable guidance, patience, and support throughout this research. Her expertise and mentorship were instrumental in shaping this thesis and my development as a researcher.

To my parents, thank you for your continuous support throughout my master's and always believing in me. To my siblings, thank you for the much-needed distractions. A very special thanks to Hannah, who provided so much encouragement and support throughout my master's, and who initially pushed me to pursue it in the first place.

To "the bois," thank you for being based as always and for keeping me sane. Thank you Chadit for setting up, deploying, and maintaining the server that hosts the pipeline presented in this work.

Last but not least, I would like to thank everyone at LAB-HA for their collaboration, insightful discussions, and for making the lab a supportive and enjoyable environment to work in.

A final acknowledgment to my dog, Riley, for being cute (I would say more but that's really his best perk).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Remote and Accessible Medical Screening

Remote and accessible medical screening is becoming increasingly important in modern healthcare. For example, remote access to healthcare screening was essential in controlling the outbreak of coronavirus disease-19 (COVID-19) when travel was prohibited or restricted during the pandemic [1, 2]. It is also vital to increase accessibility of healthcare screening for individuals living in remote communities or those with accessibility challenges, such as older individuals. Remote screening tools can serve as an efficient first step in the clinical pathway, allowing faster preliminary assessments that can be conducted as early as possible in the disease process and facilitating referrals to specialized medical services. With early-stage screening, strain on healthcare resources could be alleviated, such as by reducing unnecessary specialist consultations

Remote medical screening falls under the umbrella of point-of-care testing (POCT), which refers to any medical screening test that is performed outside of a conventional clinical laboratory, and instead at or near the location of the patient [3, 4]. The core value of POCT in the context of screening is decentralized and rapid assessment, delivering timely results that facilitate enhanced patient monitoring and triage. A significant development enabling widespread POCT is the use of wearable medical devices (WMDs) [5]. A WMD is defined as a device that can be worn on the body to perform a specific medical function, such as monitoring or collecting physiological data. These devices provide non/semi-invasive healthcare monitoring and screening, which is a key driver behind the anticipated market growth of approximately 27% from 2020 to 2027 [6].

However, the vast majority of commercially available WMDs that are currently available are focused on general health and fitness applications. These devices, such as the Apple Watch or Fitbit, excel at tracking easily measurable physiological parameters like heart rate, sleep patterns, and steps [7]. While valuable for wellness, a significant gap remains in the widespread availability of WMDs for complex disease monitoring or early screening of serious conditions. Monitoring for these conditions requires the analysis of specific biochemical markers, which are historically challenging to detect non-invasively. An example of these complex markers is the group of inflammatory biomarkers, which are measurable proteins that signal inflammation, which is a key part of the body's immune response [8]. Measuring these biomarkers is essential in diagnosing, monitoring, and predicting conditions such as cancer, neurodegenerative diseases, and infections [8, 9, 10]. Inflammatory biomarkers can be detected in blood, sweat, tissue, and other biological fluids. Unlike blood sampling, which is invasive, or tears/saliva, which are often harder to collect reliably, researchers have increasingly focused on sweat as a practical and continuous source of biomarker information [11]. This shift has led to the emergence of sweat-based wearable devices that are designed to monitor inflammatory biomarkers for the early screening and management of diseases where inflammation is a key driver [12, 13, 14]. However, a fundamental technical challenge remains in translating their signals for reliable, quantitative, and easily interpretable measurements for POCT. This requires reliably translating the device's inherent signal into precise concentration data (often measured in picograms or nanograms per milliliter) that is independent of the external environment while maintaining a process that is accessible outside a laboratory setting.

1.1.2 Sweat Patch

In this thesis, we introduce a systematic algorithmic pipeline designed for POCT. The pipeline autonomously and simultaneously measures signals from multiple inflammatory biomarkers using a smartphone-captured image of a microfluidic wearable sweat patch. This patch serves as the physical platform that generates the raw data used throughout the computational work of this thesis. It is a thin, flexible device worn directly on the skin, where it collects small quantities of sweat that flow into five sensing chamber regions (SCRs).

These five SCRs enable colourimetric detection of key inflammatory biomarkers. Three chambers are specifically functionalized to detect and simultaneously measure Interleukin-6 (IL-6), Interleukin-1 beta (IL-1beta), and C-reactive protein (CRP). The remaining two chambers serve as negative and positive controls to ensure the validity and reliability of the chemical

reactions.

Our wearable sweat patch is a type of lateral flow immunoassay (LFIA), a paper-based device in which the presence of a target analyte produces a visible colour change [15]. Specifically, our device employs colourimetric assays, a technique in which the presence and concentration of an analyte (i.e., the specific substance that is being measured) are determined by observing the intensity of a colour change in each SCR. In our design, this colourimetric response is achieved using gold nanoparticles (AuNPs)–conjugated secondary antibodies that bind to target antigens in sweat. Upon specific capture at the test line, AuNP-labeled antigen–antibody complexes accumulate, producing a visible color change. The intensity of this colorimetric signal is proportional to the biomarker concentration in the sweat sample, enabling semi-quantitative visual detection.

The patch is therefore designed so that each SCR produces a colour response when its target biomarker is present, resulting in a dark-red hue at higher concentrations. These visual signals serve as the raw data that our algorithmic pipeline later processes for quantitative POCT. The structure and layout of all five chambers (IL-6, IL-1beta, CRP, Negative Control, and Positive Control) are illustrated in Figure 1.1.

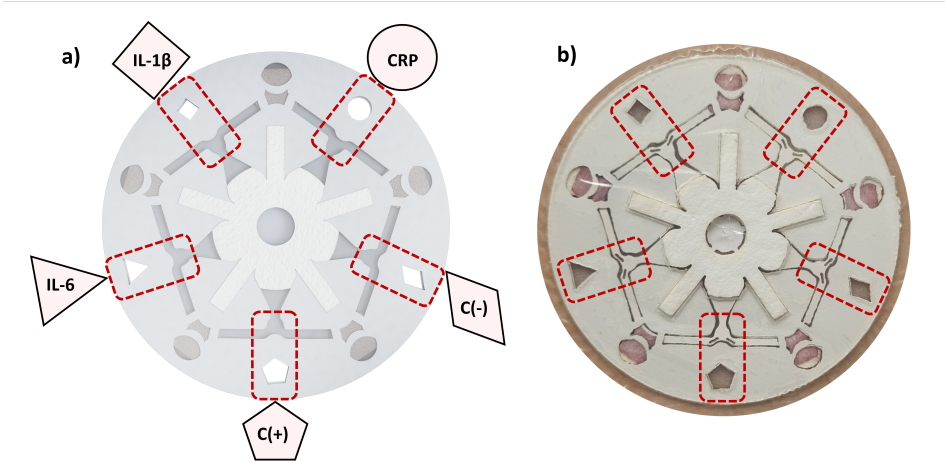


Figure 1.1: The physical platform utilized for sweat collection and colourimetric detection of inflammatory biomarkers. a) and b) show the schematic and real image of the device, respectively, with its five labelled Sensing Chamber Regions (SCRs). Three chambers target specific biomarkers (IL-6, IL-1beta, and CRP) and two chambers serve as internal controls (Negative and Positive).

1.1.3 Smartphone for Patch Analysis

For the purpose of enabling accessible and remote POCT with the wearable sweat patch, we aim to ensure that the tools used for data acquisition and processing are affordable, widely available, and easy to use. To achieve this, we rely on a standard smartphone to perform both data acquisition and processing, all through our custom mobile application. This application is designed to provide a user-friendly experience, handling everything from image capture to data processing and reporting. The ultimate goal is for a user to wear the patch for the required period, then use any smartphone to access our mobile application, where they can take a photo of the patch, submit it for processing, and receive reliable readings for all three biomarkers within seconds. These results will be immediately available for viewing, saving, and further analysis.

For data acquisition, we use the common smartphone camera as the main tool for data collection. Modern smartphone cameras offer high pixel density, typically ranging from 12 megapixels (MP) to 108 MP [16]. For instance, a common 12 MP sensor provides an image resolution of approximately 4000×3000 pixels. Therefore, modern smartphone cameras are fully capable of capturing high-resolution, detailed images. However, smartphone-based image analysis still comes with critical limitations that need to be considered when designing an effective POCT. For example, environmental factors (such as inconsistent lighting and background) and camera-specific differences (including manufacturing bias, angle, and zoom) lead to significant variability in smartphone imaging [17, 18]. These uncontrolled variables can directly undermine the quantitative accuracy that is required for successful POCT.

To address this challenge of reliable smartphone-based imaging, and to ensure proper functionality between the patch and the computational pipeline, we created a calibration layer as the topmost layer of the patch (Figure 1.2). The calibration layer is used to limit variability and precisely identify the patch and its internal regions from a smartphone-captured image. The layer features two main components:

1. ArUco markers: Black-and-white fiducial markers that provide stable reference points for precise patch extraction and biomarker region identification.

2. 49 colour swatches: These serve as internal reference standards to facilitate digital correction for colour variations introduced during image capture.

The complete details of the calibration layer design and function are further described in Section 3.2.

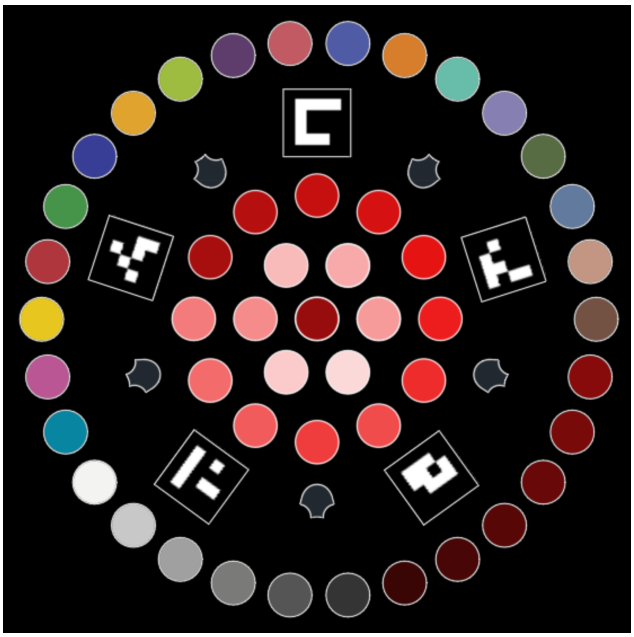


Figure 1.2: Calibration layer attached to the top of the wearable sweat patch, consisting of five ArUco markers and 49 colour reference swatches.

1.2 Thesis Contribution

In this thesis, we propose an algorithmic pipeline for the simultaneous quantification of multiple inflammatory biomarkers from colourimetric signals that are captured using a smartphone image of a wearable sweat patch. The proposed pipeline is designed to provide rapid and fully automated POCT, delivering biomarker concentration estimates within seconds through an intuitive mobile application.

To our knowledge, this work presents the first comprehensive smartphone-based pipeline that enables the simultaneous measurement of multiple inflammatory biomarkers from colourimetric sweat patch signals. Prior approaches have focused on single biomarkers, required specialized hardware, or lacked robustness to real-world imaging variability.

A key contribution of this thesis is the incorporation of an on-patch calibration layer, consisting of ArUco markers and reference colour swatches, which mitigates the variability introduced by different smartphones, lighting conditions, capture distances, and imaging environments. Following established practices in prior work, we operate in the CIELAB colour space and apply a colour correction matrix derived from the calibration layer, to normalize the captured images. To isolate the relevant signals in the SCRs, we introduce a pixel-level

scoring method that is specifically designed to identify the most informative colourimetric data.

The CIELAB colour space represents colour using three components, L^*, a^*, where L^* represents lightness, a^* represents the green to red axis, and b^* represents the blue to yellow axis. Unlike RGB, CIELAB is designed to be perceptually uniform, meaning that differences in the colour values correspond more closely to human visual perception. This makes CIELAB particularly useful for quantifying subtle colour changes and improving consistency across different devices and lighting conditions.

The effectiveness of the proposed pipeline is evaluated through several sets of experiments. Specifically, the system is assessed in terms of processing time, its ability to reduce variability under different lighting conditions, phones, capture distances, and reliability across participant data. Through quantitative results and statistical analysis, we demonstrate the practical feasibility of the proposed approach for POCT colorimetric screening of inflammatory biomarkers.

1.3 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2: Related Work provides a review of the existing literature on wearable sweat-based diagnostics and POCT. It discusses current electrochemical and optical biosensing technologies and reviews state-of-the-art methods in smartphone-based colourimetry, including RGB and CIELAB colour space approaches.

Chapter 3: Methodology details the proposed image processing pipeline. It describes the design of the calibration layer, the pre-processing and alignment techniques, and the core algorithms for image normalization, segmentation, and feature extraction. Finally, it outlines the methods for concentration prediction and the mobile application.

Chapter 4: Experimental Evaluation presents a comprehensive evaluation of the pipeline. This includes quantitative assessments of processing time, the effectiveness of image normalization, and the consistency across different smartphone devices and distances. The chapter concludes with a clinical evaluation comparing the pipeline's performance against ELISA

results using data from human participants.

Chapter 5: Conclusion summarizes the research findings, discusses the limitations of the proposed pipeline, and suggests directions for future work to further enhance the reliability and accessibility of this technology.

Chapter 2

Related Work

2.1 Wearable Sweat-Based POC Technologies

The concept of WMDs as non-invasive POCT using sweat as a key medium has previously been explored [19, 20]. Several types of wearable sweat-based POCT devices have been developed in recent years, including an increasing number of devices that make use of electrochemical and optical biosensors [19, 15, 21].

2.1.1 Electrochemical Biosensors

Electrochemical biosensors detect biochemical reactions and transduce them to measurable electrical currents [15, 22].

A recent development in the use of electrochemical biosensors for sweat-based WMDs that measure inflammatory biomarkers is the SWEATSENSER device [12]. The SWEATSENSER is a WMD that continuously monitors the inflammatory biomarkers IL-1beta and CRP in sweat and outputs measured concentrations of each. SWEATSENSER was designed to help monitor conditions such as inflammatory bowel disease [12]. The device works by using a replaceable strip and an electronic reader to convert the chemical reaction (the proteins binding to the markers on the strip) to a measurable electrical signal, which is then converted into concentrations (picograms per millilitre (pg/mL)) using a calibration curve [12]. While the SWEATSENSER device presents a promising avenue for continuous monitoring of IL-1beta and CRP in sweat, it requires each individual to remain in close proximity to an electronic reader or a paired smartphone in order to capture, process, and transmit the electrochemical signals. Furthermore, the reliance on electrical components, including using small electronics to read the signal, wireless communication modules, and a dedicated power source, increases

system complexity and can reduce cost-effectiveness because every user must depend on both having a smartphone and specialized hardware to obtain measurements.

2.1.2 Optical Biosensors

In comparison, optical biosensor methods do not require embedded electronics and instead encode biomarker concentrations as visible signal changes [15]. Most commonly, optical biosensors make use of LFIAs (such as our sweat patch) [15]. They are well known for their rapid POCT capabilities, low cost, and minimal equipment requirements, with common applications including pregnancy tests and COVID-19 diagnostics [15]. Sweat-based WMDs that use optical biosensors such as colourimetric assays have commonly been applied for the detection of pH, lactate, glucose, and chloride [23, 24, 25, 26]. For example, Bandodkar et al. developed a sweat patch for the simultaneous monitoring of sweat rate , pH, lactate, glucose, and chloride. To measure the concentration of chloride, they used the lightness channel (L^*) from the CIELAB ($L^*a^*b^*$) colour space along with a linear calibration curve [25]. They found that as the L^* channel decreases, the chloride concentration increases [25]. In our work, we also employ the CIELAB colour space and calibration curves. However, to our knowledge, no research has explored the use of colourimetric assays for detecting and obtaining reliable measurements of inflammatory biomarkers, including the simultaneous measurement of multiple biomarkers, from a sweat-based WMD.

2.2 Smartphone-based Colourimetry

The use of smartphones for colourimetric analysis has become increasingly common in biomedical and chemical applications. In this section, we focus specifically on their use for POCT of various biomarkers, reviewing the methods that are often used to capture and quantify colour changes, as well as the existing challenges for smartphone-based colourimetric measurements.

2.2.1 RGB-based Approaches

Celikbas et al. introduced a POCT paper-based LFIA for the detection of cancer using alpha-fetoprotein (AFP) and mucin-16 (MUC16) biomarkers [27]. They used an iPhone 7 to capture images of the test strips and processed the images using ImageJ software to extract colour information. They quantified the colour intensity based on the weighted contribution

of the RGB channels of the test area. This metric was then used to correlate the observed colour change with biomarker concentration in nanograms per milliliter (ng/ml). Celikbas et al. did not define exactly how colour intensity was computed, but they showed a linear correlation between colour intensity and the two biomarkers. However, their study had notable limitations. The researchers did not test the method under different lighting conditions. In addition, the analysis was performed using only a single device, an iPhone 7, which limits the generalizability of their approach to other smartphones with different camera sensors [27].

Chen et al. introduced the concept of a strongly correlated quantitative parameter (SCQP) for smartphone-based colourimetry. By deriving SCQPs and integrating them into a correction and machine learning pipeline, they mitigated variability across smartphone models, camera software and settings, ambient lighting, and skin tones in their specific case. As a demonstration, they showed that the method can noninvasively measure local oxygen saturation from lip or skin images. Their SCQP used a combination of RGB and HSV (Hue, Saturation, and Value/Brightness) colour spaces to extract complementary information from test strip images. Specifically, the RGB channels provided raw intensity values, while the HSV channels captured hue and saturation components. Their SCQP was the ratio of the green channel (G) from RGB to the value channel (V) from HSV:

$$\text{SCQP} = \frac{G}{V} \tag{2.1}$$

The SCQP parameter showed a stronger correlation with local oxygen concentration and greater robustness to variations in lighting and device type than using RGB or HSV channels alone [17].

In our work, we adopt a similar approach inspired by SCQP, but instead of RGB and HSV, we operate in the CIELAB colour space. This allows us to produce a strongly correlated metric that represents the dark-red hue of our colourimetric responses, while also reducing the variability that is introduced by different lighting conditions and devices.

2.2.2 CIELAB Colour Space

While Celikbas et al. and Chen et al. made use of RGB, other studies (including ours) deviate from the traditional RGB values due to the sensitivity of RGB to ambient lighting conditions, camera settings, and device variations [28]. A more popular approach for colour space in smartphone-based POCT is the CIELAB colour space.

Shen et al. provides a notable example of smartphone based colourimetric analysis for POCT. They captured images of commercially available colourimetric urine test strips using a smartphone camera and converted the mean RGB values from the test regions into the perceptually uniform CIELAB colour space for quantitative analysis [29]. Using CIELAB rather than raw RGB intensities improved measurement of pH and provided accuracy comparable to that obtained with a desktop scanner or silicon photodetectors [29]. Additionally, to improve calibration and compensate for environmental variability, the authors included a colour reference chart within each captured image, similar to the colour rendition charts introduced by McCamy et al., to standardise colour interpretation in imaging systems [30].

Shen et al.'s reference chart contained twelve colours, consisting of seven greyscale colours and five chromatic colours ranging from blue to red. The chromatic colours were chosen because these hues closely match the dominant colour changes in pH indicator dyes. The greyscale colours were used to correct for overall illumination shifts. By capturing both the test strip and the reference chart in the same image, the authors compared the measured RGB values of each colour with their known values to compute a colour correction matrix, which was applied to normalize the test strip region. This improved their measurements under varying lighting conditions. They also noted that printing the reference chart directly on the test strip could further improve this effect [29].

In our work, we adopt a similar approach to Shen et al. by using the CIELAB colour space and reference colours. However, for improved result, the reference colours are printed directly on the calibration layer of our sweat patch. Our calibration layer contains 49 reference colours, divided into two groups. The first group consists of 24 colours from McCamy et al., and the second group contains 25 shades of red ranging from very light to very dark. These red shades were chosen because they closely match the colour changes that occur in the SCR of our patch, similar to the approach of Shen et al. We similarly use a colour correction matrix applied to the SCRs to achieve consistent measurements under varying lighting conditions.

Chapter 3

Methodology

3.1 Pipeline Overview

This section will briefly introduce the main components of the proposed pipeline, which takes a smartphone-based image of the patch and outputs quantitative concentration readings of the three inflammatory biomarkers (IL-6, IL-1beta, CRP). This pipeline was designed to provide rapid results with minimal overhead and to overcome the known challenges of smartphone-based imaging, such as variations in lighting conditions, different camera sensors, angles, and distances. A custom mobile application was developed to integrate all stages of the pipeline into a single, user-friendly platform, from image capture to the final quantitative reading. The pipeline is made up of several key components, steps, and machine learning models. Below is a brief overview of the pipeline, consisting of five main steps:

1. Preprocessing and Alignment. This is the first computational component of the pipeline, which takes the raw image captured by the smartphone and transforms it into structured data. Here we detect the ArUco markers from the calibration layer to extract and standardize the patch to a fixed size and view.

2. Image Normalization. This step focuses on addressing the challenges associated with smartphone-based imaging. We first normalize the lighting of the extracted patch. Following this, we use the 49 colour references to normalize the colours of the patch by correcting them to their known references using a colour correction model.

3. Segmentation and Classification. Here, we extract and classify each SCR from the patch. For this we train a lightweight segmentation model to segment and extract the SCR. We then classify each region to its corresponding biomarker or control class.

4. Feature Extraction. Extract relevant pixel features from the segmented SCRs, based on their difference from the Control-Negative chamber's colour (the baseline).

5. Prediction and Output. We pass the extracted features of each of the three biomarkers to their respective calibration curves to retrieve a final prediction and result.

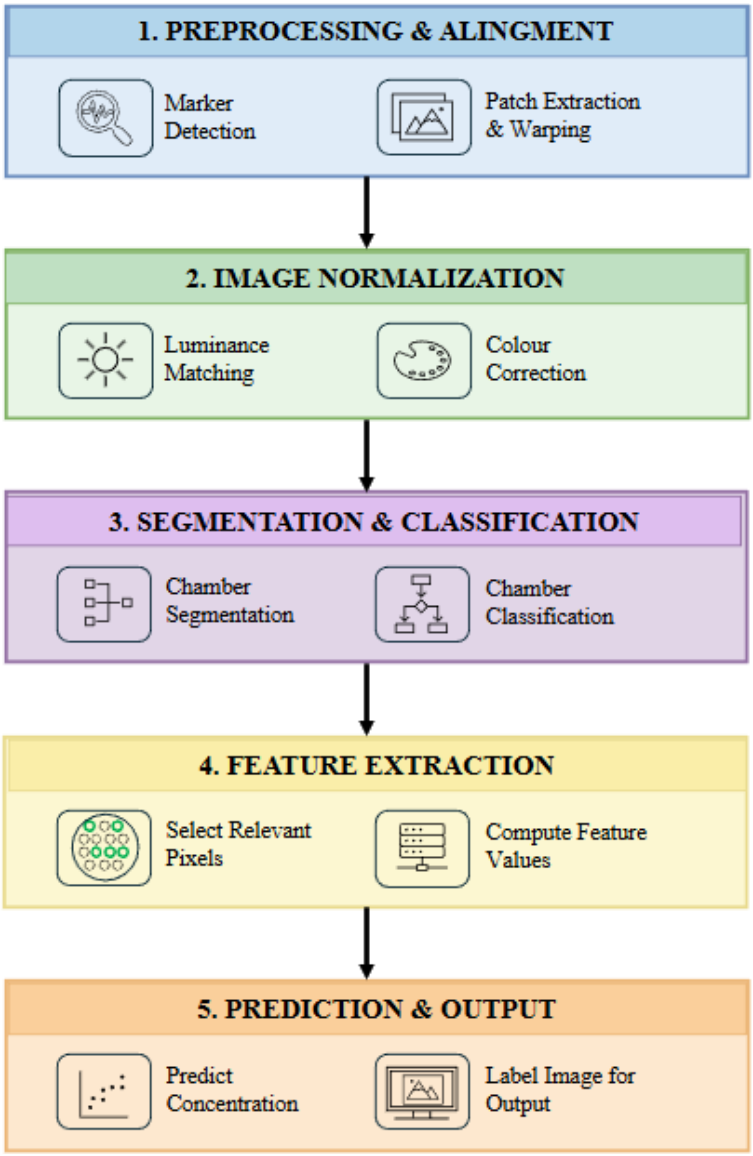


Figure 3.1: Overview of the pipeline consisting of five main components: (1) Preprocessing and alignment, (2) Image normalization, (3) Segmentation and classification of the SCRs, (4) Feature extraction, and (5) Prediction and output of the final concentration measurements.

The following sections will provide a detailed explanation of each component and step of the pipeline.

3.2 Calibration Layer

The calibration layer is designed to fit directly on top as the uppermost layer of the patch. Its main purposes are to provide reference points for image detection and to provide colour references for colour correction. The calibration layer features 49 colour references and five ArUco markers, with specific cutouts for each SCR. Notably, the calibration layer does not interfere with the functionality of the underlying layers or reduce the overall quality or effectiveness of the patch. Figure 3.2 shows the calibration layer fitted onto the sweat patch. The following subsections will detail and justify the design of each of the calibration layer's features.

Figure 3.2: The calibration layer fitted onto the sweat patch, illustrating the swatches, chamber cutouts, and ArUco markers as physical features.

3.2.1 Colour Swatches

Colour correction is a critical consideration in smartphone-based imaging systems, where variations in lighting conditions, camera sensors, angles, and distances can significantly affect the colours that are observed in the captured image. In the context of the inflammatory biomarker patch, these variations could introduce inaccuracies in the quantitative readings that are derived from the colourimetric reactions within the SCRs. To address these challenges, the calibration layer incorporates 49 colour references, which are referred to as colour swatches. These swatches represent the colours that are captured by the camera. They are later compared against their corresponding true reference values to correct for the variability introduced by the variations mentioned above. This method is inspired by the colour-rendition chart [30], which is commonly used in photography to correct observed colours to

their true values.

As shown in Figure 3.3 (Image retrieved from [31]), the colour-rendition chart is a standardized chart of 24 solid-coloured swatches that are designed to provide colour references for correcting imaging variations. The colours that are used in the standard colour-rendition chart utilize the colours proposed by McCamy et al. [30], which introduced the Macbeth colour chart as a standard reference for colour correction in photography [30]. The chart is organized into four rows, each containing six colours belonging to different categories. These categories can be named as natural colours (e.g., skin tones, foliage, sky), primary and secondary colours, grayscale colours, and miscellaneous colours. This arrangement provides a representative sample of colours that are commonly encountered in real-world imaging scenarios [30]. As such, the first 24 of the 49 colour swatches on our calibration layer correspond to the original 24 colours from the colour-rendition chart. These standard colours act as a universal reference for the camera. By including them, the colour correction process can account for different types of lighting, such as a warm, yellow light or a cool, fluorescent one, and correct for those distortions. This ensures that the final colour readings from the patch are accurate and consistent, regardless of the environment where the photo was taken.

The remaining 25 swatches were specifically chosen as a range of red shades, since the colourimetric reaction produced within the SCRs of the patch results in variations of red. Including these shades allows us to correct more precisely for different red variations, ensuring that subtle differences in intensity or hue can be reliably accounted for. Figure 3.4 illustrates both the standard colour-rendition chart colours in the top row and the additional 25 shades of red in the bottom row.

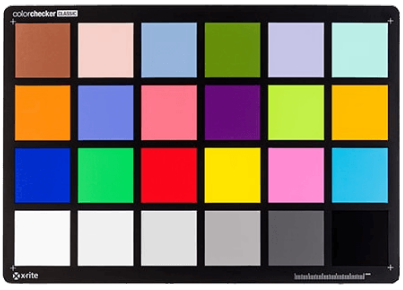


Figure 3.3: Colour-rendition chart, consisting of a set of 24 swatches. Such charts are commonly used in photography and computer vision to correct observed colours that are captured by the camera to match their true values more accurately, correcting for variations in imaging conditions.

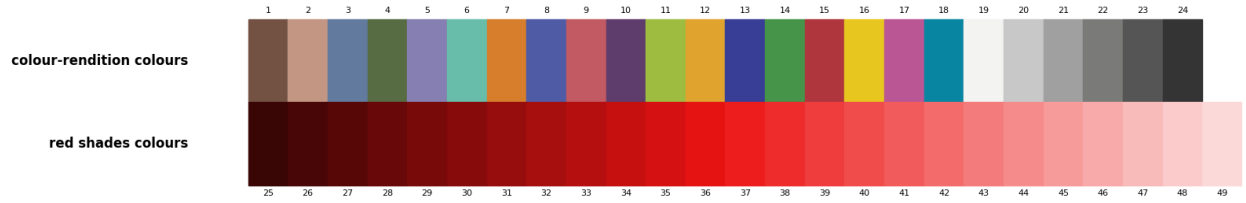


Figure 3.4: Figure of the 49 colour swatches on the calibration layer. The top row shows the 24 standard Macbeth colour-rendition colours, and the bottom row shows the 25 custom red shades specifically selected to calibrate for the colourimetric reaction of the patch.

Overall, the 49 colour swatches provide a robust colour reference that is both broad enough to correct for general imaging variations and specific enough to accurately correct for the colourimetric reaction of the patch.

3.2.2 ArUco Markers

In addition to addressing varying imaging conditions through colour swatches, the pipeline requires a method to reliably locate the patch and identify its internal structure. This involves two key steps. First, the patch must be accurately extracted from its background, as this region of interest is a critical precursor to segmenting the SCRs. Second, each segmented chamber must be correctly mapped to its corresponding biomarker, as the device features multiple SCRs .

To address these challenges, we utilized ArUco markers. An ArUco marker is a square fiducial marker with an inner binary matrix that serves as a machine-readable reference point for a camera [32]. As shown in Figure 3.2, five ArUco markers were placed around the inner part of the calibration layer, such that each SCR is positioned between two markers. By detecting these markers, the pipeline can precisely localize the device within the image and apply a perspective transform function to correct for any geometric distortions (Section 3.3 discusses this in further detail). This process standardizes the image and isolates the patch from the background, which is a crucial preparation step before the segmentation model can be applied.

After segmentation, the fixed positions of the ArUco markers relative to the SCRs provide a definitive coordinate system. This allows the pipeline to definitively identify which of the five SCRs corresponds to a particular biomarker or control. Specifically, the two markers between an SCR are used to compute its precise location, which in turn maps it to a biomarker label.

3.3 Preprocessing and Alignment

The pipeline begins with a smartphone-captured image as input. In this step, we address variations that are caused by differences in camera angle, zoom level, and background noise that arise from capturing images in uncontrolled environments. This is achieved by extracting the patch from the in and standardizing it to a fixed size with a flat, top-down view. This is important for the computational efficiency of our pipeline by standardizing to a fixed resolution, as well as for subsequent steps, such as Illumination Normalization (Section 3.4.1) and Chamber Segmentation (Section 3.5.1). Notably, the orientation of the input image is determined using the ArUco markers on the calibration layer, which provides clear positional references that indicate the orientation of the patch. As such, the orientation of the input image is not altered. The following section describes the patch extraction process in detail.

3.3.1 Patch Extraction

As introduced in Section 3.2, the calibration layer features five ArUco markers. These markers were generated using the OpenCV Python library [33]. OpenCV provides several predefined dictionaries for generating ArUco markers. In our case, the `DICT_4X4_50` dictionary was used. This dictionary defines markers based on a 4×4 grid of bits (16 bits in total) and contains 50 unique markers. From this dictionary, five distinct markers with IDs 10, 11, 12, 13, and 14 were generated for our calibration layer. The coordinates of each marker are estimated, stored in a dictionary, and then detected to provide positional information within the image. Formally, we define this dictionary as

$$D : \mathcal{I} \to \mathbb{R}^{4\times 2}, \tag{3.1}$$

where $\mathcal{I}$ denotes the set of marker IDs, and $\mathbb{R}^{4\times 2}$ represents the four corner coordinates of a marker in image space. Thus, for a marker with ID $i \in \mathcal{I}$, we have

$$D(i) = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \end{bmatrix}, \tag{3.2}$$

where (x_j, y_j) are the 2D pixel coordinates of the j-th corner of marker i.

Using the dictionary D, we can extract the patch region as follows. First, for each detected marker with ID $i \in \mathcal{I}$, we compute the marker center, c_i, by averaging its corner coordinates:

$$c_i = \frac{1}{4} \sum_{j=1}^{4} (x_j, y_j), \tag{3.3}$$

Next, we compute the patch center C as the average of all marker centers:

$$C = \frac{1}{5} \sum_{i=1}^{5} c_i, \tag{3.4}$$

Finally, we define the patch radius r as the maximum Euclidean distance between the patch centroid C and any marker center c_i, scaled by a factor α:

$$r = \alpha \max_i \|c_i - C\|_2. \tag{3.5}$$

Here, α is a scalar expansion parameter that compensates for the fact that the detected marker centers lie inside the true patch boundary. Without this adjustment, the extracted region would end at the markers and parts of the calibration layer, such as the colour swatches, would be inadvertently excluded. We found that setting $\alpha = 1.6$ provides an accurate estimation of the patch radius, reliably encompassing the entire patch including both the SCRs and the colour swatches.

After computing the patch center and radius, we apply a perspective transform. The perspective transform projects the patch so that its center aligns with the center of the output image. Specifically, we project it to a fixed size of 2000×2000 pixels that results in an image with a flat, top-down view. This projection standardizes our input and removes any variations in angles or zoom. Additionally, because the projection boundaries are defined by the patch radius, the operation effectively acts as a crop, isolating the patch region and removing most of the irrelevant background noise.

We choose a resolution of 2000×2000 pixels based on the typical camera sensors that are found in modern smartphones. Most contemporary devices utilize a 12-megapixel sensor or higher, which generally produces input resolutions of approximately 4000×3000 pixels or more. We assume that the input image resolution is at least 4000×3000 pixels. Since we retain only the patch region from the image. As a result, not all 4000×3000 pixels are required. Therefore, we found that a resolution of 2000×2000 pixels provides a good balance between preserving crucial image details and avoiding the need to maintain the entire original resolution.

To perform the perspective transformation, we define four source corner points that are bound the patch region in the original image as well as four destination corner points that correspond to the corners of the output image. The source points are defined relative to the patch center C and the estimated radius r. Specifically, the four corners of the patch region are calculated as

$$\texttt{src\_pts} = \begin{bmatrix} C_x - r & C_y - r \\ C_x + r & C_y - r \\ C_x - r & C_y + r \\ C_x + r & C_y + r \end{bmatrix}. \tag{3.6}$$

The destination points corresponding to the corners of the output image of size 2000×2000 pixels are calculated as

$$\texttt{dst\_pts} = \begin{bmatrix} 0 & 0 \\ 2000 - 1 & 0 \\ 0 & 2000 - 1 \\ 2000 - 1 & 2000 - 1 \end{bmatrix}. \tag{3.7}$$

where index $i \in \{0, 1, 2, 3\}$ of both `src_pts` and `dst_pts` corresponds to the respective corner:

$$0 = \text{top-left}, \quad 1 = \text{top-right}, \quad 2 = \text{bottom-left}, \quad 3 = \text{bottom-right}. \tag{3.8}$$

Using these two point arrays, the perspective transform matrix M is estimated with OpenCV's $cv2.getPerspectiveTransform$ [33] function. This function automatically calculates the unique homography matrix that maps the four source points to their corresponding destination points. The transformation is applied to the patch using OpenCV's $cv2.warpPerspective$ function, yielding a perspective-corrected patch image with a fixed size of 2000×2000 pixels.

While the initial perspective transform already removed a substantial amount of background noise and centered the patch, a circular masking technique is applied to remove the residual, non-patch-related background that is remaining along the edges of the normalized 2000×2000 image (such as skin, fabric, or other background that was not fully cropped by the prespective transformation). This step is essential because the segmentation model, which is specifically designed to segment and isolate the SCRs, was trained on images of the patch without any background. Given that all images have been normalized to a size of 2000×2000 pixels, we can generate a simple binary mask. A binary mask is an image where each pixel has a value of either 0 or 255 (black or white respectively), which corresponds to whether the pixel is ignored or kept. Specifically, we generate a circular mask of the same size as the output

image, centered at $(1000, 1000)$ with a radius of r. By creating a white circle on a black background and then performing a bitwise AND operation with the perspective-corrected patch image, we can isolate the patch area and effectively remove the remaining surrounding background.

Figure 3.5 illustrates the extraction process. The raw smartphone image is first processed to detect ArUco markers, which define the patch region. The patch is then normalized through perspective correction, and finally, a circular mask is applied to remove any unnecessary background. The resulting standardized patch serves as the input for subsequent stages in the pipeline.

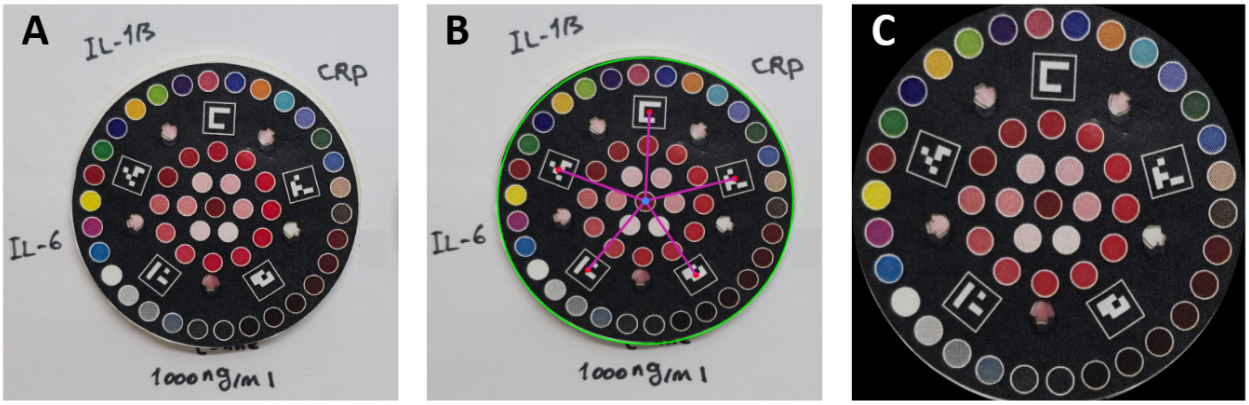


Figure 3.5: The extraction process. **(A)** Original patch image captured by a smartphone (in this case by Samsung Galaxy S24 Ultra). **(B)** **Red** points indicate ArUco marker centers. The **blue** point is the computed patch center, the **purple** lines represent the initial computed radius, and the **green** circle shows the final extraction radius after adjustment by a 1.6 ratio to encompass the entire patch. **(C)** Final warped and standardized image with a flat, top-down angle/view of the extracted patch with a size of 2000 × 2000 pixels.

3.4 Image Normalization

The goal of this stage is to ensure that all input images share a consistent visual domain, minimizing variations that could negatively affect analysis. A consistent visual domain represents a normalized state of feature consistency, where the influence of external imaging variables such as differences in lighting, colour temperature, and camera sensor characteristics is minimized. For instance, if the same colourimetric reaction is captured under two distinct lighting conditions, the observed pixel values will be inherently different. Our objective is to minimize this difference so that the same underlying features are represented consistently across all images, regardless of the conditions. To achieve this, we first normalize the illumination across the extracted patches. We then use the 49 colour references on the calibration layer

to correct and normalize colour values based on their known ground-truth references. In this context, ground-truth values correspond to the true RGB values of each of the 49 colours. The following sections provide a detailed description of this two-step normalization process.

3.4.1 Illumination Normalization

Image illumination is a well-recognized problem in image analysis, and many studies have shown that image analysis is heavily affected by illumination [34, 35, 36]. To address this issue, a variety of illumination normalization methods have been developed, which can generally be classified into global and local techniques. Global techniques, such as histogram equalization and gamma correction, operate uniformly on all pixels of an image, adjusting overall brightness or contrast based on global statistics. In contrast, local techniques focus on small regions within the image, allowing for spatially adaptive correction of illumination variations. An example of a local technique is Contrast Limited Adaptive Histogram Equalization (CLAHE), which performs histogram equalization on localized tiles and then merges the results using interpolation [37]. While these approaches can partially mitigate lighting variations, they are limited for our purposes because they fail to normalize different input images to a single, consistent visual domain and instead, normalize each image independently. Therefore, a technique like CLAHE produces a standardized distribution for each individual image but does not ensure that the final illumination distribution of two separate input images will match.

The goal of our normalization pipeline is not merely to enhance the appearance of individual images, but to ensure that all images share a consistent visual domain, thereby minimizing feature variance across different imaging conditions for the same input. To achieve an illumination distribution that is consistent across all images, we normalize the lighting of every input to a standard, digitally generated reference that is inherently free from the influence of environmental lighting. The specific reference image used for this is shown in Figure 3.6.

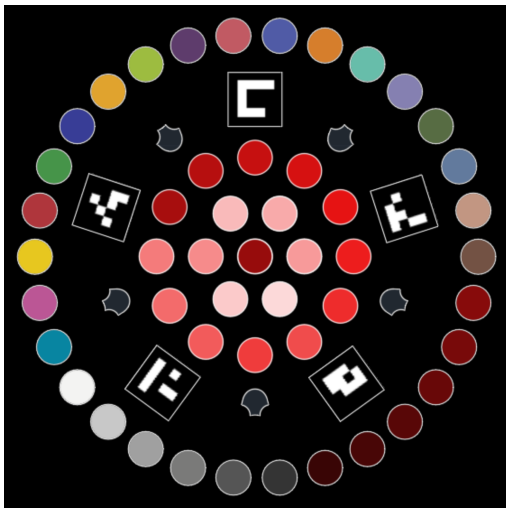


Figure 3.6: A digitally created reference image used for illumination normalization. This image serves as the standard target for luminance histogram matching, ensuring all input images are normalized to a consistent domain.

To ensure the illumination correction works as intended, we first match the orientation of the reference image to the orientation of the input. The rotation angle is computed using the ArUco markers that are embedded on each patch, which provides a reliable geometric reference for determining patch orientation (the detailed computation of marker angles is described in Section 3.5.2). The reference is then rotated around its center using an affine transformation, which preserves image size and interpolates pixel values to avoid artifacts. In our implementation, this is performed using OpenCV's $getRotationMatrix2D$ and $warpAffine$ functions [33]. Once aligned, histogram matching is performed to transform the input's luminance distribution to that of the rotated reference, producing a normalized patch that is consistent in illumination.

This process is conducted in the CIELAB ($L^*a^*b^*$) colour space, which is designed to be perceptually uniform. This means that a given change in a colour's numerical value in this space corresponds to a similar change as perceived by the human eye. Crucially for our method, CIELAB separates the luminance (or lightness) channel, denoted as L^*, from the two chromaticity channels, a^* and b^*. By applying the histogram matching algorithm exclusively to the L^* channel, we can effectively match the luminance channel while leaving the chromaticity information in the a* and b* channels unchanged. This ensures that only lighting is normalized, leaving the colourimetric data of the SCRs intact for subsequent analysis. Algorithm 1 provides a high level step-by-step description of this process.

Algorithm 1 Luminance Histogram Matching

procedure MATCHLUMINANCEHISTOGRAM(Patch, Reference) ▷ Normalize the luminance of Patch to match a Reference image.
    ▷ Rotate Reference to match Patch.
    $ReferenceRotated \leftarrow \text{ROTATE}(Reference)$
    ▷ Convert both images to CIELAB to separate luminance (L channel).
    $PatchLab \leftarrow \text{CONVERTCOLOUR}(Patch, RGB \rightarrow LAB)$
    $ReferenceLab \leftarrow \text{CONVERTCOLOUR}(ReferenceRotated, RGB \rightarrow LAB)$
    ▷ Split channels to isolate luminance.
    $L_{patch}, a_{patch}, b_{patch} \leftarrow \text{SPLITCHANNELS}(PatchLab)$
    $L_{ref}, \_, \_ \leftarrow \text{SPLITCHANNELS}(ReferenceLab)$
    ▷ Match the patch's luminance histogram to the reference.
    $L_{matched} \leftarrow \text{MATCHHISTOGRAMS}(L_{patch}, L_{ref})$
    ▷ Merge matched luminance with original colour channels.
    $PatchLabMatched \leftarrow \text{MERGECHANNELS}(L_{matched}, a_{patch}, b_{patch})$
    ▷ Convert back to RGB.
    $PatchMatched \leftarrow \text{CONVERTCOLOUR}(PatchLabMatched, LAB \rightarrow RGB)$
    return $PatchMatched$
end procedure

3.4.2 Colour Normalization

Following illumination normalization, we address the challenge of colour variations. The concentration predictions of the three biomarkers from our patch are derived directly from the colourimetric reactions in the SCRs. For example, the same sample taken under different lighting conditions can appear warmer or cooler in hue. To account for such variations, we normalize the pixel colours of the SCRs. This ensures that the colourimetric responses are standardized across inputs, reducing variability and improving feature consistency.

To implement this colour normalization, we apply a colour correction strategy. We first extract the 49 colour swatches on the calibration layer. These swatches provide our "observed" colours, which are the raw pixel values captured by the smartphone's camera. We then match each observed colour to its respective reference colour (known true value). With this, we are able to fit a Linear Regression model. This model learns the best transformation matrix that is required to map the observed colours to their true reference colours. We can then apply this learned transformation model to the pixel values of the SCRs. This normalizes the SCR

pixel values to a consistent visual domain that better matches the known reference colour values. The data fed to the downstream analysis is then more consistent and normalized. The following will detail this process.

3.4.2.1 Swatch Extraction

Before any colour correction can be performed, we must first extract the 49 colour swatches from the patch. Unlike a standard colour-rendition chart where colours are arranged in a fixed grid, our colour swatches are distributed across the calibration layer as shown in Figure 3.2. This means that the location of each swatch is unknown in the captured input and requires a more complex extraction. Our extraction method must not only locate each swatch, but also correctly match each observed swatch colour to its known reference colour value before we can apply any colour correction methods.

To extract the 49 colour swatches from the patch, we utilize blob detection, a technique provided by OpenCV [33] as the computer vision technique for identifying blobs. A blob is a contiguous group of pixels in an image that are similar in properties such as colour, intensity, or texture. In other words, blobs are effectively regions of interest that stand out from their surroundings. Since each swatch is a group of pixels that form a circular shape with a distinct colour, we can see that each swatch fits the description of a blob. Figure 3.7 illustrates the intermediate result of applying blob detection.

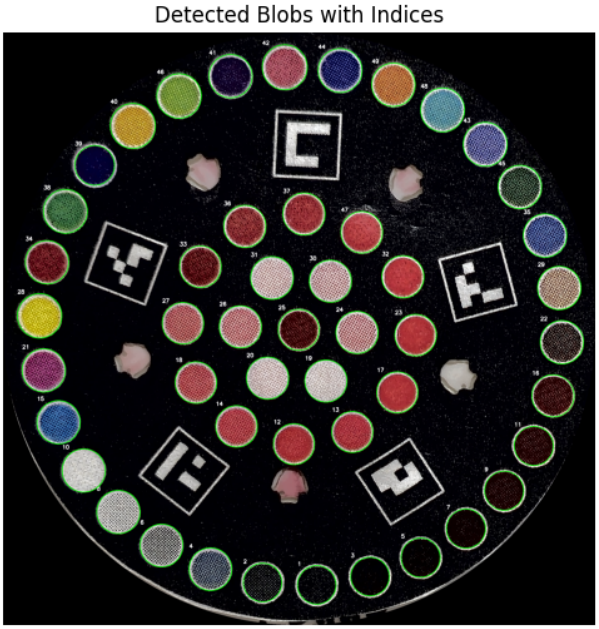


Figure 3.7: Result of Blob Detection on the Calibration Layer. The green circles indicate the regions that are successfully identified as colour swatches (blobs).

Once each blob is detected, we extract its observed colour by averaging its pixel (RGB) values. This yields a set of 49 colours that are observed from the initial input.

3.4.2.2 Swatch Matching

After extraction, we need to correctly match each of the 49 observed colours to its corresponding reference colour. Since the order of the extracted swatches is arbitrary and not arranged in a fixed grid, a simple one-to-one assignment based on position is not possible. This necessitates a more robust matching algorithm.

We formulate this problem as a Linear Sum Assignment Problem (LSAP). LSAP aims to find an optimal one-to-one pairing between N observed items and N reference items such that the total cost of all pairings is minimized. In our case, $N = 49$. To do so, we must first define a cost between a pair consisting of an observed colour and a reference colour.

Let the two sets of colours be represented as

$$S_O = \{o_1, o_2, \ldots, o_{49}\}, \quad S_R = \{r_1, r_2, \ldots, r_{49}\}.$$

Where S_O is the set of observed colours from our input, and S_R is the set of reference colours. We then define a cost matrix $C \in \mathbb{R}^{49\times 49}$, where each element C_{ij} represents the distance (or dissimilarity) between the observed colour o_i and the reference colour r_j. The goal of LSAP is to find an optimal one-to-one assignment between the 49 observed colours in S_O and the 49 reference colours in S_R such that the total sum of all pairwise distances is minimized (cost matrix C is minimized).

Mathematically, this can be expressed as follows. Let $i \in \{1, 2, \ldots, 49\}$ index the observed colours in S_O. A permutation π of these indices represents an assignment of each observed colour o_i to a reference colour $r_{\pi(i)}$. The total cost of a given assignment is then

$$\sum_{i=1}^{49} C_{i,\pi(i)}, \tag{3.9}$$

Where $C_{i,\pi(i)}$ is the cost of pairing o_i to a reference colour $r_{\pi(i)}$, and the goal of LSAP is to find the permutation π that minimizes this total cost.

We define the cost between two colours as their distance in the CIELAB ($L^*a^*b^*$) colour space. CIELAB is used because it provides a perceptually uniform representation, meaning that the Euclidean distance between two points in this space corresponds more closely to the human perception of colour difference. Specifically, we compute the CIEDE2000 colour difference standard, denoted as ΔE_{00}, between each pair of observed and reference colours. CIEDE2000 is the current CIELAB colour difference formula that was developed by the International Commission on Illumination (CIE):

$$C_{ij} = \Delta E_{00}(o_i, r_j), \tag{3.10}$$

The complete cost matrix C is then passed to the Hungarian algorithm (also known as the Kuhn–Munkres algorithm) [38], which efficiently solves the LSAP in polynomial time and finds the optimal one-to-one mapping between the observed and reference colours.

In some instances, a matched pair of observed and reference colours may still exhibit a very large perceptual difference (ΔE_{00}). This can be caused by external factors like poor image quality. These pairs can be an issue for the colour correction model because they act as outliers, which introduces a greater bias into the training process. If left unaddressed, the model may over-correct for these specific differences, which compromises its overall performance. To avoid over-correcting for such cases, any matches with a perceptual distance greater than 25 units ($\Delta E_{00} > 25$) are considered unreliable and are directly replaced with their corre-

sponding reference colours. $\Delta E_{00} > 25$ threshold was selected based on empirical testing. By doing this, we minimize the impact of such cases and maintain a more balanced correction.

As a result, we get an optimal mapping between each observed colour and its corresponding reference colour. Figure 3.8 illustrates the matched swatches for the input shown in Figure 3.7, with each reference colour displayed alongside its corresponding observed colour.

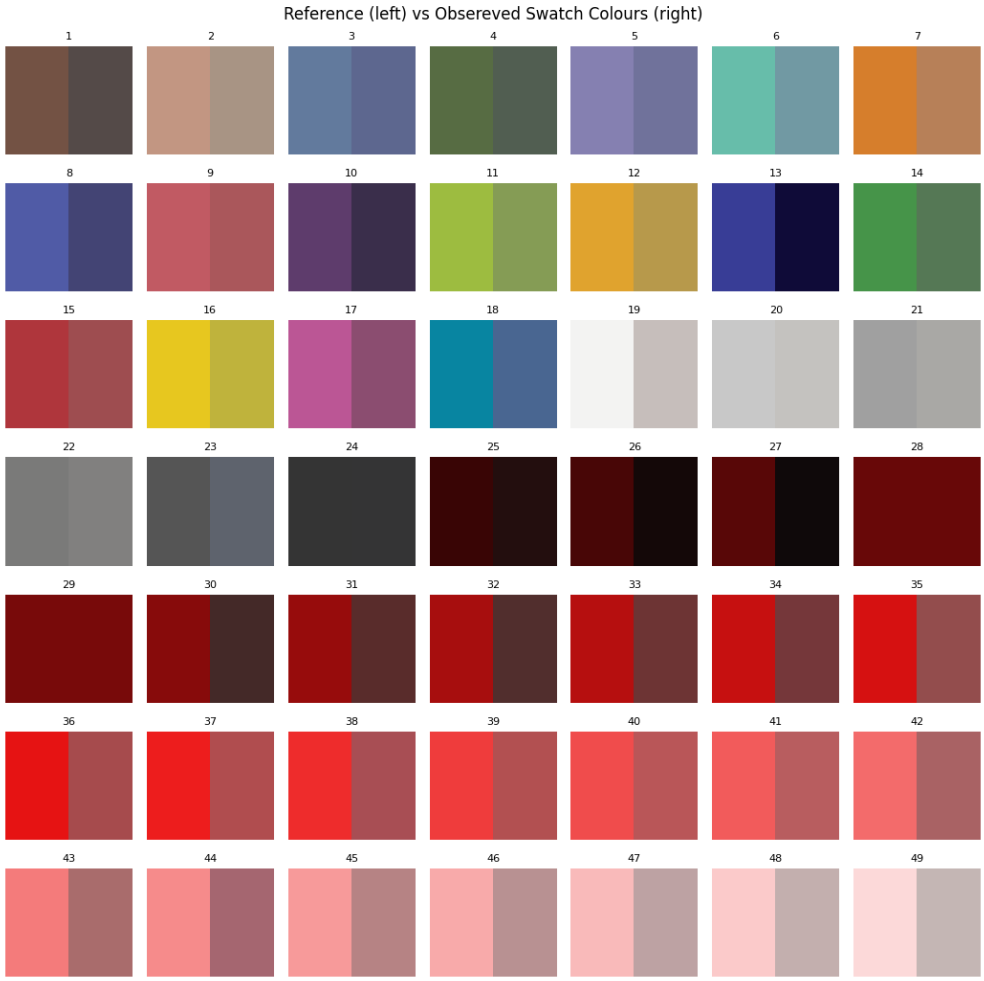


Figure 3.8: Visualization of swatch matching results. Each reference colour (left) is paired with its corresponding observed colour (right) as determined by our matching algorithm. Swatches 28 and 29 were replaced with the known reference colours due to their ΔE_{00} exceeding 25.

3.4.2.3 Colour Correction Model

Now that we have extracted the observed RGB values for each of the 49 colour swatches and established their correspondence with reference values, we can fit a regression model to learn a colour correction transformation. This transformation maps observed colours to their true

reference counterparts.

Formally, we defined $X \in \mathbb{R}^{49\times3}$ as the matrix of observed colours, and $Y \in \mathbb{R}^{49\times3}$ as the matrix of corresponding reference colours. Each row represents a swatch, and the three columns correspond to its RGB values. The previously computed mapping ensures that the i-th rows $x_i \in X$ and $y_i \in Y$ form matched pairs of observed and reference colours.

Our goal is to learn a transformation matrix M such that

$$Y \approx XM, \tag{3.11}$$

where $M \in \mathbb{R}^{3\times3}$ maps a 3-dimensional input colour space (R, G, B) to a 3-dimensional reference colour space. This can be achieved by the ordinary least-squares solution, and is obtained as

$$M = (X^\top X)^{-1} X^\top Y. \tag{3.12}$$

Once M is learned, it is applied to correct the colours of the SCRs after they are segmented (further discussed in Section 3.6). This approach improves the computational efficiency of the pipeline because because the correction is applied only to the SCR regions instead of the entire patch image. This normalization step ensures that all SCRs are represented within a consistent visual domain.

3.5 Chamber Segmentation and Classification

This stage of the pipeline focuses on isolating and identifying each SCR through a two-stage process. First, a lightweight segmentation model is used to segment each SCR. Next, each extracted region is classified into its corresponding biomarker or control class.

3.5.1 Chamber Segmentation

To segment the SCRs from the patch image, we designed a lightweight U-Net variant. U-Net is a convolutional neural network (CNN) architecture that was originally introduced for biomedical image segmentation [39]. It features a "U"-shaped structure in which the encoder progressively downsamples the input image to extract features at multiple scales, while the decoder upsamples these features to generate the final segmentation mask. Skip connections between corresponding encoder and decoder layers allow the network to retain high-resolution spatial information. The segmentation model was intentionally designed to be lightweight in

order to enable rapid processing.

3.5.1.1 Model Architecture

Our lightweight U-Net variant incorporates a pretrained MobileNetV3-Large [40] encoder from the timm library [41] to extract multi-scale feature maps. MobileNetV3-Large is efficient and lightweight model, providing high-quality features while keeping computation low. We use a version of the pretrained on ImageNet, a large dataset of natural images widely used for training vision models [42]. Specifically, we extract multi-scale feature maps at four spatial resolutions

Let the input image be $X \in \mathbb{R}^{H \times W \times 3}$, where H and W represent the image height and width in pixels, and 3 denotes the colour channels. The encoder produces feature maps

$$\{E_1, E_2, E_3, E_4\} = \text{Encoder}(X), \tag{3.13}$$

where $E_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, with spatial dimensions (H, W) decreasing and channel features (C) increasing with depth i.

The decoder reconstructs the segmentation mask by progressively upsampling the encoded features and concatenating them with the corresponding encoder features:

$$D_i = \text{DoubleConv}(\text{Upsample}(D_{i+1}) \oplus E_i), \tag{3.14}$$

where $\oplus$ denotes channel-wise concatenation and *DoubleConv* consists of two convolutional layers with batch normalization, ReLU activations, and a residual connection. The first decoder stage is given by

$$D_4 = \text{DoubleConv}(\text{Upsample}(E_4) \oplus E_3), \tag{3.15}$$

and the final decoder stage concatenates the upsampled features with the original input image:

$$D_1 = \text{DoubleConv}(\text{Upsample}(D_2) \oplus X). \tag{3.16}$$

The output mask is produced by a 1×1 convolution followed by a sigmoid activation:

$$Y_{\text{pred}} = \sigma(\text{Conv}_{1\times1}(D_1)), \quad Y_{\text{pred}} \in [0,1]^{H \times W}. \tag{3.17}$$

The final output is a binary mask in which each of the five SCRs is individually segmented.

Figure 3.9 illustrates example segmentation results, showing close alignment between the predicted masks and ground-truth annotations.

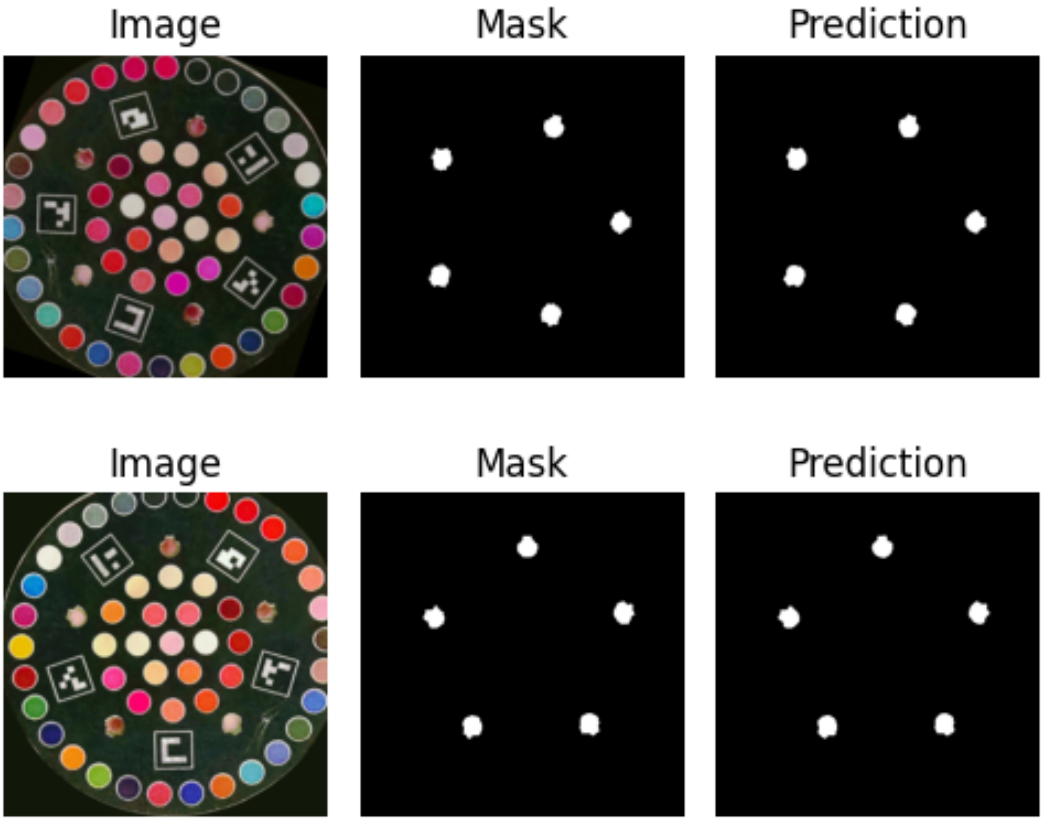


Figure 3.9: Example of two segmentation results that are produced by the proposed model. Each row shows the input patch image, the ground-truth mask, and the corresponding model prediction.

3.5.1.2 Training Details

The segmentation model was trained using the PyTorch framework [43] on a single NVIDIA P100 GPU. Input images and their corresponding ground-truth masks were resized to 320 × 320 pixels for inference and then the predictions were resized to 2000 × 2000. Importantly, the model was trained on patch images that had a uniform black background to help it learn features of the SCRs, without introducing background bias from the training data. This highlights the importance of the earlier preprocessing step (Section 3.3.1) that extracts patches from the raw images and removes background noise. Data augmentation included random rotations, flips, colour jitter, Gaussian blur, and random cropping for model generalization. All input images were normalized using the ImageNet mean and standard deviation [42]. Further results, implementation, data curation, and hyperparameter details are provided in Appendix B.

We employed a 5-fold cross-validation strategy with a batch size of 16 and trained each fold for 50 epochs using the Adam optimizer with a learning rate of 1×10^{-3}. The model was trained using a hybrid loss function that combines Dice loss [44] and binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{hybrid}} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} \tag{3.18}$$

where $\lambda_{\text{dice}} = \lambda_{\text{bce}} = 0.5$. The Dice loss is defined as:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2|P \cap G| + \epsilon}{|P| + |G| + \epsilon} \tag{3.19}$$

where P and G denote the predicted and ground truth segmentation masks, respectively, and ϵ is a small constant for numerical stability.

Model performance was evaluated using the Dice coefficient [44] that we computed on the validation folds, and the model that achieved the highest Dice score across folds was selected for subsequent use (Appendix B).

3.5.2 Chamber Classification

Following segmentation, each of the five extracted SCRs is classified to determine its identity. Each SCR is assigned to one of five classes, including three biomarkers (IL-6, IL-1beta, and CRP) and two controls (Control-Negative and Control-Positive). This classification step is crucial for three main reasons. First, it identifies the Control-Negative chamber, which serves as the baseline for all feature calculations. Second, it ensures that the features that are extracted from each SCR are directed to the correct downstream analysis for accurate concentration prediction. Third, it provides the necessary labels to generate interpretable results that can be clearly presented to the user.

Fortunately, we have a definitive method for classifying each SCR into its respective class. Recall that we have placed five ArUco markers on our calibration layer such that each SCR lies between two adjacent ArUco markers. Each ArUco marker has a unique ID. Specifically, in our case, we have five markers with IDs $[10, 11, 12, 13, 14]$. With this, we can define a relationship between each pair of adjacent marker IDs (i, j) and their corresponding biomarker label based on the physical layout of the patch:

$$
\begin{aligned}
&(10, 11) \rightarrow \text{CRP},\\
&(11, 12) \rightarrow \text{Control-Negative},\\
&(12, 13) \rightarrow \text{Control-Positive},\\
&(13, 14) \rightarrow \text{IL-6},\\
&(14, 10) \rightarrow \text{IL-1beta}.
\end{aligned}
\tag{3.20}
$$

To determine which pair a SCR belongs to, we compute its angle and assign it to the pair of adjacent ArUco markers whose angles bound the SCR angle. For example, suppose the computed angles of the ArUco markers are as follows:

Marker 10: 0°, Marker 11: 72°, Marker 12: 144°, Marker 13: 216°, Marker 14: 288°.

If a SCR has a computed angle of 75°, it lies between markers 11 and 12, and according to the mapping above, it is assigned to the Control-Negative class. Similarly, a SCR with an angle of 230° lies between markers 13 and 14 and is assigned to the IL-6 class.

To start, we first compute the angular position of each ArUco marker relative to the patch center. Let (C_x, C_y) denote the coordinates of the patch center, and let (x_i, y_i) denote the centroid of marker i (these are computed similarly as done in Section 3.3.1). The angle of the marker relative to the patch center is computed using the two-argument arctangent function:

$$
\theta_i = \operatorname{atan2}(y_i - C_y,\, x_i - C_x), \tag{3.21}
$$

which returns the signed angle in radians. These angles are then converted to degrees and normalized to the range $[0, 360)$:

$$
\theta_i = \left(\frac{180}{\pi} \cdot \operatorname{atan2}(y_i - C_y,\, x_i - C_x) + 360 \right) \bmod 360. \tag{3.22}
$$

The same process is applied to each SCR centroid (x_r, y_r) to obtain its angular coordinate θ_r. A SCR is classified as belonging to the biomarker interval (i, j) if its angle θ_r lies within the angular span between θ_i and θ_j.

This geometric method ensures that each SCR is consistently and deterministically assigned to its correct biomarker class, independent of the image orientation or capture conditions. In other words, even if the patch is rotated or captured from a different angle, each SCR is

deterministically assigned to the correct biomarker or control class. Figure 3.10 illustrates this geometric mapping, plotting the calculated angles for each SCR and ArUco marker relative to the patch center.

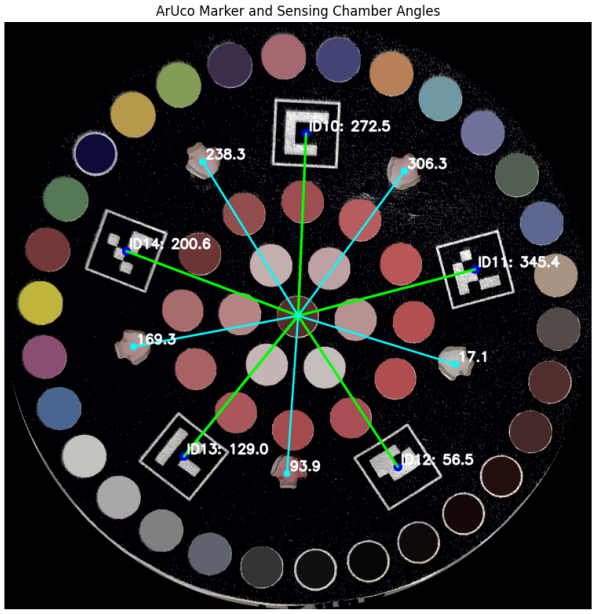


Figure 3.10: Visualization of the angular classification method for SCRs. The angle for each ArUco marker (green lines, IDs 10-14) and each SCR (cyan lines) is computed relative to the patch's center using the *atan*2 function. Each chamber is then deterministically classified based on the angular interval defined by its two adjacent ArUco markers. For instance, the SCR at 238.3° will be identified as IL-1beta because its angle falls between that of marker ID14 (200.6°) and marker ID10 (272.5°).

3.6 Feature Extraction

Following the segmentation and classification stages, each SCR has been segmented and assigned a corresponding biomarker label (IL-6, IL-1beta, or CRP) or control class. This section describes the feature extraction process. First, a preliminary step is taken due to a limitation in the manufacturing of the device. The calibration layer was laser cut to create the SCR cutouts, which often resulted in burn lines on the edges of the SCRs. To mitigate the impact of these burn lines, which could significantly affect the results, the segmentation area of each SCR was shrunk by 20%. After this adjustment, we then apply our colour correction model from Section 3.4.2 to normalize the pixels of each SCR. Finally, we extract relevant features based on the colourimetric signal in each of the three biomarker's SCR relative to the Control-Negative SCR.

3.6.1 Applying Colour Normalization

Before feature extraction, colour consistency across each SCR is ensured using the colour correction model introduced in Section 3.4.2. The learned transformation matrix M is applied to each SCR to perform colour normalization. This step ensures that the extracted features are normalized to a consistent visual domain, helping minimize variations caused by differences in camera sensors, or imaging conditions.

For a given SCR, let $\mathbf{X}_{\text{SCR}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ denote the set of all n pixel colour vectors within its region. Each pixel $\mathbf{x}_i$ is represented directly by its RGB components:

$$\mathbf{x}_i = [R_i, G_i, B_i] \in \mathbb{R}^{1\times 3} \tag{3.23}$$

The learned transformation matrix from Section 3.4.2, $M \in \mathbb{R}^{3\times 3}$, is then applied to map each pixel vector into the corrected colour domain:

$$\mathbf{x}'_i = \mathbf{x}_i M, \quad \text{where } \mathbf{x}'_i \in \mathbb{R}^{1\times 3}. \tag{3.24}$$

The colour-corrected SCR is then represented by the matrix $\mathbf{X}'_{\text{SCR}} \in \mathbb{R}^{n\times 3}$:

$$\mathbf{X}'_{\text{SCR}} = \mathbf{X}_{\text{SCR}} M, \tag{3.25}$$

where $\mathbf{X}_{\text{SCR}} \in \mathbb{R}^{n\times 3}$ denotes the matrix containing the raw RGB values of all pixels within that region. The set of corrected pixel vectors is the matrix $\mathbf{X}'_{\text{SCR}} = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_n\}$. This ensures that each pixel colour of each SCR is projected into a normalized, consistent visual domain.

3.6.2 Extracting Features

The concentration of each biomarker is directly correlated with the colourimetric reaction produced by its respective SCR. This reaction generates varying shades of red, where the concentration of the target biomarker is directly proportional to the intensity and hue of the red colour. For example, an SCR exhibiting a darker red shade indicates a relatively higher concentration of its corresponding biomarker. Therefore, the feature extraction process must effectively capture this red signal.

In order to effectively capture the relevant signal, we assign a score to each pixel. This allows us to encode, within each pixel of the SCR, how strongly it relates to the underlying concentration. Specifically, this score reflects both the redness and darkness of each pixel. To

achieve this, we make use of the CIELAB colour space, which provides two channels that are particularly relevant: the a^* and L^* channels. The a^* channel represents the red–green axis and directly encodes the degree of redness, while the L^* channel captures lightness, indicating how bright or dark a pixel appears.

To quantify how dark-red a pixel is, we combine information from the a^* and L^* channels into a pixel scoring function, as both are highly correlated with the colour characteristics we aim to measure.

We first introduce a darkness factor based on the L^* channel, defined as

$$1 - \frac{L^*}{100}. \tag{3.26}$$

The L^* channel ranges from $[0, 100]$, where higher values correspond to brighter pixels. Dividing by 100 normalizes it to $[0, 1]$, and subtracting from 1 inverts the scale so that darker pixels (lower luminance) receive higher values. This ensures that darker pixels contribute more strongly to the final score. Since L^* acts primarily as a scalar, the redness encoded by the a^* channel plays the dominant role in determining the final pixel score. The redness of a pixel is crucial, as it reflects the biomarker signal resulting from the reaction with the AuNPs in the SCR. To emphasize this contribution, we raise the a^* value to the power of 3 (an exponent of 2 was avoided to prevent sign cancellation). The final pixel score is therefore computed as

$$\text{score} = (a^*)^3 \left(1 - \frac{L^*}{100}\right). \tag{3.27}$$

This approach ensures that pixels which are strongly red-hued and dark are assigned higher scores.

With our pixel scoring function in place we can start our feature extraction process from the SCRs. We first establish a baseline using the Control-Negative SCR. The Control-Negative does not produce any reaction. If it does produce a reaction, this indicates a faulty sample and invalid results. Additionally, the Control-Negative SCR is captured under the same conditions as the other three biomarker SCRs within the patch. This allows us to confidently label the Control-Negative SCR as representing a concentration of zero for any given sample.

Using the Control-Negative as a baseline is critical because the raw pixel values of a SCR are sensitive to capture conditions such as lighting and camera settings. For example, consider

a SCR measured under warm lighting versus cool lighting. Even if the underlying biomarker concentration is identical, the pixel values of the SCR may differ significantly, potentially leading to misleading feature values. While our colour normalization helps mitigate this issue, it cannot completely remove the effect of different capture conditions. By comparing each biomarker SCR to the Control-Negative score, we extract relative differences that are much more robust to these variations, allowing us to quantify the intensity of each biomarker's colourimetric reaction more reliably.

The Control Negative is computed by first converting its pixels from RGB to CIELAB and then calculating a score for each pixel using our pixel scoring function. We then select the pixels in the lowest 25th percentile of scores to represent the Control Negative. The lowest 25th percentile is used, rather than a simple average, to ensure that any unexpected red signals, such as those caused by nanoparticle trapping or other artifacts, do not influence the measurement. Finally, the median of these selected pixels is taken to obtain a robust Control Negative value, which further reduces the impact of outliers. This yields a baseline score that represents a zero-concentration state.

Formally, if $scores_{\text{Control-N}}$ denotes the array of pixel scores for the Control Negative SCR, we compute:

$$\text{Score}_{\text{Control-N}} = \text{median}\{score_i \mid score_i \leq \text{percentile}(scores_{\text{Control-N}}, 25)\} \tag{3.28}$$

where percentile(scores, x) denotes the score value that $x\%$ of the elements in *scores* are less than or equal to.

For the three biomarker SCRs, we focus only on pixels that exhibit a visible colourimetric reaction. Simply averaging all pixel scores within the SCR would fail to accurately represent the true reaction, since it is often non-uniform across the surface. For example, a biomarker SCR may contain a small region of dark red pixels surrounded by unreacted areas. Averaging across the entire SCR in such cases would underestimate the reaction intensity, as the biomarker concentration is reflected by the intensity of red, not the area it covers.

To selectively isolate the pixels representing the reaction, we again apply our pixel scoring function. As with the Control Negative, we first convert the pixels from RGB to CIELAB colour space and then compute the pixel scores for each biomarker SCR. Recall that a high score indicates a pixel that is both strongly red and dark.

To identify the most significant pixels, we select only the pixels in the top 75th percentile, capturing the most intense and dark red regions that are likely representative of the true biomarker signal.

Formally, if $scores_{\text{SCR}}$ denotes the array of pixel scores for a biomarker SCR, we compute:

$$\text{Score}_{\text{SCR}} = \text{median}\{score_i \mid score_i \geq \text{percentile}(scores_{\text{SCR}}, 75)\} \tag{3.29}$$

This selection can be visualized by re-colouring these pixels in the original image, confirming that the algorithm accurately identifies the most pronounced reaction areas, as shown in Figure 3.11.

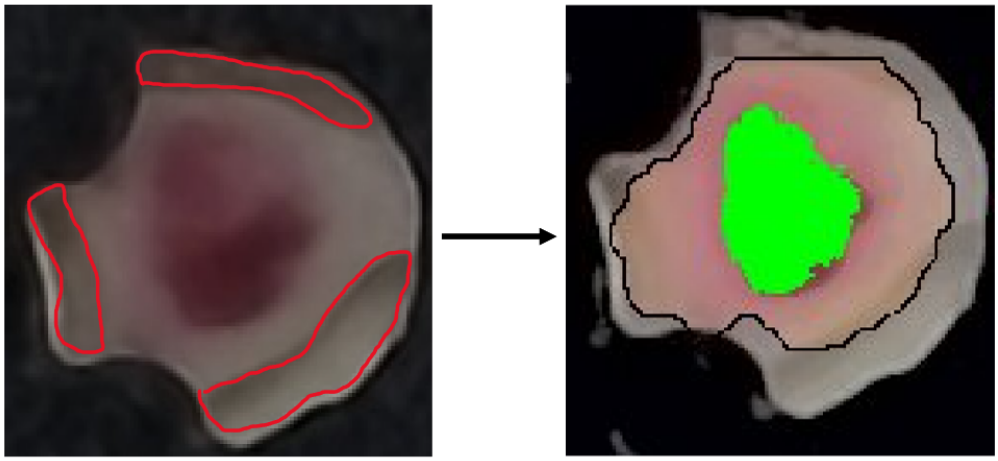


Figure 3.11: Visualization of the feature extraction process. (A) SCR before processing, with burn lines highlighted in red, resulting from the laser cutting used to create the SCR cutouts on the calibration layer. (B) The same SCR after processing, black outline indicates the segmentation (reduced by 20% to mitigate burn lines), and feature selection, where pixels in the top 75th percentile from the computed $(a^*)^3 \cdot (1 - L^*/100)$ scores are highlighted in green.

Finally, from this refined set of selected pixels, we compute the final feature representing the colourimetric response for each SCR. For each biomarker SCR, we calculate the feature ΔE as a nonnegative, log transformed difference between the score of the biomarker SCR and the score of the Control Negative SCR. Specifically:

$$\Delta E = \left(\log\left(1 + \max\left(\text{Score}_{\text{SCR}} - \text{Score}_{\text{ControlN}}, 0\right)\right), \qquad \text{SCR} \in \{\text{IL-6}, \text{IL-1beta}, \text{CRP}\}. \tag{3.30}$$

The logarithmic transformation is applied to ΔE because the biomarker concentration tends to level off as ΔE increases. Without this transformation, differences between very high ΔE

values would appear much larger than the actual differences in concentration. This transformation also helps stabilize and normalize ΔE. Additionally, ΔE is limited to positive values because negative values would indicate that the Control Negative SCR exhibits a higher signal than the biomarker SCR, which isn't an expected result but might occur due to device production faults or under conditions such as high humidity that reduce the effectiveness of the AuNPs. Algorithm 2 summarizes the complete feature extraction procedure.

Algorithm 2 Feature Extraction from SCRs using Pixel Scoring

Input: SCR data
Output: Feature set $\{\Delta E_{\text{biomarker}}\}$ for each biomarker

Step 1: Compute Control-Negative baseline
for each SCR in data **do**
    if SCR is Control-Negative **then**
        Normalize pixels to $[0, 1]$ and convert to CIELAB
        Compute pixel scores: $S_i = \texttt{get\_pixel\_scores}(pixel_i)$
        Compute 25th percentile: $Q_{25} = \text{percentile}(S, 25)$
        Select pixels $S_{\text{selected}} = \{S_i \mid S_i \leq Q_{25}\}$
        Compute Control-Negative score: $Score_{\text{Control-N}} = \text{median}(S_{\text{selected}})$
        Break
    end if
end for

Step 2: Extract features for biomarker SCRs
for each SCR in data **do**
    if SCR is Control-Negative or Control-Positive **then**
        Skip
    end if
    Normalize pixels to $[0, 1]$ and convert to CIELAB
    Compute pixel scores: $S_i = \texttt{get\_pixel\_scores}(pixel_i)$
    Compute 75th percentile: $Q_{75} = \text{percentile}(S, 75)$
    Select pixels: $S_{\text{selected}} = \{S_i \mid S_i \geq Q_{75}\}$
    Compute SCR score: $Score_{\text{SCR}} = \text{median}(S_{\text{selected}})$
    Compute preliminary difference: $d = Score_{\text{SCR}} - Score_{\text{Control-N}}$
    Clamp negative values: $d = \max(d, 0)$
    Compute feature: $\Delta E = \log(1 + d)$
    Store ΔE with biomarker identifier
end for

3.7 Concentration Prediction and Output

To map our computed ΔE feature to a concentration, we prepared data for a calibration curve. A calibration curve is a standard approach to relate measured responses to known

concentrations, allowing the prediction of unknown concentrations from new observed signals. We produced four sample patches for four concentration labels of 1, 10, 100, and 1000 ng/mL (nanograms per milliliter), yielding a total of 16 samples. To achieve a specific concentration, a droplet containing the precise amount of antigen was applied to each of the three biomarker SCRs on the patch, producing the corresponding colourimetric reaction.

After preparing the patches, the calibration layer was attached, and images were captured under controlled conditions using a Samsung Galaxy S24 Ultra. Each image was processed through the feature extraction pipeline, and the resulting ΔE values were compiled into a dataset.

To model the relationship between biomarker concentration and the measured response ΔE, we define ΔE as a function of concentration using a logarithmic model. This model is able to capture the rapidly increasing response at low concentrations and the slower rate of change at higher concentrations, reflecting the saturation behavior of the colourimetric signal. Although ΔE has already been log-transformed during feature extraction to stabilize large values and reduce the impact of extreme differences, the logarithmic model here serves as function to relate the measured response to concentration in a way that accounts for saturation at high concentrations. The model is defined as

$$\Delta E = f(C) = a \cdot \log(C) + b, \tag{3.31}$$

where C is the biomarker concentration and a and b are model parameters.

The parameters a and b were estimated for each biomarker using nonlinear least-squares fitting via the $curve\_fit$ function from the *scipy.optimize* library [45]. For each concentration level, the mean ΔE value was used as the representative response, and the best of fit curve was quantified using R^2. Three separate calibration models were fitted, one for each biomarker (IL-6, IL-1beta, and CRP). These models allow the prediction of biomarker concentration from any newly measured ΔE value.

To estimate the concentration corresponding to a new observed response ΔE we invert the logarithmic model so that the concentration can be expressed as a function of ΔE, allowing us to predict the biomarker concentration from the measured colourimetric change ΔE. This is defined as

$$C = f(\Delta E) = \begin{cases} 0, & \text{if } \Delta E = 0, \\ \exp\left(\frac{\Delta E - b}{a}\right), & \text{otherwise,} \end{cases} \tag{3.32}$$

where a and b are the fitted calibration parameters for the respective biomarker. When making predictions, if $\Delta E = 0$, the predicted concentration C is set to zero. Otherwise, C is computed using the fitted logarithmic calibration model for the corresponding biomarker.

Addtionaly, because our calibration dataset only includes concentrations up to 1000, predictions above this range are not reliable. To keep all estimates within the supported domain, we cap the predicted concentration at 1000 (later in the mobile application we report these as ≥ 1000 - Section 3.8):

$$C = \min(C, 1000) \tag{3.33}$$

Thus, if $\Delta E = 0$ the predicted concentration is set to zero. Otherwise, C is computed from the fitted model and then limited to a maximum value of 1000 to reflect the bounds of the calibration data.

The calibration curves for each biomarker, along with their fitted equations and corresponding R^2 values, are shown in Figures 3.12, 3.13, and 3.14. These plots illustrate how the model captures the relationship between concentration and the transformed ΔE feature.

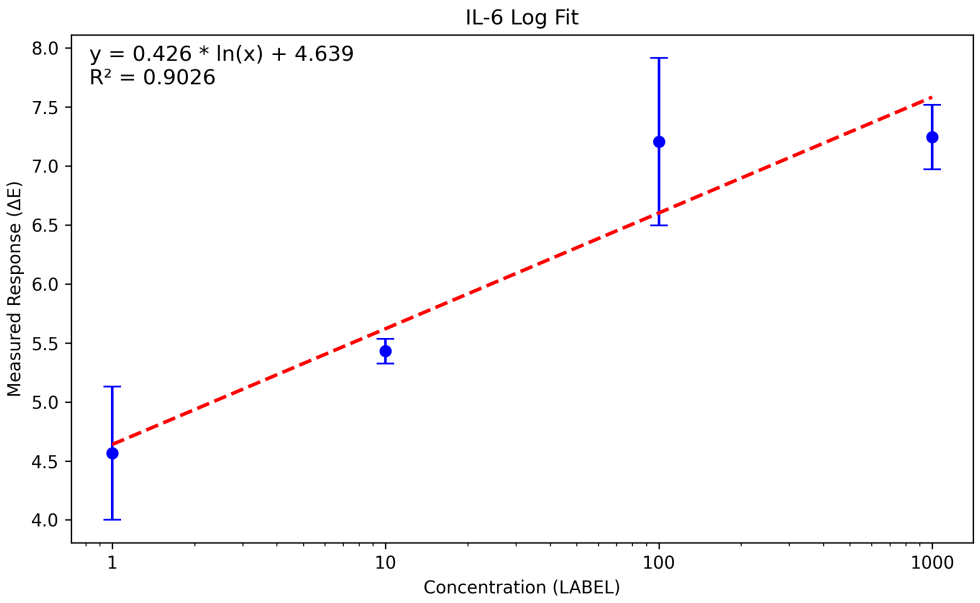


Figure 3.12: Calibration curve for IL-6. Points represent mean ΔE with standard deviation error bars and the red dashed line shows fitted result.

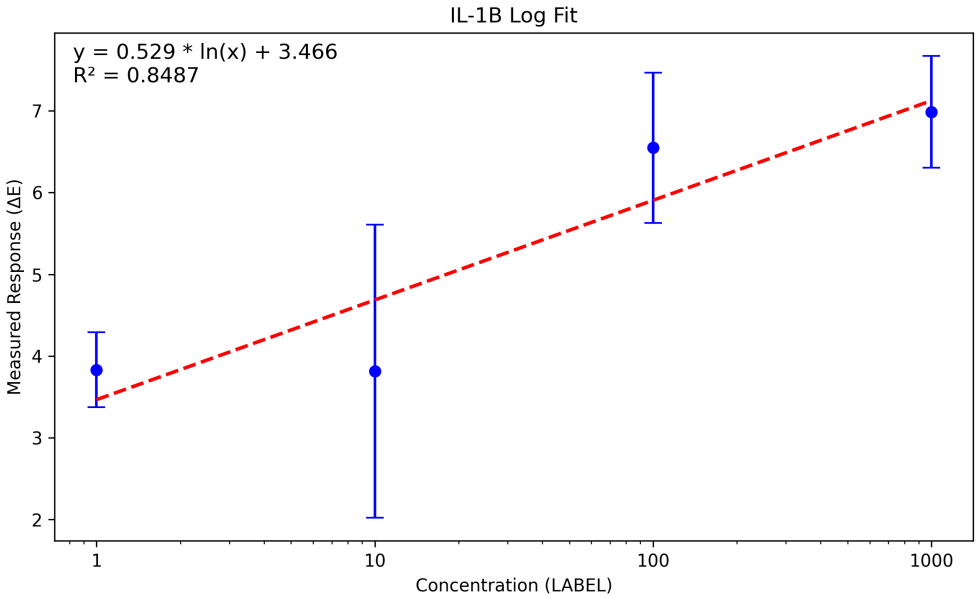


Figure 3.13: Calibration curve for IL-1beta. Points represent mean ΔE with standard deviation error bars and the red dashed line shows fitted result.

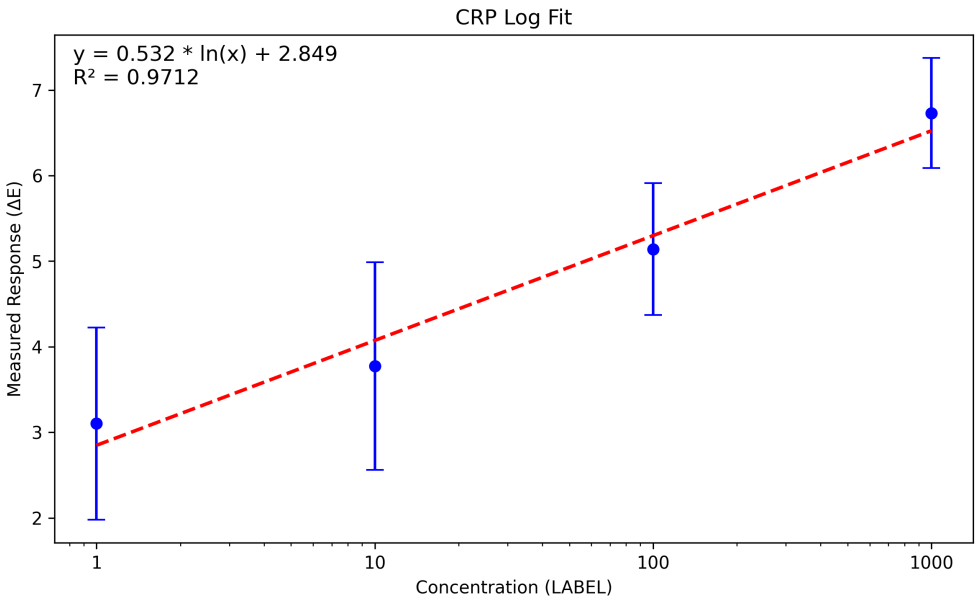


Figure 3.14: Calibration curve for CRP. Points represent mean ΔE with standard deviation error bars and the red dashed line shows fitted result.

3.8 Procedures and Mobile Application

This section outlines the steps required to use the sweat patch and introduces the mobile application, which provides a user-friendly interface for capturing, processing, and visualising the results produced by our pipeline. The combination of the sweat patch and mobile application allows for non-invasive, near real-time monitoring of biomarker levels through colourimetric responses. To use the sweat patch, it should be applied to an area of the body where sweat production is naturally higher, such as the back or the hands (Figure 3.15).

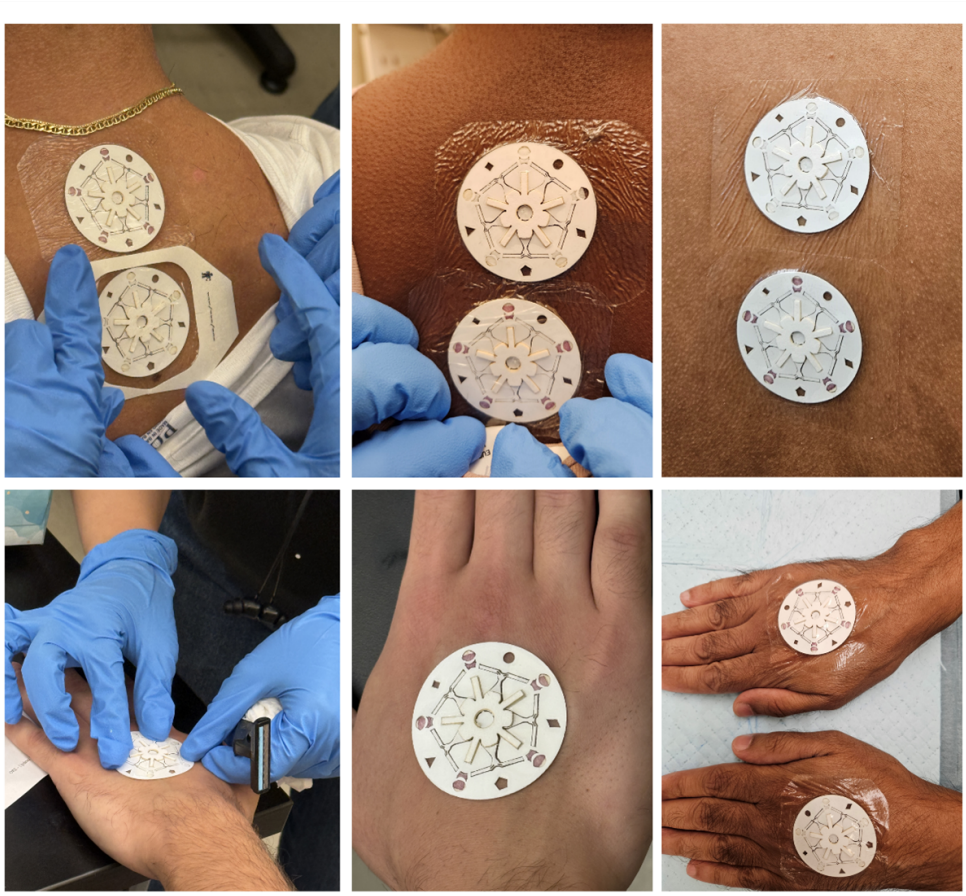


Figure 3.15: Two devices were applied to each participant's back or hand: a wearable sensing device for real-time colorimetric cytokine quantification and a wearable collector device for ELISA-based validation.

Sweat can be generated naturally through normal activity or induced through physical exercise, such as walking or running on a treadmill, or in a heated environment, such as a sauna. In this study, participants were instructed to complete a standardized activity protocol consisting of 30 minutes of brisk walking on a treadmill (4 miles per hour (mph)), with 5-minute rest intervals following every 10 minutes of walking (Figure 3.16).

Figure 3.16: On-body clinical evaluation of the device in healthy and patient cohorts with sweat-collection protocol during brisk walking on a treadmill (4 miles per hour (mph)).

When sweat is produced, it enters the device and flows through the microfluidic channels into the sweat collection region (SCR). In our testing, it typically takes approximately 10–15 minutes for enough sweat to accumulate in the SCR to trigger a colourimetric response. Even if sweat is not actively induced, the patch can still function. However, the time required for sweat to accumulate sufficiently may be longer. Based on our testing, passive sweat accumulation can take approximately 30 minutes before the colourimetric response becomes detectable.

After 10–30 minutes, once a colourimetric response is visible in the control positive SCR, the user can capture an image of the patch. The control positive SCR is a reference region included in the patch that should always produce a colourimetric response, confirming that the device is functioning correctly. If the patch was applied to the back, it can be removed and placed on a flat surface for imaging. If it was applied to the hand, the user can take a photo of the patch directly.

This process is facilitated by the mobile application, which guides the user through capturing and submitting the image for processing. Once submitted, the image is sent to our servers where the pipeline is executed, taking only 1–3 seconds to process each image.

The application then displays the biomarker concentrations derived from the extracted ΔE. The value of ΔE itself is not shown to the user, as interpreting it requires technical knowledge of the pipeline. The processed image is also displayed to the user, showing the segmented SCRs and highlighting in green the pixels in each SCR that were selected for analysis. It is also important to note that, as our calibration curves are capped at a concentration of 1000 ng/mL, any predicted concentrations at or exceeding this value are reported as ≥ 1000 ng/mL.

Figure 3.17 illustrates the complete workflow within the application, displaying the home screen, the image capture interface, and the final results screen with the processed visual output.

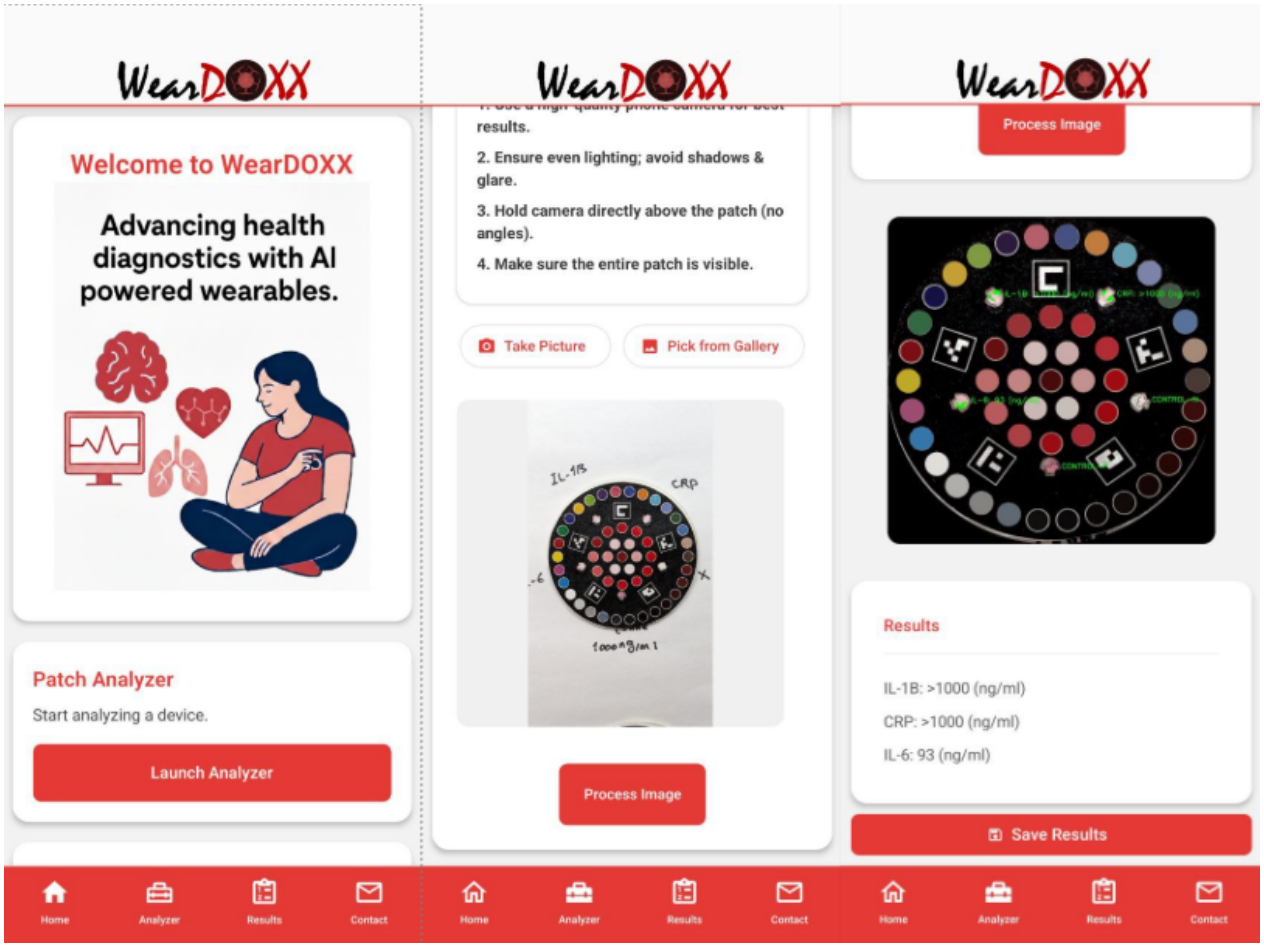


Figure 3.17: The WearDOXX mobile application interface. Left: Home screen allowing users to launch the analyzer. Center: Image capture interface where users can take a photo or upload from the gallery, accompanied by guidelines for optimal capture. Right: Results screen showing the processed patch image with analyzed pixels highlighted in green, alongside the calculated concentrations for CRP, IL-1beta, and IL-6.

Chapter 4

Experimental Evaluation

In this chapter, we present a comprehensive experimental evaluation designed to validate the proposed pipeline's performance, consistency, and accuracy. We begin by assessing the computational efficiency of the system to ensure that it meets the expectations for rapid POCT. Subsequently, we evaluate the robustness of our image normalization and feature extraction algorithms under varying conditions, including differences in lighting, smartphone hardware, and capture distances. Finally, we compare our pipeline smartphone-based measurements against Enzyme-Linked Immunosorbent Assay (ELISA) results using data that has been collected from human participants.

4.1 Evaluating Processing Time

To validate the near real-time processing speed of our pipeline for POCT, we evaluated its processing time performance on desktop processors that are comparable to the server hardware that will host the deployed pipeline. This server will handle processing requests that are submitted by users through the mobile application. We aim to show that the pipeline can process samples in only a few seconds, demonstrating its rapid processing speed and potential for near real-time results to users.

4.1.1 Experimental Setup

We measured the computational processing time of the pipeline on two desktop processor architectures: an Intel Ultra 5 125H (a modern mid-range desktop processor) and an AMD Ryzen 5 7600 (a high-performance desktop processor). Both processors were tested under identical software conditions, including the same operating system, environment, and all pipeline dependencies.

The dataset that was used in these experiments consisted of 30 sample images of patches, which were drawn from a combination of the calibration dataset and other datasets used in other experiments. For each processor, the complete dataset was processed 10 times to account for variability in runtime due to system load or caching effects. The processing time for each sample was recorded in seconds, from image loading through feature extraction and predicted concentration computation. The mean processing time over the 10 runs was then taken as the representative processing time for each sample. The resulting data were aggregated to compute descriptive statistics, including mean, median, standard deviation, and range for each processor.

This experiment provides insight into the practical deployment of the pipeline in POCT settings, allowing us to quantify performance differences across hardware platforms and assess whether processing speed meets near real-time expectations.

4.1.2 Results

The processing time results for the 30 sample images are summarized in Table 4.1 and Figure 4.1. Table 4.1 presents descriptive statistics for each processor, including the mean, standard deviation, minimum and maximum values, as well as the 25th, 50th (median), and 75th percentiles. Percentiles indicate the value below which a certain percentage of observations fall, for example, the 25th percentile represents the processing time below which 25% of samples were processed. Figure 4.1 shows the mean processing time per sample for both processors across the dataset. Each point represents the average processing time of a sample over 10 runs.

Table 4.1: Mean processing times per sample (in seconds) for each processor.

	AMD	Intel
Number of Samples	30	30
Mean	1.00	2.20
Std	0.04	0.31
Min	0.91	1.88
25%	0.97	2.00
50%	1.00	2.11
75%	1.03	2.33
Max	1.08	3.40

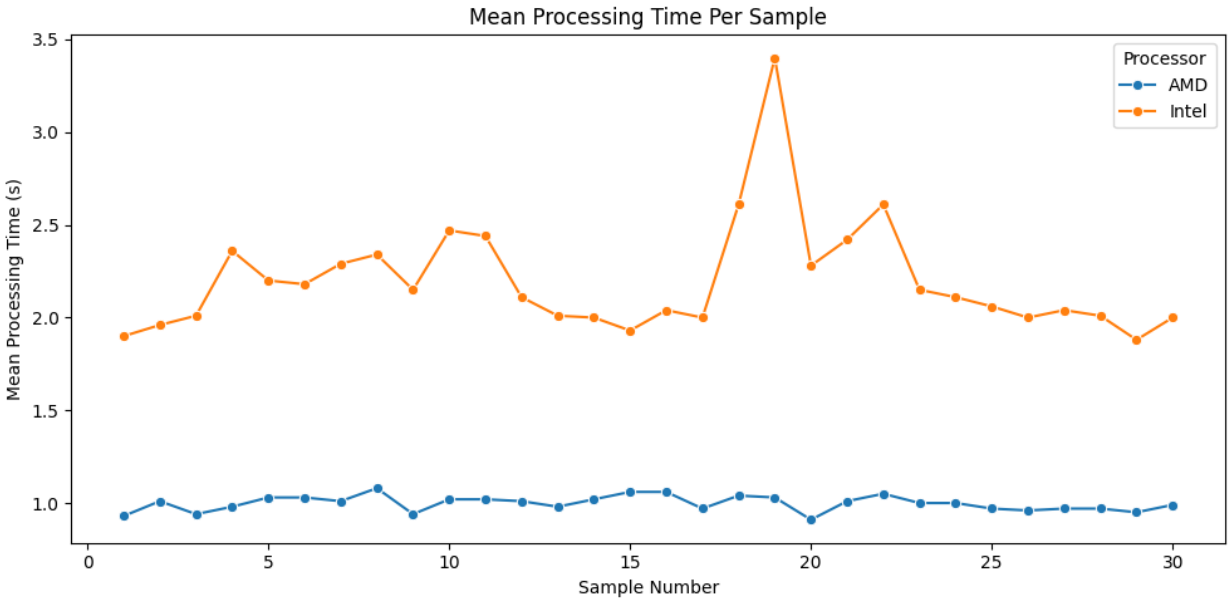


Figure 4.1: Mean processing time per sample for the AMD and Intel processors. Each point represents the mean over 10 runs for a given sample.

4.1.3 Discussion

The processing time results demonstrate that the pipeline is capable of operating within a near real-time range, which is essential for POCT. Across all samples, the maximum observed processing time was 3.40 seconds on the Intel processor. Even when using this peak value as an upper bound, the response time remains well within an acceptable processing time for rapid feedback. In practice, most samples were processed much faster than the maximum. The 75th percentile values show that the majority of samples fell between 1.03 seconds on the AMD processor and 2.33 seconds on the Intel processor, meaning that typical processing times were only one to two seconds. The AMD processor consistently achieved lower times, which is expected given that it is a generally faster processor than the Intel model that is used in this experiment.

The results also show that the pipeline exhibits relatively consistent performance across repeated runs. The standard deviations for mean processing time were approximately 0.04 seconds for the AMD processor and 0.31 seconds for the Intel processor. These low values indicate minimal fluctuation across the 10 repeated executions, meaning that common runtime variations such as caching or background processes have limited impact on overall speed. Although one sample on the Intel system reached 3.40 seconds, this appears to be an outlier rather than a typical case.

Overall, the numerical results confirm that the processing pipeline is both fast and stable. Even in the slowest recorded case (3.4 seconds), the processing time remained only a few seconds, and the typical processing time for most samples was between one and two seconds. This level of performance supports the use of the pipeline for POCT settings, where timely responses are expected.

4.2 Evaluating Image Normalization

The purpose of this experiment is to evaluate the effect of our two-step image normalization process on the consistency of the extracted colourimetric feature ΔE from the SCRs. Specifically, we aim to determine how applying illumination normalization followed by colour normalization impacts the variability in our extracted feature ΔE across multiple samples with varying lighting conditions. This assessment is important to understand how normalization affects the reliability and consistency of biomarker measurements across different lighting conditions.

4.2.1 Experimental Setup

To quantify variability, we calculate the Coefficient of Variation (CV) for each biomarker in five samples across six variations. The CV is defined as the ratio of the standard deviation to the mean for the measured ΔE values within each sample:

$$\text{CV} = \frac{\sigma}{\mu} \tag{4.1}$$

where σ is the standard deviation, and μ is the mean of ΔE. A lower CV indicates more consistent and reliable measurements across variations, while a higher CV suggests greater variability.

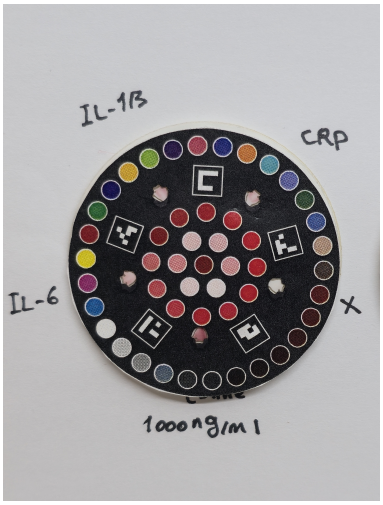
We use the CV here because enabling or disabling normalization changes the scale of the data. Specifically, the ΔE values differ depending on whether normalization is applied, due to both the luminance matching to the reference and the colour correction model transformations applied to the pixels. By using CV, we obtain a scale-independent measure of variability, allowing a meaningful comparison between the normalized and unnormalized datasets.

The dataset for this experiment was generated by applying a series of controlled augmentations to five sample images of concentrations 0, 10, 50, 100, and 1000. The augmentations include:

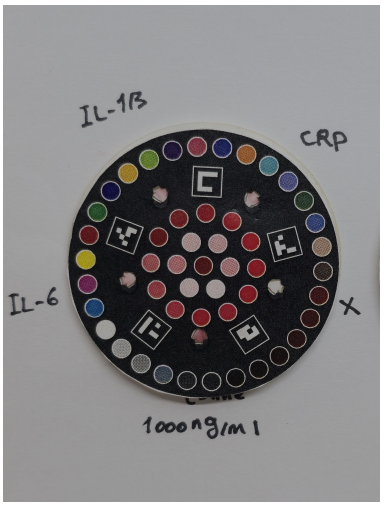
- Brightness adjustments (0.8× and 1.2×)
- Contrast adjustments (0.8× and 1.2×)
- Warm and cool colour casts

For each sample, the six augmented variations were analyzed together with the original image, resulting in a total of seven samples per concentration for the subsequent analysis. These variations simulate common environmental and acquisition differences that may occur in practical settings. The original images and augmented variations were processed through our pipeline twice: once with the two-step normalization applied and once without any normalization.

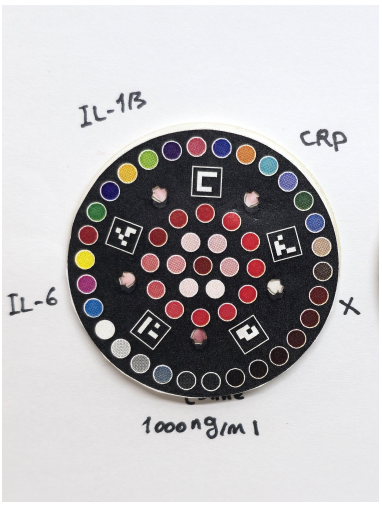
For each, the ΔE values for all SCRs were extracted and grouped by biomarker. Mean and standard deviation values were computed for each biomarker and sample, and the CV was subsequently calculated. Comparison of the CVs between normalized and unnormalized conditions allows us to assess the effectiveness of the two-step normalization in reducing variability and improving the robustness of our measured response ΔE. Figure 4.2 shows a sample image with concentration 1000 and its six augmented variations.



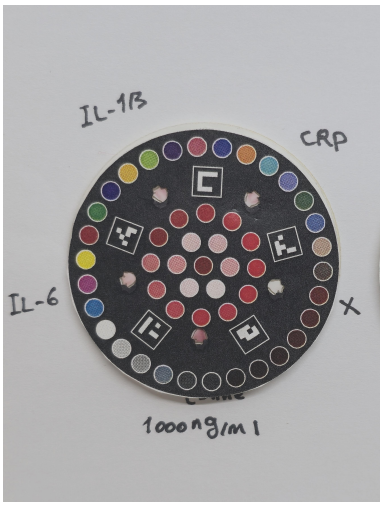
(a) Original



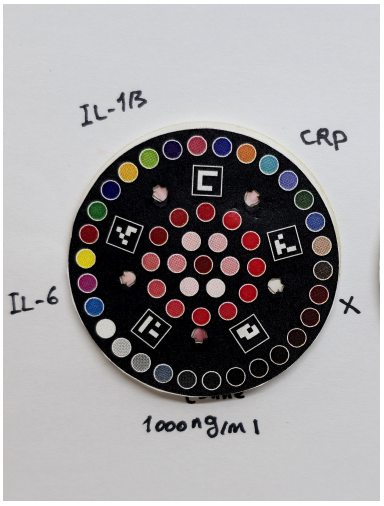
(b) Bright ×0.8



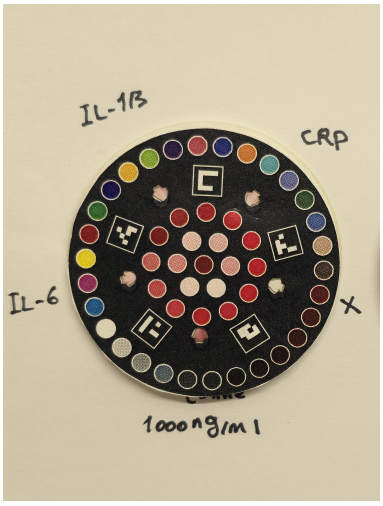
(c) Bright ×1.2



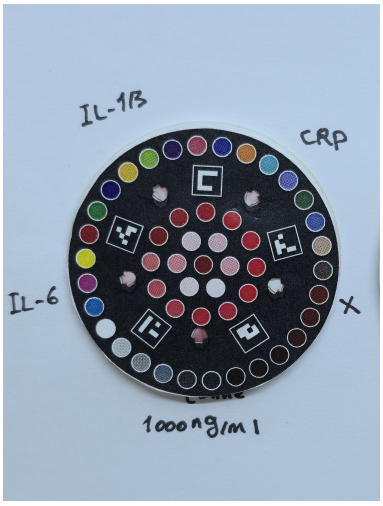
(d) Contrast ×0.8



(e) Contrast ×1.2



(f) Warm cast



(g) Cool cast

Figure 4.2: Comparison of the original image (top row) and its six variations (bottom two rows) for a single sample. Each shows one variation that was applied to the original image.

4.2.2 Results

The detailed numerical results, including the specific ΔE values for each sample and condition, for both normalized and unnormalized data, are provided in Appendix A (Tables 1 and 2). Figures 4.3 through 4.5 show the comparison of the CV for the CRP, IL-1beta, and IL-6 biomarkers. These plots demonstrate the variability in ΔE measurements across all samples with (blue) and without (orange) normalization.

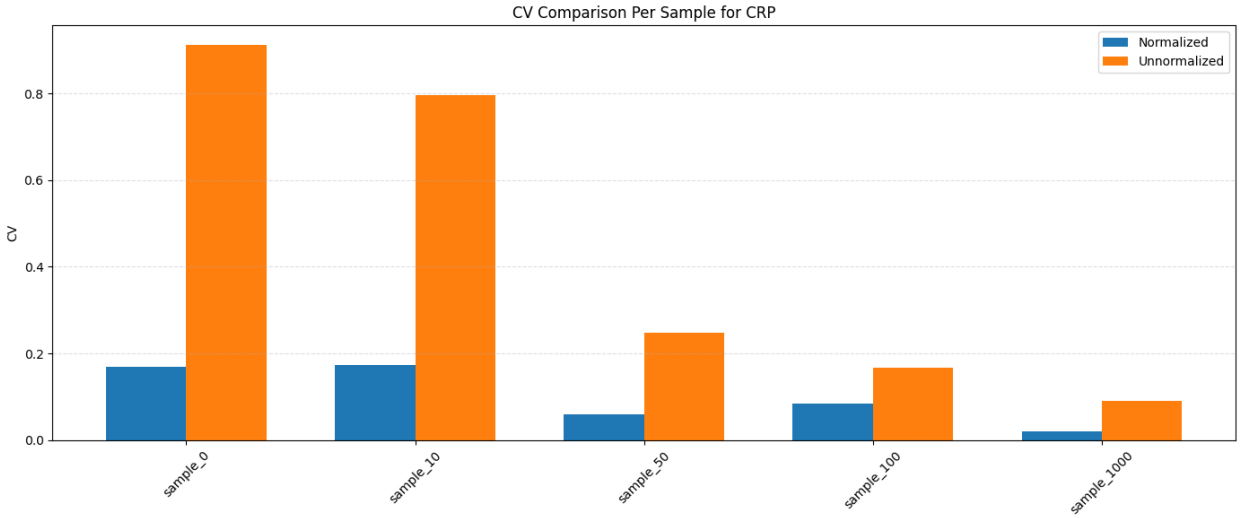


Figure 4.3: Coefficient of Variation (ΔE) cComparison for CRP across all samples, showing the reduction in variability after two-step image normalization.

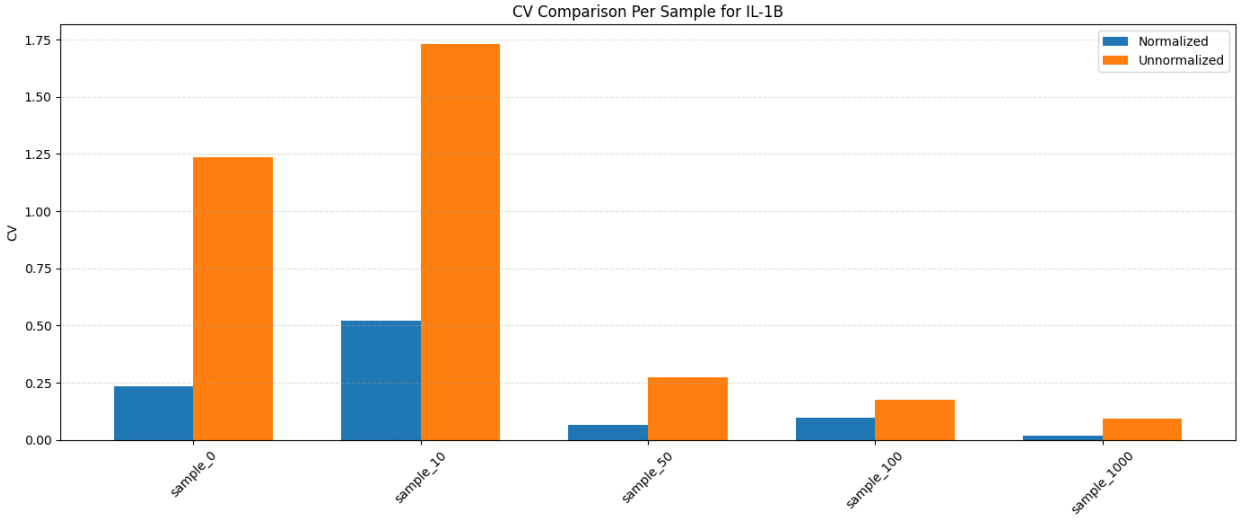


Figure 4.4: Coefficient of Variation (ΔE) comparison for IL-1beta across all samples, illustrating the improved consistency after image normalization.

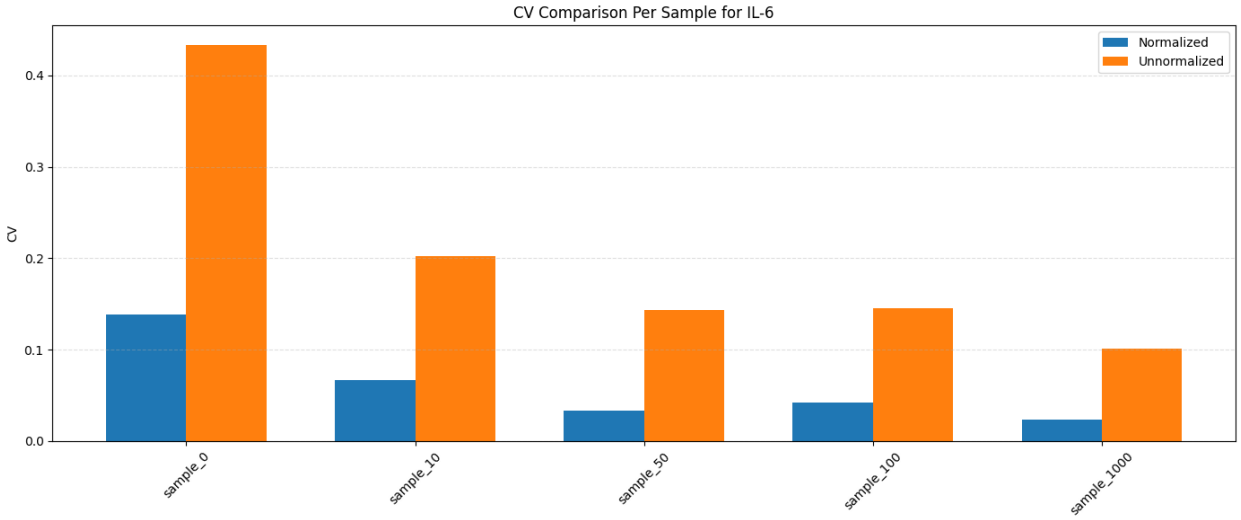


Figure 4.5: Coefficient of Variation (ΔE) comparison for IL-6 across all samples, demonstrating the significant reduction in ΔE variability due to the normalization process.

The full statistical results, including all CV values for normalized and unnormalized conditions for each sample, are provided in Table 4.2.

Table 4.2: Coefficient of Variation (CV) Statistics for Normalized vs. Unnormalized ΔE

Biomarker	Sample	CV_{norm}	CV_{unnorm}	Percent Reduction (%)	Avg Percent Reduction (%)	Lowest Percent Reduction (%)	Highest Percent Reduction (%)
CRP	sample_0	0.17	0.91	-81.42			
CRP	sample_10	0.17	0.80	-78.17			
CRP	sample_50	0.06	0.25	-75.96	-72.45	-49.03 (sample_100)	-81.42 (sample_0)
CRP	sample_100	0.08	0.17	-49.03			
CRP	sample_1000	0.02	0.09	-77.65			
IL–1B	sample_0	0.23	1.24	-81.11			
IL–1B	sample_10	0.52	1.73	-69.89			
IL–1B	sample_50	0.07	0.27	-76.07	-70.57	-45.63 (sample_100)	-81.11 (sample_0)
IL–1B	sample_100	0.10	0.18	-45.63			
IL–1B	sample_1000	0.02	0.09	-80.15			
IL–6	sample_0	0.14	0.43	-68.11			
IL–6	sample_10	0.07	0.20	-67.04			
IL–6	sample_50	0.03	0.14	-76.73	-71.93	-67.04 (sample_10)	-76.73 (sample_50)
IL–6	sample_100	0.04	0.14	-71.20			
IL–6	sample_1000	0.02	0.10	-76.56			

4.2.3 Discussion

For all three biomarkers analyzed, CRP, IL-1beta, and IL-6, the Normalized condition (blue bars) consistently show a lower CV compared to the Unnormalized condition (orange bars) across all tested samples (concentrations 0, 10, 50, 100, and 1000). Looking closer at the results we see that for CRP, the two-step normalization process yielded a substantial average CV reduction of approximately 72.45%. Meaning on average, for CRP, CV was reduced by 72.45%, indicating the normalization greatly increased consistency among sample variations. The greatest reduction for CRP was 81.42% in the *sample_0* concentration, while the lowest

reduction was 49.03% in the *sample_100* concentration. Similarly, the normalization process achieved an average CV reduction of 70.57% for IL-1beta, with the highest reduction observed at 81.11% again in *sample_0* and the lowest at 45.63% in *sample_100*. Finally, IL-6 demonstrated a comparable average CV reduction of 71.93%, peaking at 76.73% in *sample_50* and having its lowest reduction at 67.04% in *sample_10*. These findings collectively demonstrate that image normalization is highly effective across all biomarkers, providing a significant average reduction in variability.

We consistently observed approximately 70% improvement in consistency across all three biomarkers, suggesting that the two-step normalization procedure can effectively reduce the variability introduced by different lighting conditions. By reducing measurement variability, normalization ensures that subsequent analyses, such as concentration estimation or comparisons across samples, are more reliable. Overall, these results confirm that the normalization approach substantially enhances the consistency of the colourimetric measurements across different lighting conditions.

Interestingly, we observed that CV generally decreases as concentration increases for all biomarkers. This trend can be easily observed in Table 4.2 for both normalized and unnormalized data, and is likely attributable to both our pixel selection process and the use of a log transformation when computing ΔE. At higher concentrations, the colourimetric reaction in the SCR is more pronounced, producing pixels with a stronger red hue. As a result, selecting the top 25th percentile of pixels using our scoring method is more consistent across sample variations, since it is easier to identify the reaction pixels. Additionally, the log transformation compresses larger differences between the biomarker SCR and the control negative SCR, reducing the relative effect of very high values. This further reduces variability in the measurements at higher concentrations, contributing to the observed decrease in CV. In contrast, at lower concentrations, fewer pixels exhibit the red hue, and differences in SCR signals are smaller but more variable, making the top 25th percentile selection less consistent. Overall, these effects indicate that measurements at higher concentrations are more robust across different capture conditions, which is particularly advantageous for POCT.

4.3 Evaluating Different Devices

To assess the consistency of measurements across different smartphone devices, we conducted an experiment using images captured from three smartphone models: Galaxy S24 Ultra, Galaxy Z Fold 5, and Pixel 10. The aim of this experiment is to evaluate whether the mea-

surements extracted from these images are consistent across devices.

4.3.1 Experimental Setup

Images of 19 patch samples with varying concentrations were collected using three different devices. For each sample, an image was captured with each device, resulting in three images per sample. All images were taken under consistent lighting, zoom, and a relative capture distance of 10–15 cm from the patch. This process produced a total dataset of 57 images (3 devices $\times$ 19 samples). Each patch image was processed through our pipeline, and for each biomarker, we recorded the ΔE values and the predicted concentration obtained from the corresponding predictive model.

To evaluate consistency across devices, we computed the intraclass correlation coefficient (ICC) [46] for each biomarker and measurement type. ICC is a statistical measure that quantifies the consistency or agreement of measurements made by different devices (acting as raters) on the same patch samples (subjects). In this context, it assesses whether the same patch yields consistent measurements across devices and is reported on a scale from 0 to 1, where 0 indicates no agreement and 1 indicates perfect agreement. There are different variations of ICC. For guidance on selecting the appropriate model, we followed the recommendations of Koo and Li [47], who provide a detailed framework for choosing and reporting ICC in reliability research.

Based on the Koo and Li framework, we focused on two types of ICC for each measurement (ΔE and Concentration): ICC(2,1) and ICC(3,1). Both quantify absolute agreement, meaning they reflect how closely the devices match each other in their actual values, rather than simply following the same trend or pattern (as Pearson correlation would, which only measures whether the devices' measurements are linearly correlated). The key difference between ICC(2,1) and ICC(3,1) is that ICC(2,1) treats the devices as random effects, whereas ICC(3,1) treats them as fixed effects.

Formally, ICC(2,1) is a two-way random-effects model for single measurements, where both the patch samples and devices are treated as random effects. This allows generalization to devices not included in this experiment. ICC(2,1) is calculated as

$$\text{ICC(2,1)} = \frac{MS_{\text{between samples}} - MS_{\text{residual}}}{MS_{\text{between samples}} + (k-1)MS_{\text{residual}} + \frac{k}{n}(MS_{\text{between devices}} - MS_{\text{residual}})}, \quad (4.2)$$

where $MS_{\text{between samples}}$ is the mean square between samples, $MS_{\text{between devices}}$ is the mean square between devices, MS_{residual} is the residual mean square, k is the number of devices, and n is the number of samples [46, 47]. ICC(2,1) evaluates whether devices preserve the relative order of samples, meaning that samples with higher values on one device tend to have higher values on the other devices.

ICC(3,1) is a two-way mixed-effects model for single measurements, where the patch samples are treated as random effects and the devices are treated as fixed effects. This evaluates the absolute agreement of measurements for the specific devices used in this study without generalizing to other devices. ICC(3,1) is computed similarly to ICC(2,1), but the between-device term is treated as fixed rather than random.

$$\text{ICC(3,1)} = \frac{MS_{\text{between samples}} - MS_{\text{residual}}}{MS_{\text{between samples}} + (k-1)MS_{\text{residual}}}, \quad (4.3)$$

For instance, consider a single sample measured across three devices. If the measured ΔE values are similar (e.g., device A = 5, device B = 5.1, device C = 4.9), this reflects strong absolute agreement across devices. ICC(2,1) captures this agreement assuming the devices are pulled randomly from a larger population, meaning it estimates how well other devices not included in this study would agree on the same sample. ICC(3,1), in contrast, captures the agreement only for the specific devices used, without generalizing to other devices. Conversely, if the predicted concentrations vary substantially across devices (e.g., device A = 2.3, device B = 4.1, device C = 3.8), both ICCs would be low, but ICC(2,1) emphasizes that this poor agreement might extend to other devices as well.

We chose to report both ICC(2,1) and ICC(3,1) because they provide complementary insights. ICC(2,1) allows us to evaluate how well measurements generalize to a broader population of devices, while ICC(3,1) shows how reliably a sample can be measured using relatively modern devices like the three we tested. Using both helps us understand both generalizability and performance with current technology. Figure 4.6 shows a sample captured with the Pixel 10, Galaxy Z Fold 5 and Galaxy S24 Ultra.

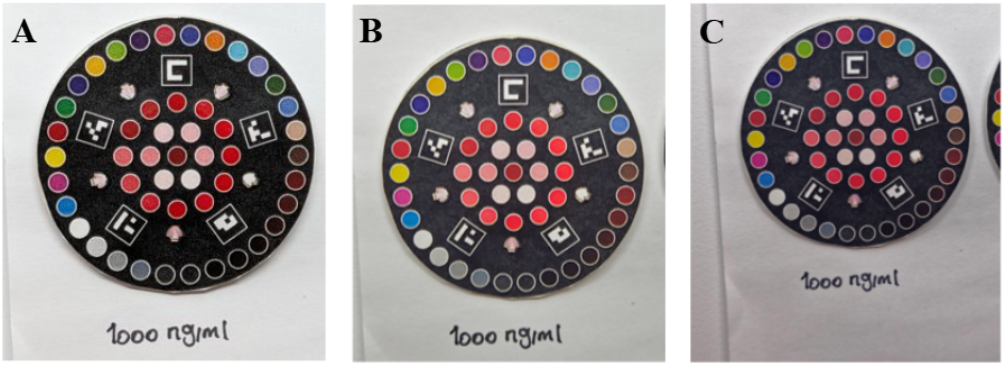


Figure 4.6: Example sample captured with three devices. (A) Pixel 10, (B) Galaxy Z Fold 5, (C) Galaxy S24 Ultra.

4.3.2 Results

ICC(2,1) and ICC(3,1) values were computed using the Pingouin library [48] to assess the reliability and device-specific consistency of measurements across the three devices (Galaxy S24 Ultra, Galaxy Z Fold 5, Pixel 10). Table 4.3 summarizes the results for the ΔE feature, while Table 4.4 summarizes the ICC values for the predicted concentrations derived from ΔE.

Table 4.3: ICC(2,1) and ICC(3,1) values for the ΔE feature across devices.

Biomarker	ICC(2,1)	ICC(3,1)	p-value (2,1)	p-value (3,1)
CRP	0.6665	0.8182	$\leq$0.05	$\leq$0.05
IL-1beta	0.8140	0.9365	$\leq$0.05	$\leq$0.05
IL-6	0.6932	0.8467	$\leq$0.05	$\leq$0.05

Table 4.4: ICC(2,1) and ICC(3,1) values for predicted concentrations derived from ΔE.

Biomarker	ICC(2,1)	ICC(3,1)	p-value (2,1)	p-value (3,1)
CRP	0.4538	0.4926	$\leq$0.05	$\leq$0.05
IL-1beta	0.4617	0.5218	$\leq$0.05	$\leq$0.05
IL-6	0.3891	0.4736	$\leq$0.05	$\leq$0.05

4.3.3 Discussion

According to Koo and Li [47], ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good

reliability, and values greater than 0.90 indicate excellent reliability. We will use these guidelines to assess the results.

First, we can observe that all ICC values for both ΔE and Concentrations are statistically significant ($p \leq 0.05$), confirming that the following observations are unlikely to have occurred by chance.

The ICC(2,1) values for the ΔE feature, which reflect the reliability of measurements for generalization to a broader population, indicate moderate to good consistency across devices. For example, IL-1beta shows an ICC of 0.81, suggesting strong agreement in sample measurements across devices. CRP and IL-6 have ICCs of 0.67 and 0.69, respectively, indicating lower agreement, but they are still considered moderately reliable.

In contrast, the ICC(2,1) values for the predicted concentrations are lower, ranging from 0.39 to 0.46, indicating much weaker agreement between devices. This suggests that while the raw ΔE measurements are fairly reliable, converting these measurements into predicted concentrations introduces additional variability. A key factor contributing to these reduced ICC values is the limited size of the calibration dataset, which included only 16 samples. This small dataset restricts the predictive models' ability to generalize to new samples, resulting in poor consistency across devices. For example, IL-1beta has an ICC of 0.46 for predicted concentrations, whereas its corresponding ΔE ICC is 0.81, highlighting a substantial discrepancy between the reliability of ΔE measurements and concentration predictions.

For ICC(3,1), which treats the devices as fixed effects and evaluates the consistency of measurements for the specific devices used in this experiment, the values are higher for both ΔE and predicted concentrations. For example, IL-1beta shows an ICC(3,1) of 0.94 for ΔE and 0.52 for predicted concentrations. The increase in ΔE ICC values for ICC(3,1) is expected, as here we are not generalizing to a broader population of devices, but focusing purely on the reliability between the three devices used.

While the ICC(3,1) values for predicted concentrations are higher than their ICC(2,1) counterparts, that increase is relatively smaller compared to the increase observed in the ΔE measurements. This suggests that even when considering only the devices in the experiment, variability in predicted concentrations still remains largely driven by limitations of the predictive model. Therefore, improving model performance and expanding the calibration dataset are critical for achieving higher reliability in predicted concentrations.

Even though ICC(2,1) values show only moderate reliability when generalizing to a broader population of devices, the three devices used in this study represent modern smartphone cameras. The good reliability observed in ICC(3,1) gives us confidence that, when using at least a relatively modern device, measurements of the extracted feature ΔE indicate good reliability. Similarly, predicted concentrations are expected to become more reliable as the calibration dataset increases.

Overall, these results suggest that ΔE measurements are moderately to highly reliable across devices. However, predicted concentrations show lower reliability due to model limitations and the small size of the calibration dataset.

4.4 Evaluating Capture Distances

To evaluate how variations in capture distance affect our measurements, we conducted an experiment capturing images of samples at three distances: 10 cm, 15 cm, and 20 cm. The goal of this experiment is to determine whether the measurements (ΔE and predicted concentrations) remain consistent across distances. This allows us to assess whether small variations in camera distance affect the results, which is crucial for practical scenarios where precise positioning cannot be guaranteed. It also provides a way to validate our Processing and Alignment (Section 3.3) step, determining whether it effectively reduces variability across different capture distances or zoom settings.

4.4.1 Experimental Setup

Images of 10 samples of varying concentrations were captured using a Samsung Galaxy S24 Ultra at three different distances: 10 cm, 15 cm, and 20 cm. To ensure consistent distances, a ruler was placed next to the sample, and the phone was positioned along the ruler while taking each picture. All images were captured under the same lighting conditions. This process resulted in a total dataset of 30 images (10 samples $\times$ 3 distances). Each image was processed through our imaging pipeline, and for each biomarker, we recorded the ΔE and the predicted concentration values. Figure 4.7 shows an example of a single sample captured at the three distances of 10 cm (A), 15 cm (B), and 20 cm (C).

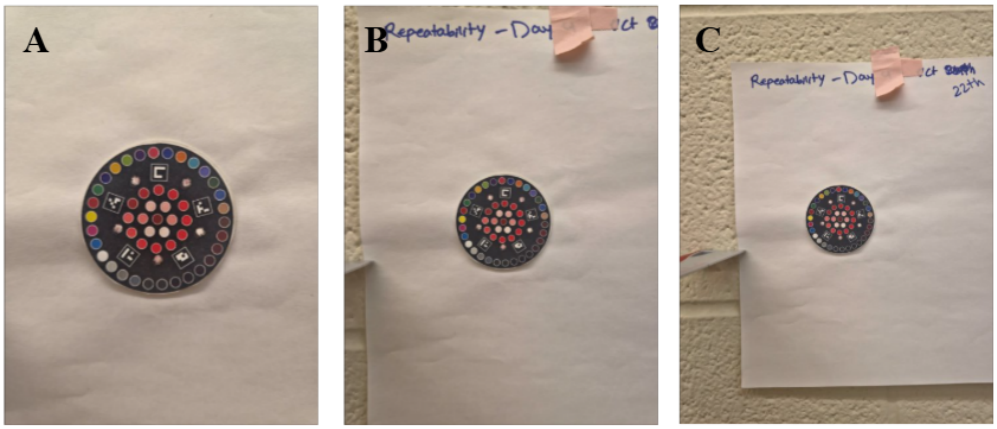


Figure 4.7: Example sample captured at different distances. (A) 10 cm, (B) 15 cm, (C) 20 cm.

To evaluate consistency across capture distances, we again computed ICC(2,1) and ICC(3,1) for each measurement (ΔE and predicted concentration). As in the previous experiment, ICC provides a quantitative measure of agreement, indicating whether the same sample produces similar measurements at different distances. ICC(2,1) assesses absolute agreement while treating the capture distances as random effects, providing insight into generalizability to other potential distances. ICC(3,1), in contrast, treats the distances as fixed effects, evaluating the reliability of measurements specifically for the distances tested in this experiment (10 cm, 15 cm, 20 cm).

Additionally, to specifically assess reliability between the more practical capture distances, we focused on 10 cm and 15 cm. Distances in this range represented a more typical usage, as 20 cm required noticeably more movement to capture the sample, and distances under 10 cm were not included because images began to blur. We applied the Wilcoxon signed-rank test [49] to evaluate whether the differences in paired measurements between 10 cm and 15 cm were statistically significant, without assuming normality.

Formally, to compute the Wilcoxon signed-rank test, let d_i represent the difference between measurements at 10 cm and 15 cm for sample i, and let R_i be the rank of $|d_i|$ ignoring the sign. The Wilcoxon signed-rank statistic W is then computed as:

$$W = \sum_{i:d_i>0} R_i \tag{4.4}$$

where the sum is taken over all positive differences. The null hypothesis of the test is that the median difference between paired measurements is zero, and in this case would indicate no systematic change between 10 cm and 15 cm capture distances. If $p \leq 0.05$, we reject

the null hypothesis, suggesting that the measurements differ significantly between the two distances. Conversely, a non-significant p-value ($p > 0.05$) indicates that the measurements are consistent, showing no evidence of a systematic difference.

This combination of ICC and the Wilcoxon signed-rank test allows us to quantify the reliability of both the overall and the more typical capture distances.

4.4.2 Results

The results are presented in Tables 4.5 to 4.8. Tables 4.5 and 4.6 report ICC(2,1) and ICC(3,1) values for the ΔE feature and the predicted concentrations derived from ΔE, respectively. Tables 4.7 and 4.8 show the results of the Wilcoxon signed-rank test for paired measurements between the 10 cm and 15 cm capture distances, for ΔE values and predicted concentrations, respectively.

Table 4.5: ICC(2,1) and ICC(3,1) values for the ΔE feature across distances.

Biomarker	ICC(2,1)	ICC(3,1)	p-value (2,1)	p-value (3,1)
CRP	0.8339	0.8348	$\leq$0.05	$\leq$0.05
IL-1beta	0.7445	0.7418	$\leq$0.05	$\leq$0.05
IL-6	0.6126	0.6225	$\leq$0.05	$\leq$0.05

Table 4.6: ICC(2,1) and ICC(3,1) values for predicted concentrations derived from ΔE.

Biomarker	ICC(2,1)	ICC(3,1)	p-value (2,1)	p-value (3,1)
CRP	0.8526	0.8516	$\leq$0.05	$\leq$0.05
IL-1beta	0.6168	0.6692	$\leq$0.05	$\leq$0.05
IL-6	0.4769	0.4957	$\leq$0.05	$\leq$0.05

Table 4.7: Wilcoxon signed-rank test results for ΔE values between 10 cm and 15 cm capture distances.

Biomarker	p-value	p-value > 0.05
CRP	0.85	Yes
IL-1beta	0.19	Yes
IL-6	0.08	Yes

Table 4.8: Wilcoxon signed-rank test results for predicted concentrations between 10 cm and 15 cm capture distances.

Biomarker	p-value	p-value > 0.05
CRP	0.83	Yes
IL-1beta	0.19	Yes
IL-6	0.08	Yes

4.4.3 Discussion

The goal of this experiment was to determine whether variations in camera distance introduce measurable changes in the extracted ΔE feature and in the concentrations predicted from them. Across the three evaluated distances (10 cm, 15 cm, and 20 cm), the results indicate that measurements remain reasonably stable, particularly at the feature (ΔE) level.

For the ΔE feature, ICC(2,1) and ICC(3,1) values ranged from approximately 0.61 to 0.83 across biomarkers, corresponding to moderate to good reliability. We also observe that for both ICC(2,1) and ICC(3,1), the p-values were below 0.05, indicating that the results are statistically significant rather than due to chance. The similarity between ICC(2,1) and ICC(3,1) values suggests that whether we generalize to a broader range of distances or focus specifically on the 10 to 20 cm range used in this experiment, the reliability of the measurements remains consistent. This observation, along with the p-values, is seen for both the ΔE and predicted concentration values. This further validates that the Processing and Alignment procedure can effectively reduce the variability introduced by moderate changes in capture distance.

Among biomarkers for ΔE values, CRP showed the highest reliability, while IL-6 exhibited the lowest, although still within an acceptable range for practical use. For predicted concentration values, IL-1beta showed the highest reliability, while IL-16 showed the lowest. This observation is less meaningful, as it could easily reflect bias introduced by the small dataset and does not necessarily indicate that one biomarker is measured more reliably than another. Future work with a much larger dataset could investigate whether certain biomarkers produce more reliable results.

As expected, and consistent with the previous experiment (Section 4.2), predicted concentration values exhibited lower ICC values for IL-1beta and IL-16. Interestingly, this effect was not observed for CRP. Overall concentration predictions are varied from poor (IL-6 at 0.47)

to good (CRP at 0.85) reliability. This varied reliability can be due to small variations in ΔE that can produce disproportionately larger changes once passed through the predictive model, especially given that the model is trained on a relatively small calibration set. Despite this, concentration estimates remained stable enough to indicate overall moderate reliability, suggesting that the model does not become dramatically affected by distance-related differences.

Focusing specifically on the capture distances of 10 cm versus 15 cm, we aim to observe whether these more typical capture conditions produce statistically reliable measurements. The Wilcoxon signed-rank tests between the two distances showed no statistically significant differences for either ΔE or predicted concentrations across all biomarkers (all p-values greater than 0.05). This further supports that, within a reasonable capture range, if the user moves the phone a few centimeters closer or farther, this does not significantly affect the resulting measurements. Although precise positioning was not enforced in this study, this suggests that a future feature addition to the mobile application to encourage a capture distance of 10 to 15 cm may be advantageous to guarantee an increase in capture reliability.

Overall, this experiment demonstrates that moderate variability in capture distance does not meaningfully affect measurement quality, validating both the robustness of the pipeline and its suitability for non-controlled environments such as POCT.

4.5 Evaluating Participant Data

This experiment evaluates the performance of our pipeline measurements by comparing them to those measurements obtained from an Enzyme-Linked Immunosorbent Assay (ELISA) kit using data collected from human participants. In addition to direct comparison with ELISA results, we assess the ability of the pipeline to discriminate between patients and healthy controls using standard metrics from binary classification, namely the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves.

4.5.1 Experimental Setup

The study included eight participants in total. This cohort was divided into two equal groups:

- Four individuals served as healthy controls. These participants were considered blind healthy controls, meaning that they had not undergone any prior medical testing for the biomarkers of interest, but were generally healthy.

- Four individuals were designated as patients exhibiting a condition associated with elevated systemic inflammation.

Table 4.9 shows the participant information for both the healthy and patient cohorts enrolled in this study, including their sex, age, ethnicity, and type of disease for patients.

Table 4.9: Demographic information for the healthy and patient cohorts.

Patients				
ID	Sex	Age	Ethnicity	Type of disease
P1	M	21	European	Psoriasis
P2	F	40	European	Severe atopic dermatitis (eczema)
P3	F	22	West Asian	Psoriasis
P4	F	39	Middle Eastern	Moderate atopic dermatitis (eczema)
Healthy				
ID	Sex	Age	Ethnicity	Type of disease
H1	M	32	Middle Eastern	N/A
H2	M	26	European	N/A
H3	M	26	European	N/A
H4	M	32	Middle Eastern	N/A

To induce sweating, participants were instructed to run on a treadmill while wearing the patch. Immediately following the exercise, images of the patch were captured using a Samsung Galaxy S24 smartphone. To ensure the consistency of the data, all images were taken within the same laboratory setting. Furthermore, additional sweat samples were collected concurrently from all participants. These samples were analyzed using an enzyme-linked immunosorbent assay (ELISA) kit to measure the concentrations in picograms per milliliter (pg/mL). ELISA is a widely used laboratory technique that detects and quantifies specific proteins or molecules. The measurements obtained from the ELISA kits serve as the benchmark for comparison with the results from our pipeline.

On top of comparing against ELISA results, we evaluate the discriminative performance of the pipeline using two metrics from binary classification, the Receiver Operating Characteristic (ROC) and the Precision-Recall (PR) curves. Since ROC and PR metrics require binary labels, each participant was assigned a label based on their health status. Patients were labeled as $y_i = 1$ and healthy controls as $y_i = 0$, i.e.,

$$y_{\text{true}} = \{\text{P1:1, P2:1, P3:1, P4:1, H1:0, H2:0, H3:0, H4:0}\}.$$

For each biomarker measurement x_i, we compute the area under the ROC curve (ROC AUC) by taking the true positive rate (TPR) against the false positive rate (FPR) at varying thresholds and calculating the area under the ROC curve:

$$\text{ROC AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{4.5}$$

which quantifies the ability of the pipeline to correctly discriminate between patients and healthy controls. An ROC AUC of 1 indicates perfect discrimination, while an AUC of 0.5 indicates performance equivalent to random guessing.

Similarly, the PR curve plots precision versus recall across different thresholds, where precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4.6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4.7}$$

with TP, FP, and FN representing true positives, false positives, and false negatives, respectively. The area under the PR curve, known as the average precision (AP), is calculated as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$\text{AP} = \sum_k (R_k - R_{k-1}) P_k, \tag{4.8}$$

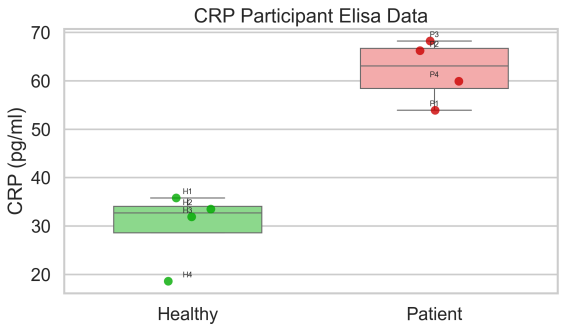
where P_k and R_k are the precision and recall at the k-th threshold. This metric summarizes the trade-off between precision and recall across all thresholds. An AP of 1 indicates that all patients (y=1) are ranked above all healthy controls (y=0), meaning perfect precision and recall at all thresholds. Lower AP values indicate that some healthy controls are ranked higher than patients, reflecting a decrease in the discriminative ability of the measurement.

4.5.2 Results

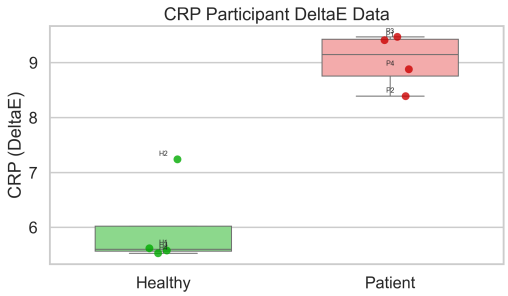
We compare the results obtained from the ELISA assays with our extracted feature ΔE and predicted concentrations. Figures 4.8–4.10 show box plots comparing the ELISA values, the extracted ΔE values, and the predicted concentrations for CRP, IL-1beta, and IL-6 biomarkers, respectively. Each box plot displays the Healthy and Patient groups with indi-

vidual participant data points overlaid. The box represents the interquartile range, spanning from the 25th to the 75th percentile, with the line inside the box indicating the median. The whiskers extend to show the overall range of the data, excluding extreme outliers. Each participant is additionally shown as a dot labeled with their participant ID. Table 4.10 reports the exact numerical values corresponding to these measurements for all participants and biomarkers.

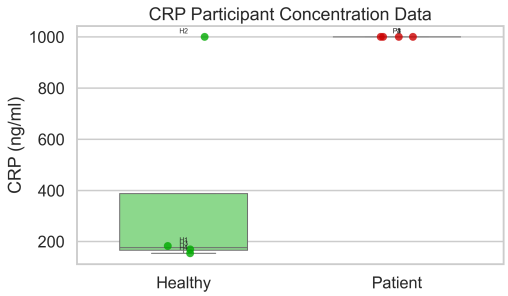
Table 4.11 summarizes the discriminative performance of the pipeline for each biomarker and measurement type. These metrics were computed for the ELISA measurements, the extracted ΔE values, and the predicted concentrations. The table reports the ROC AUC and the AP for each biomarker and measure.



(a) ELISA measurements

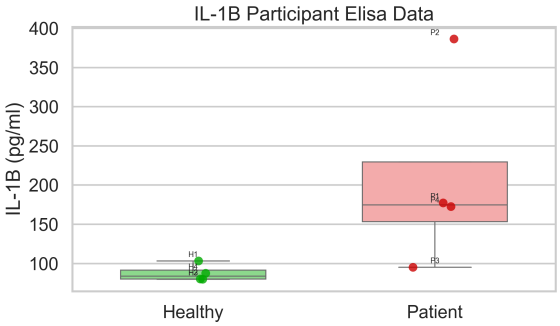


(b) ΔE measurements

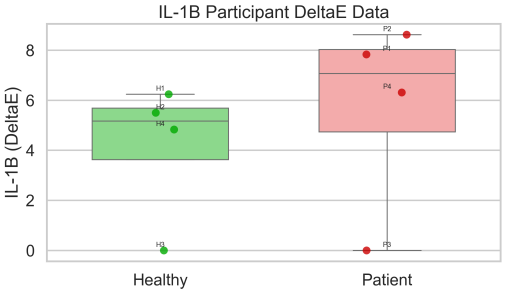


(c) Concentration measurements

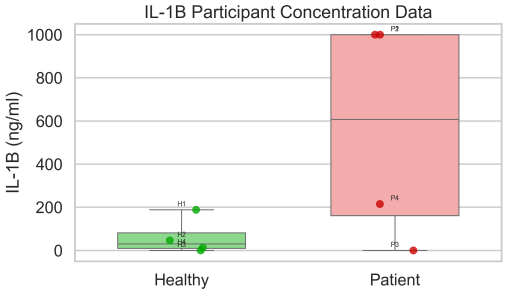
Figure 4.8: CRP measurements across participants. Top: ELISA measurements. Bottom: ΔE (left) and concentration measurements (right) for the same participants.



(a) ELISA measurements

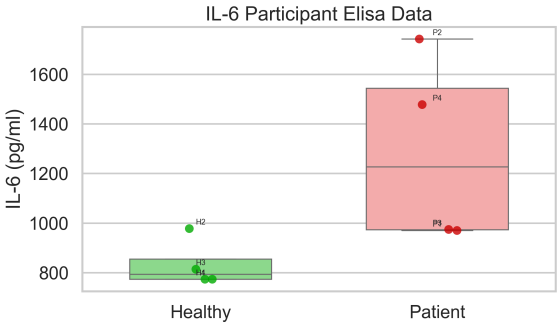


(b) ΔE measurements

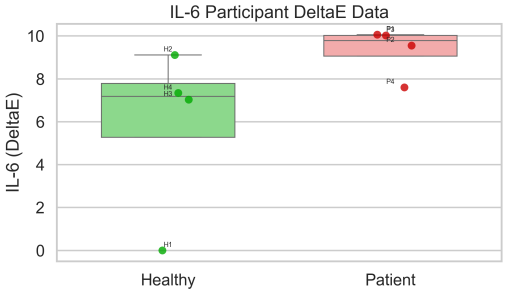


(c) Concentration measurements

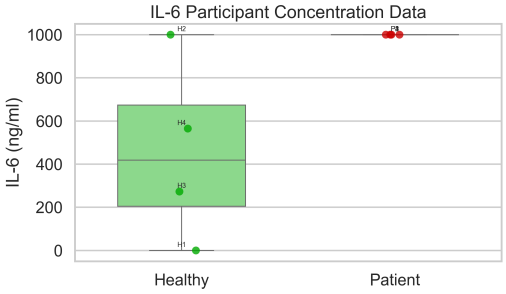
Figure 4.9: IL-1beta measurements across participants. Top: ELISA measurements. Bottom: ΔE (left) and concentration measurements (right).



(a) ELISA measurements



(b) ΔE measurements



(c) Concentration measurements

Figure 4.10: IL-6 measurements across participants. Top: ELISA measurements. Bottom: ΔE (left) and concentration measurements (right).

Table 4.10: Participant ELISA, ΔE, and predicted concentration data values for CRP, IL-1beta, and IL-6.

ParticipantID	Healthy	Biomarker	Elisa	ΔE	Concentration
P1	No	CRP	53.9	9.41	> 1000
P2	No	CRP	66.22	8.39	> 1000
P3	No	CRP	68.22	9.47	> 1000
P4	No	CRP	59.9	8.88	> 1000
P1	No	IL-1beta	177.24	7.83	> 1000
P2	No	IL-1beta	386.4	8.62	> 1000
P3	No	IL-1beta	95.1	0.00	0.00
P4	No	IL-1beta	172.6	6.31	215.1
P1	No	IL-6	970.45	10.06	> 1000
P2	No	IL-6	1742.21	9.55	> 1000
P3	No	IL-6	974.6	10.02	> 1000
P4	No	IL-6	1477.9	7.60	> 1000
H1	Yes	CRP	35.8	5.62	182.91
H2	Yes	CRP	33.5	7.24	> 1000
H3	Yes	CRP	31.9	5.58	169.66
H4	Yes	CRP	18.6	5.53	154.45
H1	Yes	IL-1beta	103.2	6.24	188.46
H2	Yes	IL-1beta	80.2	5.50	46.58
H3	Yes	IL-1beta	79.9	0.00	0.00
H4	Yes	IL-1beta	87.5	4.83	13.14
H1	Yes	IL-6	774.0	0.00	0.00
H2	Yes	IL-6	978.1	9.11	> 1000
H3	Yes	IL-6	814.3	7.03	272.94
H4	Yes	IL-6	774.0	7.34	564.82

Table 4.11: ROC AUC and Average Precision (AP) for each biomarker and measurement type. Metrics were computed for ELISA, ΔE, and predicted concentrations.

Biomarker	Measure	ROC AUC	AP
CRP	DeltaE	1.0	1.0
IL-1beta	DeltaE	0.78125	0.875
IL-6	DeltaE	0.9375	0.95
CRP	Concentration	0.875	0.8
IL-1beta	Concentration	0.78125	0.875
IL-6	Concentration	0.875	0.8
CRP	Elisa	1.0	1.0
IL-1beta	Elisa	0.9375	0.95
IL-6	Elisa	0.875	0.8875

4.5.3 Discussion

We first compare our ΔE and predicted concentration measurements for each participant and biomarker with the ELISA measurements, as presented in Table 4.10 and shown in Figures 4.8–4.10.

For CRP, the pipeline successfully distinguished the Patient cohort from the Healthy cohort in most cases, with the exception of participant H2. In the Patient group, P3 recorded the highest ELISA concentration (68.22 pg/mL) and correspondingly the highest ΔE (9.47). However, the inner group ranking was not fully preserved. For instance, P2 showed the second-highest ELISA value (66.22 pg/mL) but the lowest ΔE (8.39) among patients. Within the Healthy cohort, H2 presented a notable outlier, exhibiting a high ΔE (7.24) and a predicted concentration ($>$ 1000 pg/mL), despite a moderate ELISA reading (33.5 pg/mL). Conversely, H4 correctly displayed the lowest values across all metrics (ELISA: 18.6 pg/mL; ΔE: 5.53).

The IL-1beta results demonstrated the highest consistency between the ELISA method and our pipeline. The ranking of patient severity was perfectly preserved, with P2 exhibiting the highest values (ELISA 386.4 pg/mL, ΔE 8.62), followed by P1, P4, and finally P3, who showed minimal levels in both ELISA (95.1 pg/mL) and ΔE (0.00). This strong correlation extended to the Healthy cohort, where values remained low. H3, for example, recorded the lowest ELISA (79.9 pg/mL) and a ΔE of 0.00. While there was a minor ranking swap between H2 and H4 in the pipeline compared to ELISA, the general trend in Patients versus

Healthy controls was accurately captured.

For IL-6, the pipeline effectively identified elevated inflammation in patients but struggled to rank them according to ELISA results. While all Patients (P1–P4) showed significantly higher ΔE values (7.60–10.06) compared to the Healthy controls (0.00–9.11), the relative ordering was inconsistent. P2 had the highest ELISA concentration (1742.21 pg/mL) but the third-highest ΔE (9.55). Conversely, P1 had the lowest ELISA among patients (970.45 pg/mL) but the highest ΔE (10.06). Among the healthy controls, H2 again appeared as a potential outlier with a high ΔE (9.11) and predicted concentration, consistent with their CRP results.

While not always consistent with ELISA in preserving the exact inner group ranking, the ΔE measurements and predicted concentrations are able to reproduce the group-level separation between Healthy and Patient participants. The predicted concentrations at > 1000 ng/ml for several participants suggests the pipeline is highly sensitive to elevated biomarkers.

Moving our attention to the ROC AUC and AP metrics presented in Table 4.11, we see that both the extracted ΔE values and the predicted concentrations effectively discriminate between Patient and Healthy participants, performing comparably to the ELISA measurements. For CRP, ΔE achieves perfect separation with a ROC AUC of 1.0 and an AP of 1.0, while the predicted concentrations slightly underperform with a ROC AUC of 0.875 and an AP of 0.8, indicating that the model predictions introduce minor variability as expected. For IL-1beta, both ΔE and predicted concentrations yield a ROC AUC of 0.78125 and an AP of 0.875, slightly below the ELISA results (ROC AUC 0.9375, AP 0.95), though not substantially. For IL-6, however, the ΔE feature achieves a higher ROC AUC (0.9375) and AP (0.95) than ELISA (ROC AUC 0.875, AP 0.8875), suggesting that the extracted ΔE was more effective in capturing elevated IL-6 levels and distinguishing patients from healthy controls than ELISA. The predicted concentrations also provide strong discrimination, though slightly below both ΔE and ELISA. Altogether, these results demonstrate that both the extracted ΔE values and the predicted concentrations offer very close performance to the ELISA measurements and are expected to improve further with better calibration, model tuning, and future enhancements to the pipeline.

Chapter 5

Conclusion

5.1 Summary

This thesis presented a complete end-to-end pipeline for the remote, non-invasive screening of three inflammatory biomarkers (CRP, IL-1beta, and IL-6) using a wearable sweat patch and smartphone-based image processing. The primary motivation for this work was to bridge the gap between complex biochemical monitoring and accessible POCT. While wearable devices for measuring biomarkers from sweat exist, reliable quantification in uncontrolled, non-laboratory environments remains a significant challenge due to variations such as lighting and camera hardware.

To address these challenges, we introduced a novel algorithmic pipeline integrated with a custom-designed calibration layer. The calibration layer, featuring five ArUco markers and 49 colour reference swatches, enabled the pipeline to autonomously standardized the patch geometry and normalize colourimetric data to a consistent visual domain. We developed a lightweight U-Net segmentation model to precisely isolate the SCRs and introduced a feature extraction method that utilizes the CIELAB colour space to compute ΔE, a metric desgined to overcome environmental and camera variability.

Experimental evaluations demonstrated the efficacy of this approach. The two-step image normalization process reduced measurement variability by approximately 70% across different lighting conditions. The pipeline proved robust across different smartphone devices and capture distances, maintaining high reliability for the extracted ΔE feature. Furthermore, when evaluated against the standard ELISA method using human participant data, our pipeline achieved comparable results. Specifically, the ΔE feature demonstrated excellent discriminative ability, achieving an ROC AUC of 1.0 for CRP and 0.9375 for IL-6, proving

that smartphone-based colourimetry can effectively screen for elevated inflammation levels.

5.2 Limitations

This study faced several limitations that impacted specific aspects of the pipeline's performance and generalizability.

First, the physical manufacturing of the calibration layer introduced noise into the data. The layer was fabricated using a laser cutter to create the cutouts for the SCRs. This process frequently resulted in burn lines along the edges of the chamber openings (Can refer back to Figure 3.11). To mitigate the risk of these artifacts influencing the colourimetric analysis, the segmentation mask for each SCR had to be reduced by 20%. While this successfully removed most the burn lines, it also discarded a significant portion of potentially valid SCR data, effectively reducing the sample size of pixels available for feature extraction.

Second, the dataset used to generate the calibration curves was limited in size, consisting of only 16 samples across four concentration levels (1, 10, 100, and 1000 ng/mL). While this was sufficient to establish a logarithmic relationship between ΔE and concentration, the small sample size restricted the predictive model's ability to generalize. This limitation was quantitatively observed in our experimental results. For example in the device comparison experiment, we noticed that the ICC for predicted concentrations was notably lower than that of the raw ΔE feature. This suggests that the current mathematical model for concentration prediction is less robust than the underlying feature extraction method.

Finally, the calibration layer itself was produced using a standard consumer-grade colour printer. Consequently, the "observed" colours of the 49 colour swatches were subject to printer bias. Unlike professional photographic colour cards (e.g., Macbeth ColorCheckers) which are manufactured with high precision and consistency under different lighting conditions, standard printed ink colours may appear less consistent and different under varying light sources. This printer bias introduces a baseline error in the colour normalization process that cannot be fully corrected algorithmically.

5.3 Future Work

To advance this technology from a research prototype to a deployable medical tool, several areas of future work should be prioritised.

A primary objective is to improve the robustness of the concentration prediction model by expanding the calibration dataset. Future data collection should involve a significantly larger number of samples at a wider range of concentration intervals (e.g., adding 50, 250, and great than 1000 ng/mL set points). A larger, high-density dataset would allow for the training of more complex regression models or machine learning regressors, which would improve the consistency of quantitative predictions.

On the software side, the mobile application can be enhanced to actively assist the user in capturing high-quality data. Our experiments showed that capture distances between 10 cm and 15 cm yielded the most consistent results. Future iterations of the application could implement Augmented Reality (AR) guides or real-time distance estimation to strictly enforce this optimal capture range, rejecting images taken from too far away or too close. Additionally, further testing should be conducted on a wider array of smartphone devices and under more extreme lighting environments to ensure universal accessibility.

The segmentation model also presents an opportunity for optimization. While the current MobileNetV3-based U-Net is efficient, future work could explore newer, even more lightweight architectures to further improve robustness of segmentation predictions.

Finally, the physical limitations of the calibration layer should be addressed through professional manufacturing, if possible. Future iterations of the layer should be produced using high-fidelity printing processes similar to those used for professional photography colour cards. Eliminating printer bias and laser-cut burn lines would allow for the utilization of the full SCR surface area and ensure that the reference colours remain accurate and stable.

Bibliography

[1] E. Monaghesh and A. Hajizadeh, "The role of telehealth during covid-19 outbreak: a systematic review based on current evidence," *BMC public health*, vol. 20, no. 1, p. 1193, 2020.

[2] A. Ahmed, M. Mutahar, A. A. Daghrery, N. H. Albar, I. Q. I. Alhadidi, A. M. Asiri, N. Boreak, A. A. S. Alshahrani, M. Shariff, M. A. Shubayr, *et al.*, "A systematic review of publications on perceptions and management of chronic medical conditions using telemedicine remote consultations by primary healthcare professionals april 2020 to december 2021 during the covid-19 pandemic," *Medical science monitor: international medical journal of experimental and clinical research*, vol. 30, pp. e943383–1, 2024.

[3] P. B. Luppa, C. Müller, A. Schlichtiger, and H. Schlebusch, "Point-of-care testing (poct): Current techniques and future perspectives," *TrAC Trends in Analytical Chemistry*, vol. 30, no. 6, pp. 887–898, 2011.

[4] A. Larsson, R. Greig-Pylypczuk, and A. Huisman, "The state of point-of-care testing: a european perspective," *Upsala journal of medical sciences*, vol. 120, no. 1, pp. 1–10, 2015.

[5] D. Prakashan, R. PR, and S. Gandhi, "A systematic review on the advanced techniques of wearable point-of-care devices and their futuristic applications," *Diagnostics*, vol. 13, no. 5, p. 916, 2023.

[6] T. Kamei, T. Kanamori, Y. Yamamoto, and S. Edirippulige, "The use of wearable devices in chronic disease management to enhance adherence and improve telehealth outcomes: a systematic review and meta-analysis," *Journal of telemedicine and telecare*, vol. 28, no. 5, pp. 342–359, 2022.

[7] D. Fuller, E. Colwell, J. Low, K. Orychock, M. A. Tobin, B. Simango, R. Buote, D. Van Heerden, H. Luan, K. Cullen, *et al.*, "Reliability and validity of commercially

available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review," *JMIR mHealth and uHealth*, vol. 8, no. 9, p. e18694, 2020.

[8] D.-G. Macovei, M.-B. Irimes, O. Hosu, C. Cristea, and M. Tertis, "Point-of-care electrochemical testing of biomarkers involved in inflammatory and inflammatory-associated medical conditions," *Analytical and Bioanalytical Chemistry*, vol. 415, no. 6, pp. 1033–1063, 2023.

[9] J. Das, U. Bhui, S. Chowdhary, S. Sarkar, I. K. Ghoshal, S. Nayak, R. Bishayee, N. Khurana, B. Kumar, and N. Sharma, "Biomarkers unveiling neurodegeneration: Keys to progression and therapeutic insights," *Indian Journal of Pharmaceutical Education and Research*, vol. 59, no. 1s, pp. s1–s15, 2025.

[10] R. D. Kehm, J. A. McDonald, S. E. Fenton, M. Kavanaugh-Lynch, K. A. Leung, K. E. McKenzie, J. S. Mandelblatt, and M. B. Terry, "Inflammatory biomarkers and breast cancer risk: a systematic review of the evidence and future potential for intervention research," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5445, 2020.

[11] F. Criscuolo, I. N. Hanitra, S. Aiassa, I. Taurino, N. Oliva, S. Carrara, and G. De Micheli, "Wearable multifunctional sweat-sensing system for efficient healthcare monitoring," *Sensors and Actuators B: Chemical*, vol. 328, p. 129017, 2021.

[12] B. Jagannath, K.-C. Lin, M. Pali, D. Sankhala, S. Muthukumar, and S. Prasad, "A sweat-based wearable enabling technology for real-time monitoring of il-1β and crp as potential markers for inflammatory bowel disease," *Inflammatory Bowel Diseases*, vol. 26, no. 10, pp. 1533–1542, 2020.

[13] Q. Hua and G. Shen, "A wearable sweat patch for non-invasive and wireless monitoring inflammatory status," *Journal of Semiconductors*, vol. 44, no. 10, p. 100401, 2023.

[14] T. Wu and G. Liu, "Non-invasive wearables in inflammation monitoring: From biomarkers to biosensors," *Biosensors*, vol. 15, no. 6, p. 351, 2025.

[15] S. R. S. Pour, D. Calabria, A. Emamiamin, E. Lazzarini, A. Pace, M. Guardigli, M. Zangheri, and M. Mirasoli, "Electrochemical vs. optical biosensors for point-of-care applications: A critical review," *Chemosensors*, vol. 11, no. 10, p. 546, 2023.

[16] V. Blahnik and O. Schindelbeck, "Smartphone imaging technology and its applications," *Advanced Optical Technologies*, vol. 10, no. 3, pp. 145–232, 2021.

[17] J. Chen, D. Zhao, H.-W. Shi, Q. Duan, P. Jajesniak, Y. Li, W. Shen, J. Zhang, J. Reboud, J. M. Cooper, *et al.*, "Inclusive and accurate clinical diagnostics using intelligent computation and smartphone imaging," *ACS sensors*, vol. 9, no. 10, pp. 5342–5353, 2024.

[18] H. Nejati, V. Pomponiu, T.-T. Do, Y. Zhou, S. Iravani, and N.-M. Cheung, "Smartphone and mobile image processing for assisted living: Health-monitoring apps powered by advanced mobile imaging algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 30–48, 2016.

[19] S. Tonello, G. Abate, M. Borghetti, N. F. Lopomo, M. Serpelloni, and E. Sardini, "How to assess the measurement performance of mobile/wearable point-of-care testing devices? a systematic review addressing sweat analysis," *Electronics*, vol. 11, no. 5, p. 761, 2022.

[20] J. Xu, Y. Fang, and J. Chen, "Wearable biosensors for non-invasive sweat diagnostics," *Biosensors*, vol. 11, no. 8, p. 245, 2021.

[21] Y. Cheng, S. Feng, Q. Ning, T. Li, H. Xu, Q. Sun, D. Cui, and K. Wang, "Dual-signal readout paper-based wearable biosensor with a 3d origami structure for multiplexed analyte detection in sweat," *Microsystems & Nanoengineering*, vol. 9, no. 1, p. 36, 2023.

[22] A. Salek-Maghsoudi, F. Vakhshiteh, R. Torabi, S. Hassani, M. R. Ganjali, P. Norouzi, M. Hosseini, and M. Abdollahi, "Recent advances in biosensor technology in assessment of early diabetes biomarkers," *Biosensors and Bioelectronics*, vol. 99, pp. 122–135, 2018.

[23] J. Choi, A. Bandodkar, J. Reeder, T. Ray, A. Turnquist, S. Kim, N. Nyberg, A. Hourlier-Fargette, J. Model, A. Aranyosi, *et al.*, "Soft, skin-integrated multifunctional microfluidic systems for accurate colorimetric analysis of sweat biomarkers and temperature, acs sens. 4 (2019) 379–388."

[24] A. Koh, D. Kang, Y. Xue, S. Lee, R. M. Pielak, J. Kim, T. Hwang, S. Min, A. Banks, P. Bastien, *et al.*, "A soft, wearable microfluidic device for the capture, storage, and colorimetric sensing of sweat," *Science translational medicine*, vol. 8, no. 366, pp. 366ra165–366ra165, 2016.

[25] A. J. Bandodkar, P. Gutruf, J. Choi, K. Lee, Y. Sekine, J. T. Reeder, W. J. Jeang, A. J. Aranyosi, S. P. Lee, J. B. Model, *et al.*, "Battery-free, skin-interfaced microfluidic/electronic systems for simultaneous electrochemical, colorimetric, and volumetric analysis of sweat," *Science advances*, vol. 5, no. 1, p. eaav3294, 2019.

[26] Y. He, L. Wei, W. Xu, H. Wu, and A. Liu, "Laser-cutted epidermal microfluidic patch with capillary bursting valves for chronological capture, storage, and colorimetric sensing of sweat," *Biosensors*, vol. 13, no. 3, p. 372, 2023.

[27] E. Celikbas, A. E. Ceylan, and S. Timur, "based colorimetric spot test utilizing smartphone sensing for detection of biomarkers," *Talanta*, vol. 208, p. 120446, 2020.

[28] A. Horta-Velázquez, G. Ramos-Ortiz, and E. Morales-Narváez, "The optimal color space enables advantageous smartphone-based colorimetric sensing," *Biosensors and Bioelectronics*, vol. 273, p. 117089, 2025.

[29] L. Shen, J. A. Hagen, and I. Papautsky, "Point-of-care colorimetric detection with a smartphone," *Lab on a Chip*, vol. 12, no. 21, pp. 4240–4243, 2012.

[30] C. S. McCamy, H. Marcus, J. G. Davidson, *et al.*, "A colour-rendition chart," *J. App. Photog. Eng*, vol. 2, no. 3, pp. 95–99, 1976.

[31] xrite, "Colorchecker classic." https://www.xrite.com/categories/calibration-profiling/colorchecker-classic. Accessed: 2025-12-18.

[32] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Mar´in-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[33] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[34] Y. Soda and E. Bakker, "Quantification of colourimetric data for paper-based analytical devices," *ACS sensors*, vol. 4, no. 12, pp. 3093–3101, 2019.

[35] G. Zhang, Q. Chen, and Q. Sun, "Illumination normalization among multiple remote-sensing images," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, pp. 1470–1474, 09 2014.

[36] W.-C. Wu, H. Zhao, W.-W. Liu, and W. Tao, "Effects of illumination on image quality in precision vision measurement," *Shanghai Jiaotong Daxue Xuebao/Journal of Shanghai Jiaotong University*, vol. 43, pp. 931–934+939, 06 2009.

[37] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*, pp. 474–485, 1994.

[38] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[40] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[41] R. Wightman, "Pytorch image models." https://github.com/rwightman/pytorch-image-models, 2019.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[44] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[45] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[46] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability.," *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[47] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[48] R. Vallat, "Pingouin: statistics in python," *Journal of Open Source Software*, vol. 3, p. 1026, Nov. 2018.

[49] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[50] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: an open source differentiable computer vision library for pytorch," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3674–3683, 2020.

Appendices

A Image Normalization Experiment Supplementary

Table 1: Normalized ΔE values for different biomarkers and samples

Biomarker	ΔE_mean	ΔE_std	Sample
CRP	4.28	0.72	sample_0
IL-1beta	3.97	0.93	sample_0
IL-6	4.84	0.67	sample_0
CRP	3.16	0.55	sample_10
IL-1beta	1.86	0.97	sample_10
IL-6	5.73	0.38	sample_10
CRP	5.47	0.33	sample_50
IL-1beta	5.65	0.37	sample_50
IL-6	6.89	0.23	sample_50
CRP	5.75	0.49	sample_100
IL-1beta	5.49	0.53	sample_100
IL-6	6.79	0.28	sample_100
CRP	8.11	0.16	sample_1000
IL-1beta	8.11	0.15	sample_1000
IL-6	8.10	0.19	sample_1000

Table 2: Unnormalized ΔE values for different biomarkers and samples

Biomarker	ΔE_mean	ΔE_std	Sample
CRP	0.66	0.60	sample_0
IL-1beta	0.38	0.47	sample_0
IL-6	1.23	0.53	sample_0
CRP	0.55	0.44	sample_10
IL-1beta	0.30	0.52	sample_10
IL-6	3.28	0.66	sample_10
CRP	3.75	0.93	sample_50
IL-1beta	3.78	1.03	sample_50
IL-6	5.19	0.74	sample_50
CRP	4.02	0.67	sample_100
IL-1beta	3.65	0.64	sample_100
IL-6	5.15	0.75	sample_100
CRP	6.24	0.57	sample_1000
IL-1beta	6.28	0.58	sample_1000
IL-6	6.18	0.62	sample_1000

B Segmentation Model

B.1 Data Curation and Augmentation Implementation

We trained the segmentation model on a limited dataset of 26 manually annotated patch images. To prevent overfitting and improve the model's generalization to varying capture conditions, each image-mask pair in the training set was augmented approximately 30 times, effectively expanding the dataset to a size of 800 image-mask pairs.

Preprocessing

- **Resizing:** All input images and masks were resized to fixed spatial dimensions of 320×320 pixels.
- **Normalization:** Input images were normalized using standard ImageNet statistics (these were also applied at inference):
  - Mean: $[0.485, 0.456, 0.406]$
  - Standard Deviation: $[0.229, 0.224, 0.225]$

- **Mask Binarization:** Masks were converted to binary tensors using a threshold of 0.5.

Augmentation Pipeline We employed a sequential augmentation pipeline using the Kornia library [50] to apply both photometric and geometric transformations.

1. **Photometric Distortions (Applied to Image Only):**
    - **Colour Jitter ($p = 0.5$):** Brightness (0.2), Contrast (0.2), Saturation (0.2), Hue (0.1).
    - **Random Contrast ($p = 0.3$):** Contrast factor range $[0.7, 1.3]$.
    - **Random Sharpness ($p = 0.3$):** Sharpness factor 0.2.
    - **Random Grayscale ($p = 0.2$):** Converted to grayscale to reduce colour dependency.
    - **Gaussian Noise ($p = 0.2$):** Mean 0.0, Std 0.01.
    - **Solarize ($p = 0.2$):** Threshold 0.5.
2. **Geometric Transformations (Shared by Image and Mask):**
    - **Random Rotation:** ± 30.0 degrees.
    - **Flips:** Random horizontal and vertical flips.
    - **Gaussian Blur:** Kernel size $(3, 3)$, sigma range $(0.1, 3.0)$.
    - **Crops ($p = 0.8$ total):** Random crop ($p = 0.5$) or Center crop ($p = 0.3$) with a reduction of 20 pixels, followed by resizing back to 320×320 to simulate zoom and scale variations.

B.2 Cross-Validation Results

The model performance was evaluated using a 5-fold cross-validation strategy. Table 3 details the Dice coefficient scores obtained for each fold. The model demonstrated high stability across folds, achieving an average Dice score of 0.9164 with a low standard deviation of 0.0030.

Table 3: 5-Fold Cross-Validation Dice Scores for the Segmentation Model

Fold	**Dice Score**
Fold 1	0.9206
Fold 2	0.9116
Fold 3	0.9155
Fold 4	0.9178
Fold 5	0.9163
Average	**0.9164**
Std Dev	**0.0030**