©Copyright 2022
Chunjong Park

Approaches for Improving Data Acquisition in Sensor-Based mHealth Applications

Chunjong Park

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Shwetak Patel, Chair
James Fogarty
Tadayoshi Kohno

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Approaches for Improving Data Acquisition in Sensor-Based mHealth Applications

Chunjong Park

Chair of the Supervisory Committee:
Washington Research Foundation Entrepreneurship Endowed Professor Shwetak Patel
Paul G. Allen School of Computer Science & Engineering

Mobile health (mHealth) applications enable people with little or no clinical experience to measure vital signs and screen for different health conditions with mobile devices. While such applications provide accessible health to the general population, recent mHealth applications leverage diverse sensors, require complex data acquisition procedures, and rely on complicated, black-box algorithms. Users are often uncertain about the quality of acquired data and resulting health-related predictions. They are also exposed to silent failures that lead to inaccurate results that could impact their medical decisions.

To bridge the gap between underlying algorithms and *non-expert* users, this thesis explores approaches to design feedback for *non-expert* users to ensure that the screening algorithms only process high-quality data to provide accurate, reliable, and trustworthy results. Different feedback strategies should be applied at different stages of mHealth applications usage. During data acquisition, real-time, sensor-driven feedback is explored through two projects: (1) RDTScan, a smartphone-based rapid diagnostic test (RDT) reader, which employs real-time image quality assurance and guidance to capture high-quality RDT images, and (2) CapApp, a smartphone-based capillary refill time (CRT) assessment, which uses sensor-driven feedback to guide users to apply and release pressure for obtaining high-quality camera-based finger PPG signals. After data acquisition, the estimated uncertainty can inform users whether health algorithms can accurately process and interpret the acquired data. I explored

leveraging state-of-the-art out-of-distribution detection methods for health ML models to improve the reliability of the results and user trust. By providing such information, users can only receive reliable predictions from the algorithms and only trust the diagnostic results with low uncertainty. This thesis provides a design space of feedback strategies to help *non-expert* users receive accurate and reliable health predictions from mHealth applications through these projects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I want to thank my amazing partner. I am infinitely grateful for her unconditional love and support. I could not imagine I could reach where I am without her. We have gone through so many things together during my PhD and I am looking forward to our next adventure. I also would like to thank my parents for constant caring and support. Even though we are physically apart, I always feel like you are right next to me. And, I also thank my brother, Jae. Having you in Seattle meant so much for me.

My sincere thanks goes to my advisor, Shwetak. You have been a perfect advisor, mentor, and role model for me throughout my PhD. You also have created such a safe environment where I can talk to you anything about my research, career, and personal matters. I also learned so much from how you advise, mentor me and lead the lab. When I am in a leadership role in the future, I would always look back and think about how you mentored me and try to do the same things for my team.

Alex, no words can explain how much grateful I am for what you have done for me. You have been my friend, senior student, co-authors, PhD thesis committee, and unofficial co-advisor. Fun memories working with you and hanging out with you. Thanks for help me adjusting to new environment (e.g., USA, UW, Ubicomp lab), going through research projects and PhD, and on thousands of small little things I cannot even list them all out.

I would like to thank previous and past Ubicomp lab members. Joining the lab is the "second" best decision I made in my life. Thanks everyone for creating such a supportive and friendly environment. When it comes to research, I love how everyone becomes so resourceful, creative, and critical. I am so proud of being a part of the lab. And, I hope this culture can last forever. Eric, I still remember you were my visit days scheduler and how you helped

me and showed me around the lab and lab member had so much impact on my decision to join UW and Ubicomp lab. Although we have not worked together when you were around, I've learned so much from your enthusiam and standard for high-quality research projects. Edward, there was so much to learn from how you execute research projects, present your work, and mentor students. Morelle, thanks for being my "mentor" when I first came. I have received so much moral support from you. And, it was really fun working with you on the sleep project. Farshid, I admire your energy and passion for everything. Everytime I talk to you, I feel like I am getting so much joy and energy from you. Manuja, we came in the same year and went through so many things together. Thanks for all the chit-chat and listening to my complaints about my PhD life. Libby, you are my lifesaver. I could not have made any progress on my research if you weren't here. Thanks for taking care of all the project management. Xin, Matt, Richard, Shirley, Alvin, Ishan, Jason, Joe, Girish, and Anand, you guys are already becoming superstars in our field. Thanks for the moral support. I hope we cross path in the future and see each other in upcoming Ubicomp gathering. Hanchaun, Lilian, Josh, Mohit, Ruth, Elliot, Keyu, Gabe, and Sidhant, I am really grateful for being a part of Ubicomp lab and having a kind, caring lab alumni. You guys always have been willing to help me in research, internship, and job search. Thanks for responding to my random emails. Your help always led me to the right direction.

I also want to thank friends in UW CSE. Tony, so many good memories in my first year of PhD. It would be 10 times more difficult if you were not around. Thanks for the friendship and best of luck to your successful career. Ravi, it is so weird not to put you as the Ubicomp lab member. It is always fun talking to you and hearing silly jokes. I am so glad you got what you want for your job. Your advice in every small things always helped me go through difficulties in my PhD. Younghoon, Junha, and Keunhong, our non-free lunch on the ave was always fun. Weekly badminton and golf was fun as well. Thanks for all good memories.

I want to thank all my undergraduate students: Eric, Anas, Hung, Hugh, Devsh, and

Sixuan. Thanks for your fresh perspectives and contribution to the research projects.

I also would like to acknowledge my research collaborators. Tim, our sleep paper had gone through the one of the wildest rejection experiences. It was worth going through at least once during PhD. I learned so much from your advice on technical details and presentation of the work. Thanks for your help and efforts to finally make it work at the end. Yoshi, I learned what it means to collaborate with researchers in different field. The synergy of diverse perspective was definitely critical to the success of our NeurIPS paper. Thanks for your contribution in our project and for always being positive. Matthew, Barry, and Jon, thanks for your valuable feedback from clinical and bioengineering perspectives. Mobile health projects won't be possible if you all are not open-minded about innovations in healthcare.

I want to thank all my industry, NGO, and university partners who all came together to make rapid diagnostic tests more usable using smartphones: Arunan, Isaac, Ari, Peter, Anuraj, Bill and Melinda Gates Foundation, Ona, Medic Mobile, Keyna Wellcome Trust, University of Adelaide, and Audere.

Lastly, I want to thank all my internship managers and mentors who have provided me opportunities to explore diverse research areas: Fahim, Chulhong, Sourav, and Gil from Nokia Bell Labs, Andrés from Snap Research, Ken, Michel, Teddy, Nic, Danial, Miah, Becky, and Tiffany from Microsoft Research. The internships during my PhD had so much impact on my career path. Positive experiences in my previous internships made me pursue industry research positions after my PhD. And, my mentors created such a great environment and provided me with so much research feedback and advice on career path. Thanks for the opportunities and hope we cross path in the future.

Chapter 1

INTRODUCTION

1.1 Overview

Mobile health (mHealth) uses mobile devices, such as smartphones, wearables, and other IoT devices, to support medical practice and public health, usually for data collection and patient information management. With the increasing adoption of smartphones and wearable devices, more mHealth applications are designed and available to ordinary smartphone users. These applications help the users to track symptoms, goals, food intake, and medications by getting responses from the users. Healthcare providers deliver and notify the users with the right information to manage their health conditions.

As smartphones and mobile devices are equipped with better sensors (e.g., camera, microphone, and inertia sensor) and new sensors (e.g., PPG, SpO2, and ECG sensors), more mHealth applications are capable of sensing new vital signs and symptoms. Moreover, their significant improvement in computing power and energy consumption enables them to run AI-powered medical diagnostic algorithms and provide medical screening results instantly. For instance, the Apple Watch can detect atrial fibrillation using ECG sensors. Google has released applications that can measure respiratory and heart rate using cameras. These applications can be used for continuously monitoring pre-existing conditions and early screening for new health conditions. They would potentially improve access to healthcare for those who are uninsured or have limited access due to time, geographical, and socioeconomic constraints.

Emerging mHealth applications with sensing capabilities usually receive input data acquired by the users with little or no clinical experience. For example, images or audio captured by such users can be used to diagnose skin cancer or pulmonary diseases. While the algorithms expect input data with a certain level of quality and criteria, the data acquired by

the users can vary due to their limited understanding of the expected data or environmental factors (e.g., ambient lighting, background noise, device heterogeneity). With low-quality data, the screening algorithms are likely to provide inaccurate diagnostic results that can lead the users to make wrong medical decisions; the repercussion can be fatal as medical conditions could be left untreated (false negative) or medical resources can be allocated to healthy individuals (false positive).

There exists a disconnect between the users and the mHealth application for (1) the quality of data expected from algorithms and data acquired by the users, and (2) the accuracy and reliability of the results provided by algorithms and expected by the users. This gap would become wider as more mHealth applications leverage multiple sensors and deep learning algorithms for diverse health conditions. To close this gap, intelligent guidance and appropriate information should be designed for the users to use mHealth applications accurately and reliably. Specifically, this thesis explores different approaches for improving data acquisition by the users in various sensor-based mHealth applications and obtaining interpretable information on the result provided by the health algorithms.

1.2 Target mHealth Applications and Users

In recent years, diverse types of mHealth applications that target different users have been and are introduced in the market recent years. This thesis focuses on emerging mHealth applications that leverage sensors on mobile devices to extract vital signs and screen for health conditions. Specifically, the target mHealth applications require users to interact with mobile device sensors to acquire the data (e.g., image, audio, motion, and physiological signal) in both healthcare workflow and at-home usage. Such mHealth applications are available and targeted to the general mobile device user populations, where varying levels of expertise in medicine and technology. The target users in this work are *non-expert* users – mobile device owners with limited clinical experiences and technical background – who are exposed to health and safety risks posed by failures of mHealth applications. Unlike *expert* users who have prior knowledge to get reliable health-related results and discern mHealth applications

behaving unreasonably and inaccurately, *non-expert* users are exposed to inaccurate health recommendations and can make irrational medical decisions in such situations. In this section, I describe the scope of target mHealth applications and users for this work.

1.2.1 Sensor-Based mHealth Applications

Healthcare Workflow

Community healthcare workflow in low- and mid-income settings significantly benefits from mHealth applications and inexpressive diagnostic testings. Over the past couple of decades, rapid diagnostic tests (RDTs) have emerged as a potential solution to the pressing need for accessible medical testing. RDTs utilize biochemistry to transduce a load of a biological sample (E.g., blood, urine, nasal swab) to an analog colorimetric output. As such, RDTs enable point-of-care diagnostics without the need for expensive equipment. Pregnancy tests are one of the most well-known types of RDTs, but RDTs exist for many other health purposes, such as malaria [10, 132, 117], influenza [77, 164], HIV [87, 202], and COVID-19 [1]. The procedures for completing an RDT are reasonably simple, allowing community health workers (CHWs) who may not have the same medical expertise as a physician to still provide vital health services in the field.However, errors are frequently made by CHWs, which range from a misinterpretation of result lines [86, 117, 163] to not waiting for the correct duration [172]. To review and reduce the errors, mHealth applications [149, 150] are assisting CHWs to capture high-quality RDT images using a smartphone camera and provide accurate interpretation results compared to people with laboratory and clinical experiences. Captured RDT images and interpretation results are not only used for diagnosis but also for documentation and insights on disease transmission. Furthermore, to prevent users from making mistakes in RDT administration, a mHealth application [168] uses a smartphone and 3D-printed box to monitor and guide users to correctly perform the procedures.

Smartphones also have become an essential part of the clinical workflow. Patients check the

[1] `https://www.cdc.gov/coronavirus/2019-ncov/testing/self-testing-videos.html`

availability of their primary physicians, make appointments, receive prescriptions, review their previous visits, and receive information about managing and preventing health conditions. In recent years, more telehealth services are available via mobile devices (e.g., smartphones, tablets, laptops). Especially, during the COVID-19 pandemic when visiting doctors' offices is often limited, telehealth [2] has become a convenient replacement for the physical visits. A long-term vision of telehealth is to provide clinical and medical services remotely for people who have time, geographical, and socioeconomic constraints to visit clinics and hospitals. For example, routine checks and diagnoses of health conditions that do not require medical tests are feasible to be done through telehealth services. The more advanced form of telehealth services [3] support a video camera to visually check patients' status. Other services have communication channels to share image and audio data with clinicians. Sensor-based mHealth functionalities are expected to play a crucial role in collecting clinically relevant data (e.g., an image of body parts, video of limb movement, heart/lung sound, tremor motion). Furthermore, recent mHealth applications can now extract vital signs (e.g., heart rate, respiratory rate) from commodity sensors in mobile devices, providing crucial data for diagnosis and management.

At-Home Health Sensing

From the patient's end, being able to obtain core vital signs at home with mobile devices is crucial for disease management and communicating with clinicians via telehealth. Recently, numerous research works and products that sense various vital signs by a novel use of sensors in mobile devices have been released. Researchers have demonstrated various applications that perform photoplethysmography (PPG) — a technique that optically measures the blood volume changes — by recording a video of a fingertip as it covers both the camera and an illumination source (flash or screen) in a smartphone. PPG signal can be also measured remotely by detecting subtle color changes of body parts due to blood volume changes from

[2] `https://www.mychart.com/`

[3] `https://amazon.care/`

smartphone camera [108]. Using the extracted PPG signals, heart rate [67], hemoglobin [194] can be measured. A smartphone camera coupled with an accelerometer can measure blood pressure [195]. Microphones on a smartphone can perform spirometry [100], track cough counts of different members in a household [199], and detect tuberculosis [166]. Respiratory rate [11] is also measured from a smartphone camera by detecting upper body movement due to breathing. Smart speakers and wristbands track people's sleep duration and quality based on the detected sleep stages. As more vital signs can be measured using a smartphone, mobile device users can be benefited from at-home health screening, effective disease management, and telehealth services.

1.2.2 Non-Expert Users

The sensor-based mHealth applications aim to target the general public, enabling them to easily and conveniently measure vital signs, screen for health conditions, and check the effectiveness of disease management outside of clinical settings. Sensors in mobile devices collect the relevant data and health algorithms process and provide vital signs as clinicians or medical devices do. However, varying levels of users' expertise and experience cause some users to get reliable and accurate health-related results and others do not. The people who fail to benefit from mHealth applications often lack prior knowledge and experiences in (1) clinical domain, (2) mobile device usage, and/or (3) mHealth application usage. For instance, a smartphone-based heart rate application that measures heart rate from a user's fingertip placed on a smartphone camera and flashlight. The measurement process requires completely covering the camera and flashlight with the finger, applying only a slight amount of pressure, and staying rested. The measurement can be easily performed by *experts* users who know they are collecting PPG signals, are familiar with interacting with a smartphone camera, and/or have used the application before to get accurate heart rate measurements. On the other hand, the application can easily fail to provide an accurate measurement for *non-expert* users who lack knowledge in the underlying process of measuring heart rate and who are not experienced with the app or smartphone camera; the *non-expert* users can misplace their

finger on the camera and/or flashlight or press too hard against the camera to restrict the blood flow. When encountered with inaccurate results (e.g., heart rate of 200bpm), *expert* users can discern the errors; but, *non-expert* users can make irrational medical decisions based on unreasonable results, exposed to safety and health risks. In this thesis, we focus on such *non-expert* users who need explicit guidance to get accurate and reliable health-related results from the sensor-based mHealth application.

1.3 Failure Modes of Sensor-Based mHealth Applications

In general, the sensor-based mHealth applications require the users to acquire data and provide the data to the health algorithm. However, failures can occur in each phase, resulting in providing unreliable and inaccurate results for the users. While the algorithm expects high-quality data from the users, the users can input varying-quality data. It is mainly because, during the data acquisition phase, the users do not have a clear understanding of what types and quality of the data they should collect for the target health screening task to be correctly performed. While the users expect the algorithm to provide accurate and reliable results, the health algorithm can fail silently and provide unreasonable results. This expectation gap would get wider as the sensor-based mHealth applications (1) evolve to measure more diverse vital signs that require more complicated data acquisition procedures and (2) leverage more complicated "black box" neural network-based algorithms. In this section, I detail the failure scenarios during the data acquisition and algorithm processing phase when there is an expectation gap between the users and mHealth applications.

1.3.1 Uncertainty of Data Quality

Users with varying understanding of what they should collect during the data acquisition phase can result in data of various quality. Previous work [149] asks CHWs to capture images of RDTs for testing result interpretation. CHWs provided images of RDTs at different distances, under different lighting conditions, and with motion artifacts. With those images, it is difficult for an algorithm to process the image for accurate interpretation. A similar

problem exists in a telehealth setting. When clinicians want to visually inspect a body part, vague instruction (e.g., send me a picture of your eyes) would result in similar data quality issues. Due to this barrier, telehealth is mainly used for repeated visits and prescription updates and physicians often recommend physical visits for new health conditions.

Often vital sign measurement requires more complex data acquisition procedures. As mentioned previously, smartphone-based PPG signals [194, 195] can be obtained by placing a finger on a smartphone camera and flashlight with slight pressure. A smartphone-based blood pressure measurement [195] is done by simultaneously performing PPG signal extraction and placing the smartphone on the chest for seismocardiography. Capillary refill time (CRT) assessment, which is used for dehydration and Raynaud's syndrome screening, would require the users to place their finger on a smartphone camera and apply/release it at the right time. Any failures during the data acquisition would result in low-quality data. The health algorithm processes such data would provide inaccurate and unreliable results.

1.3.2 Uncertainty of Health Algorithms

Advances in artificial intelligence and machine learning have made medical diagnostic and screening tools more accurate and accessible. AI-powered diagnostic tools [8, 35] are intended to assist medical personnel by making unbiased decisions based on thousands of examples. In recent years, these models [36, 108, 118, 79] are even becoming available to consumers through the growth of mobile health. Despite the potential benefits of health AI systems, there are concerns about their performance in real-world settings. Data-driven models learn from examples, making them heavily reliant on the data upon which they have been trained. Previous work [106, 209] has found that machine learning models behave unpredictably on the unseen data. When health applications are put into the hands of consumers with a limited understanding of the underlying algorithms, they may provide poor quality data that lies outside the distribution of the data that was collected by experts. For example, consumers who are using a health application that involves image processing may take photographs in poor lighting conditions or framing of the target object. Even when the data is high-quality, it

may be captured with a smartphone that has different hardware specifications than the devices that were used to collect the training dataset. Unless the models are explicitly designed or trained to detect invalid data, the models will provide an inaccurate result without any warnings to the users. This problem [5, 183] is especially critical for the sensor-based mHealth applications since the users are not likely to have prior knowledge to discern such failures.

1.4 Design Space of mHealth Application Feedback

In general, the sensor-based mHealth applications require the following steps: (1) prepare data acquisition, (2) acquire target data, (3) input the acquired data to an algorithm, and (4) receive estimated vital signs or screening results. The expectation gap between the users and mHealth applications can occur in each step, especially when the users lack knowledge of target health conditions and how the mHealth application is developed. The users would not understand the types and quality of data expected from health algorithms. Even if they have a clear understanding, whether the acquired data meet the quality criteria is not easily determined. Furthermore, if the model provides uncertain and unreasonable results even with high-quality data, there is no way for the users to know such failures with limited background and experience. Throughout the mHealth application usage, constant interchange of information between users and the app is essential for reliable and accurate use. In this thesis, I present the feedback strategies at different stages of the sensor-based mHealth application for the users to receive accurate and reliable results and protect them from unexpected silent failures of the health medical algorithms.

The reliability and trust of prediction are impacted by both data and model uncertainty: (1) data uncertainty when the data is noisy or low quality and (2) model/algorithm uncertainty when the underlying predictive process does not understand input data and reduces the confidence of the prediction. To mitigate data uncertainty, different approaches should be employed at different stages. Before the data acquisition, it is crucial that the users know the target data (i.e., what signal they are collecting, the size and duration of the target data, which sensor they are using, and how the data should be acquired). This can be done with

an interactive tutorial that explains the data acquisition procedures with visual aids (e.g., video, animated images) [23, 143]. During the data acquisition, an intuitive user interface can also be helpful for users to understand and follow the data acquisition procedures. For images, where the target should be located can be visualized with a transparent viewfinder (e.g., bank check application). For audio and motion signals, the magnitude of signals can be visualized in real-time to indicate sound and motion activities happening during the acquisition. If the features of target data (e.g., sharpness, size of a target object in an image, frequencies of target audio signal, pattern of target motion, signal-to-noise ratio) are clearly defined for the algorithm, real-time quality assurance and guidance to satisfy the criteria can help users acquire high-quality data. When the data is acquired, the data can be visualized and interacted with the users for them to confirm the data quality. Providing examples of high-quality data would be an effective method for the users to decide whether to retry. From a computation perspective, the similarity of the acquired data to the distribution of known high-quality data (e.g., train dataset) for screening algorithms can be computed to provide a metric-based evaluation of the quality of the acquired data; if it is too dissimilar, the application can guide users to retry the data acquisition.

While data uncertainty can be resolved by intelligent guidance and feedback during data acquisition procedures, algorithm uncertainty is inherent to the underlying signal processing and machine learning algorithms. Although the algorithm uncertainty cannot be corrected by the users, they should be informed about how the algorithm processed the acquired data. It is found that the algorithms encountered with a novel, unseen data are likely to provide predictions that are inaccurate and uncertain with high confidence. In other words, even if the users provide high-quality data, the algorithm could fail if the data are not within the coverage of the dataset that the algorithm or machine learning model is trained on. Especially, deep neural network models that gain popularity for health prediction tasks are notorious for their silent failures and the difficult-to-interpretable underlying process. Such failures will occur even for a highly robust and reliable model since it is extremely unlikely that the model would have complete coverage over all existing data and have 100% accuracy. This is

particularly concerning in health applications where high-stakes decision-making is involved. From deep learning literature, researchers are actively investigating methods to estimate the uncertainty of the models to improve the model's reliability and trust. Providing such information plays a critical role for users to trust the predictions with high confidence and low uncertainty, and vice versa.

While intuitive tutorials and user interfaces for mHealth applications are widely studied [23, 143], much effort is needed for data quality-based feedback and guidance and providing uncertainty of algorithms' predictions in the mHealth context. Thus, this thesis focuses on exploring approaches to improve data acquisition, compute the quality of the acquired data, and inform the users of algorithm uncertainty. These approaches would help the users can benefit most from mHealth applications without sacrificing accuracy and reliability.

1.5 Thesis and Research Questions

In this thesis, I argue that **real-time and post-hoc feedback can enable *non-expert* users to acquire high-quality data while using sensor-based mHealth applications, improving the accuracy and reliability of health screening algorithms.** To support this thesis, I address the following research questions:

RQ.1: How can we design sensor-driven real-time feedback to improve data acquisition by *non-expert* users? (§ 3, § 4)

RDTScan in Chapter 3 explores real-time image processing-based quality assurance to guide the users to capture high-quality rapid diagnostic test images. CapApp in Chapter 4 explores multi-modal sensing to the users to follow each step of data acquisition procedures for capillary refill time assessment.

RQ.2: How can we design post hoc feedback for *non-expert* users to receive reliable and trustworthy results from black-box deep learning algorithms? (§ 5)

In Chapter 5, I leverage state-of-the-art out-of-distribution detection methods to assess whether the model would provide reliable and trustworthy health prediction results with the data acquired by the users.

RQ.3: How does the feedback for *non-expert* users help achieve performance that is comparable to the *experts'*? (§ 3, § 4, § 5)

In Chapter 3 and Chapter 4, I demonstrate that the accuracy of disease screening and vital sign measurement using the data acquired by the users are comparable to medical experts' abilities. In Chapter 5, I demonstrate that the users' ability to distinguish uncertain and unreliable health results is comparable to machine learning experts' ability.

Chapter 2

BACKGROUND AND RELATED WORK

2.1 Health-Related Applications

2.1.1 For Medical Experts

Computer-aided diagnosis [130, 66] provides clinicians with information and knowledge for medical decision-making. In recent years, machine learning has been widely integrated into computer-aided diagnosis to help doctors diagnose patients easier, faster, and more accurately. Machine learning models that learn from large-scale medical datasets are able to detect various symptoms and conditions, outperforming human experts in detecting subtle symptoms. To name a few, retinal diseases [35] and lung cancer [8] can be early diagnosed using fundus and x-ray images. Breast cancer [54] can be detected and classified using deep convolutional networks from whole slide breast histopathology images. Although studies [53, 125] have demonstrated that these applications can improve diagnosis accuracy, reduce manual work, and reduce human errors, lack of user acceptance and trust prevents these applications to be widely adopted in the diagnosis practice. This is mainly due to these tools failing to provide relevant information or explanation of outcome [92, 205]. To address this challenge, Cai et al. [22, 23] proposed tools to support medical decision-makers by providing relevant information and interactivity to the expert users. As more health applications are becoming to non-medical users, these types of feedback and information should be also provided since their limited background in the clinical and medical field would lead to making wrong medical decisions.

2.1.2 For Non-Medical Experts

Community Healthcare Settings

In mid- and low-income settings, access to healthcare and medical infrastructure is limited for underserved populations. mHealth applications run on mobile devices play an essential role in improving the access and quality of healthcare. Mobile devices are used for facilitating communications between clinics and health workers [152, 97], translating paper-based screening survey into electronic version [41, 126], and disseminating health-related information [21]. As more individuals own mobile devices, location information is used for predicting disease transmission [197, 191].

In recent years, rapid diagnostic tests (RDTs) have been widely used for diagnosing various infectious diseases (e.g., malaria, leishmaniasis, influenza, HIV). Although RDTs are becoming increasingly user-friendly, subjective interpretation of their output leaves room for diagnostic uncertainty and error. This topic has mostly been explored in the context of malaria RDTs. Harvey et al. [69] observed CHWs as they administered malaria RDTs and noted that only 54% were able to correctly interpret the test results; the most common mistakes were failures to identify faint positive lines or invalid results. Harvey et al. found that multi-day training programs improved interpretation accuracy to 93%, but such programs are impractical in many cases since they can take CHWs away from other responsibilities. Although there have been attempts to improve procedural adherence and mitigate confusion by standardizing terminology, labeling, and instructions across RDT manufacturers, these efforts have typically been slow to implement and have been focused around RDTs for the same condition [80].

Many researchers have proposed standalone devices and smartphone adapters to control the imaging environment (i.e., ambient lighting, shadows, camera position) for automatic RDT interpretation [144]. An example of a standalone device for this purpose is Fio's battery-operated Deki Reader[1], which includes an internal chamber with controlled lighting. Herrera et al. [74] compared the RDT interpretation accuracy of the Deki Reader against visual

[1] `http://fio.com/`

inspection of the RDTs by experts and saw 99% concordance between the two. Mudanyali et al. [131] translated many of the Deki Reader's features to a smartphone-based system, utilizing an LED array and other optical components for imaging. Targeting a less expensive solution, Dell et al. [38, 39] and Ozkan et al. [145] independently proposed the use of plastic stands for consistent positioning between the smartphone's camera and the RDT. However, incorporating hardware imposes additional financial costs to end-users and thus reduces the ease of access and ubiquity that RDTs engender in the first place. When transferring the burden of image quality control from hardware to software, real-time guidance is needed to ensure that high-quality photographs are taken while being mindful of computational overhead for low-end smartphones. Besides mHealth applications for RDTs, as more mHealth applications rely on input data acquired by users, the design of feedback presented to the users would play a more important role in providing accurate health-related results.

At-Home Settings

With the increasing ubiquity of smartphones and advances in their computing power, machine learning-based health screening can be done on mobile devices. Various machine learning-based mobile health applications have been proposed to detect health conditions (e.g., traumatic brain injury [119], pancreatic cancer [118], jaundice [36], mental health [59, 182]) and vital signals (e.g., heart rate [108], respiratory rate [108], heart rate variability [79], blood pressure [195, 170], SpO2 [76]). Such mobile health applications can benefit nurses, health workers, and the general population for easier medical screening. A smartphone camera is being used to detect subtle yellowness from skin and sclera to estimate bilirubin levels [118, 36]. Subtle color changes due to blood flow on a face can be used to extract photoplethysmography (PPG) signals using deep neural networks [108]. A combination of a camera and external light source (e.g., flash light, custom LEDs, screen) can extract PPG [67], SpO2 [20, 76], and hemoglobin levels [194]. On top of these sensors, inertia sensors are used to perform seismocardiography to estimate blood pressure [195]. More sensors are being used simultaneously by the *non-expert* users and low-quality data can cause significant

performance drops. Billicam [36] and BilliScreen [118] address this issue by using a color calibration card to correct the input images. PupilScreen provides a 3D-printed box to control distance and light for capturing pupillary light reflex. MetaPhys [109] found that the image-based physiological sensing suffers from a performance drop when encountering unseen data (e.g., users with different skin tones), proposing a domain adaption method to address this challenge. Seismo [195] observed that the placement of smartphones, shaky hands, and ticker tissues prevent it from providing accurate blood pressure measurements. These prior works show the potential benefit and feasibility of sensor-based mHealth applications for easier medical screening by *non-expert* users. At the same time, they highlight the importance of data quality to maximize the screening algorithm to provide accurate screening results and vital measurements.

2.1.3 Trustworthy AI for Health

Machine learning systems are deployed in real-world settings to billions of users, making significant impacts on high-stake decision-making such as healthcare, policy, economy, and transportation. Since failures in such machine learning systems can cause fatal consequences, building trustworthy AI is one of the most important problems in the machine learning community. However, analyzing and interpreting deep learning models is extremely challenging because of their complexity and hidden internal state. In recent years, there are active and ongoing efforts aimed at making machine learning systems causal [9, 139], explainable [2, 93, 112, 84, 89], fair [3, 157, 43], robust [44, 50, 71, 45, 188], and privacy-preserving [1, 148, 124, 187]. Bhatt et al. [16] proposed to provide estimated uncertainty to the users in making a decision and having trust in machine learning models. This work explores a similar approach where we adopt out-of-distribution detection as a method to measure uncertainty in the context of health.

The health machine learning models often show high accuracy on their test datasets. However, their performance is questionable in real-world settings where the input data can vary drastically, resulting in unreliable prediction results [185, 162]. Researchers have investigated

the dataset shift problem for medical imaging (e.g., x-ray [26, 24], fundus eye images [26], CT scans [192], dermatology [162, 146]), focusing on developing and evaluating out-of-distribution detection methods for specific domains. However, as more diverse input data types (e.g., images, audio, motion data, touch, locations) are used for mHealth applications, the effectiveness of out-of-distribution detection methods has not been investigated in such diverse settings. We took a step further from the existing work to investigate and quantify its effect on improving reliability and trustworthiness in the context of machine learning for health.

2.2 Data Quality Assurance Methods

2.2.1 Guided Data Acquisition

Whether a person is taking a picture of an RDT, body part, or another object, post-processing can only do so much to improve the quality of a poorly captured image. In light of this issue, researchers have explored ways of introducing real-time guidance for media capture in various domains. NudgeCam [27] leverages the smartphone's inertial sensors to track the camera's stability and orientation. NudgeCam also assesses the video content itself, using real-time image processing to check the overall brightness of the scene and to detect faces. EasySnap [198, 82] provides text-to-speech audio cues to help blind and low-vision photographers take well-framed pictures. For pictures involving people, EasySnap uses face detection to continuously monitor the size and position of the subject's face in the photo. For pictures involving objects, the photographer can walk up to a target object so that EasySnap can register its visual features; once the object is registered, the photographer can move away from the object and EasySnap ensures that the object remains in view.

Guided image capture has also been created for specific object categories. One notable example is bank check recognition for online banking[2,3]. To the best of our knowledge,

[2] `https://play.google.com/store/apps/details?id=com.wf.wellsfargomobile&hl=en_US`

[3] `https://play.google.com/store/apps/details?id=com.infonow.bofa&hl=en_US`

these apps rely on optical character recognition to localize the consistent features of the bank check (e.g., routing number, check ID) and use those regions to infer the remaining contents [78]. Because countries have unique bank check formats, companies often develop one system per country [64, 65], which limits generalizability. Chen et al. [29] created an app called SmartDCap to help people scan paper documents with a smartphone. SmartDCap continuously checks the framing and sharpness of the document, consolidating these metrics into a score that is shown to the end-user along with audio feedback. Inspired by the previous work, this thesis aims to design and implement feedback to the users to acquire input data that can be accurately and safely computer for medical screening.

2.2.2 Dataset Shift Detection for Deep Learning Models

When machine learning models are deployed in real-world settings, it is known to fail when encountered with input data that are different from the training dataset. However, machine learning models fail silently by providing a high probability of an incorrect result. The information is limited to inform whether the model provides unreasonable results due to dataset shift. Therefore, machine learning models are notoriously bad at telling the users that they do not know. Unless the models are explicitly trained with the "other" class, they are designed to classify the input data into one of the predefined classes. Since "other" class can be any data that are not relevant to the classes of interest, training with "other" class requires a significantly large dataset that covers a wide range of "other" class data. Even if they include the "other" class, the performance of detecting the class is limited unless they can access such a large-scale dataset.

Recently researchers have proposed various methods to estimate the models' uncertainty due to dataset shift. The proposed methods leverage the output of the models to effectively detect *out-of-distribution* input that is different from the known distribution, *in-distribution*. Softmax confidence [72] has been the baseline for the out-of-distribution detection. Several work has been proposed for out-of-distribution detection using deep ensemble [98], Mahalanobis distance [102], Gram matrices [169], energy score [107], temperature scaling [105, 169], input

perturbation [105, 102], mean and variance of channels activations [156]. Alternate training strategies [73, 101, 116, 127] have been proposed to enable the model to detect out-of-distribution. Generative models [136, 159, 129, 208] are proposed to detect out-of-distribution examples. The dataset shift is a critical piece of information to reduce the gap between black-box algorithms and the users. In this work, we explore the utility and implication of dataset shift information for the *non-expert* users for using deep learning-based mHealth applications accurately and safely.

Chapter 3

THE DESIGN AND EVALUATION OF A MOBILE SYSTEM FOR RAPID DIAGNOSTIC TEST INTERPRETATION

3.1 Introduction

Over the past couple of decades, rapid diagnostic test (RDTs) have emerged as a potential solution to the pressing need for accessible medical testing. RDTs utilize biochemistry to transduce the load of a biological sample (e.g., blood, urine, nasal swab) to an analog colorimetric output. As such, RDTs enable point-of-care diagnostics without the need for expensive equipment. Pregnancy tests are one of the most well-known types of RDTs, but RDTs exist for many other health purposes, such as malaria [10, 132, 117], influenza [77, 164], and HIV [87, 202]. Recently, RDTs are being developed to support convenient COVID-19 testing during the pandemic[1,2]. RDTs are viable to produce at scale, making them an inexpensive (∼$1 USD each) alternative to laboratory tests and ideal for point-of-care medical screening [114, 193]. RDTs are often associated with community healthcare settings in low- and middle-income countries where resources and access to sophisticated testing facilities are limited, but RDTs are used worldwide in clinics and homes as well [147, 17].

As RDTs become more commonplace, one concern is that people may misinterpret the visual results that appear on their tests—overlooking faint lines, thinking they are seeing lines that are not actually present, or misunderstanding the lines' meanings [69]. A system that interprets RDTs on the user's behalf would limit such errors, leading to improved test accuracy and higher utility among end-users. The ability to digitally document RDT results could also facilitate community-wide reporting, contact tracing, and surveillance networks

[1] `https://cellexcovid.com/`

[2] `https://en.wondfo.com.cn/product/wondfo-sars-cov-2-antibody-test-lateral-flow-method-2/`

during disease outbreaks. The ideal system would satisfy the following requirements:

- **High interpretation accuracy:** The ideal system would have comparable diagnostic accuracy to an expert directly reading the RDTs themselves. In addition, the ideal system would be consistent across settings (e.g., smartphone model, ambient environment). Such a system would improve the effective diagnostic accuracy of RDTs in the field, particularly amongst novice users, while removing the potential for subjective, biased, or rushed decision-making.

- **Smartphone-only:** RDTs are an attractive option for community use because all of the equipment needed to run an RDT comes within an inexpensive kit. Introducing hardware or smartphone accessories, as past researchers have proposed [144, 38, 39, 74, 131], hinders deployability. Roughly 45% of the global population owns a smartphone as of April 2020 [12]. Specifically in sub-Saharan Africa, the Global System for Mobile Communications (GSMA) estimates that the fraction of people who have a SIM connection will grow from 77% in 2019 to 86% in 2025, and the fraction of those connections coming from smartphones will rise to 65% in 2025 [68]. Therefore, requiring a smartphone does not necessarily introduce significant burden.

- **Configurability to new RDT designs:** Like many other products, lateral flow RDTs are produced by manufacturers with no overarching design standard beyond having results displayed as a set of lines. Existing smartphone-based RDT readers are catered to specific RDT brands[3,4], and a machine learning approach would require a dataset with hundreds of images in diverse settings for model training to accommodate new RDT designs. The ideal system would be quickly configurable so that new RDTs can be used in response to epidemics.

---

[3] `https://www.novarumdx.com/`

[4] `http://www.albagaia.com/hydrosense-app`

With these requirements in mind, we present RDTScan, an open-source system that supports automatic RDT detection and interpretation on smartphones. RDTScan provides real-time guidance to users so that they capture a high-quality photograph of their completed RDT; RDTScan then analyzes that photograph to infer the RDT's result. RDTScan builds off of our previous work [149] by improving the robustness of our feature-based template matching approach to RDT detection, extending our system to accommodate new RDT designs, adding new quality assurance checks to ensure accurate interpretation across different RDT form factors, and rigorously testing our system through multiple studies. We first evaluated our image processing pipeline in a controlled laboratory study to demonstrate RDTScan's ability to automatically interpret RDTs. After we established that our pipeline could either match or exceed human interpretation capabilities, we engaged with community health surveillance programs to deploy RDTScan in two contrasting scenarios: (1) at-home influenza testing in Australia and (2) malaria testing by CHWs in Kenya. Participants in both studies had success capturing a high-quality photograph of their completed RDT, reaching an overall success rate of 83.3% and 91.9% in the Australia Influenza Study and Kenya Malaria Study, respectively. RDTScan showed comparable or slightly better interpretation performance to that of experts. In the Australia Influenza Study, RDTScan achieved 97.5% (85.7% sensitivity, 98.7% specificity) compared to the experts and 83.6% accuracy (33.9% sensitivity, 98.4% specificity) compared to gold-standard clinical measurements. In the Kenya Malaria Study, RDTScan achieved 96.3% accuracy (95.5% sensitivity, 98.7% specificity) and 85.5% accuracy (92.9% sensitivity, 65.0% specificity) compared to the experts and gold-standard clinical measurements, respectively. In summary, our research delivers the following contributions:

1. An open-source, smartphone-based RDT interpretation system[5] that achieves comparable accuracy to experts who are experienced in RDT administration without the need for additional hardware,

2. An in-lab validation study showing RDTScan's interpretation accuracy across RDTs,

[5] `https://github.com/cjpark87/rdt-scan`

analyte concentrations, smartphone devices, and lighting conditions, and

3. Two field evaluations that demonstrate the efficacy of RDTScan in vastly different settings.

3.2 Related Work

RDTs leverage an assortment of techniques to detect medical conditions [114, 193]. Covering these techniques is out of the scope of this work, so we instead focus on how RDTs have been incorporated into healthcare workflows around the world. We then describe past approaches to automatic RDT interpretation, after which we discuss object interpretation for broader object categories.

3.2.1 Current Practices with RDTs

RDTs are typically used in settings outside of hospital environments, namely community and primary care settings. RDTs are popular in these areas because of their low cost, ease of use by non-lab technicians, and portability. RDTs have been used both for illnesses that are endemic in certain low- and middle-income countries (e.g., malaria in sub-Saharan Africa [10, 132, 117], leishmaniasis in India [177]) as well as illnesses that are clinical priorities in high-income countries (e.g., group A streptococcus [30], influenza [196, 206], HIV [87, 202]). Some RDTs are already seeing use for at-home testing [147, 17], and we anticipate this trend will grow as more RDTs gain regulatory clearance.

Although RDTs are becoming increasingly user-friendly, subjective interpretation of their output leaves room for diagnostic uncertainty and error. This topic has mostly been explored in the context of malaria RDTs. Harvey et al. [69] observed CHWs as they administered malaria RDTs and noted that only 54% were able to correctly interpret the test results; the most common mistakes were failures to identify faint positive lines or invalid results. Harvey et al. found that multi-day training programs improved interpretation accuracy to 93%, but such programs are impractical in many cases since they can take CHWs away from

other responsibilities. Although there have been attempts to improve procedural adherence and mitigate confusion by standardizing terminology, labeling, and instructions across RDT manufacturers, these efforts have typically been slow to implement and have been focused around RDTs for the same condition [80]. Our work seeks to shift interpretation burden from novice users to an automated analysis platform while being flexible enough to accommodate new designs with significantly less overhead than what would be required for a data-driven model.

3.2.2 Automated RDT Capture and Interpretation

Image interpretation is most successful when a clear image has been taken of the target object. Therefore, many researchers have proposed standalone devices and smartphone adapters to control the imaging environment (i.e., ambient lighting, shadows, camera position) for automatic RDT interpretation [144]. An example of a standalone device for this purpose is Fio's battery-operated Deki Reader[6], which includes an internal chamber with controlled lighting. Herrera et al. [74] compared the RDT interpretation accuracy of the Deki Reader against visual inspection of the RDTs by experts and saw 99% concordance between the two. Mudanyali et al. [131] translated many of the Deki Reader's features to a smartphone-based system, utilizing an LED array and other optical components for imaging. Targeting a less expensive solution, Dell et al. [38, 39] and Ozkan et al. [145] independently proposed the use of plastic stands for consistent positioning between the smartphone's camera and the RDT. Because hardware and accessories enforce constraints on the imaging environment, the accompanying software can be highly tuned and efficient; however, incorporating hardware imposes additional financial costs to end-users and thus reduces the ease-of-access and ubiquity that RDTs engender in the first place. In contrast, RDTScan transfers the burden of image quality control from hardware to software while still being mindful of on-device computational limits.

---

[6] `http://fio.com/`

There are commercial products for automatic RDT interpretation, as well. For example, Ellume[7] produces a custom RDT cartridge with embedded sensors that can read a custom immunoassay made with fluorescent nanoparticles. Apps like Novarum's DX Mobile Reader[3] and Albagaia's Hydrosense app[4], on the other hand, use computer vision to analyze lateral flow RDTs; unfortunately, there is no documentation about their algorithms or performance because they are proprietary apps. Regardless, these products are catered to specific RDT brands, so the underlying software can rely on design-specific features to interpret those RDTs. Our approach is unique in that we aim to accommodate new RDT designs with a single template image and metadata, which is significantly less overhead than the large image datasets that would be needed to support a machine learning approach. Our prior work [149] demonstrates the first step we took towards supporting RDT interpretation: a smartphone app that uses real-time image processing to ensure high-quality image capture. We have since improved upon RDTScan in a few ways, including a better method for RDT detection and additional quality assurance methods for blood and glare detection, fiducial tracking, and color-aware line interpretation. We also rigorously evaluate our automatic result interpretation algorithm across RDTs, analyte concentrations, smartphone devices, and lighting conditions through both an in-lab evaluation and two case studies.

3.2.3 Guided Media Capture

Whether a person is taking a picture of an RDT or another object, post-processing can only do so much to improve the quality of a poorly captured image. In light of this issue, researchers have explored ways of introducing real-time guidance for media capture in various domains. NudgeCam [27] leverages the smartphone's inertial sensors to track the camera's stability and orientation. NudgeCam also assesses the video content itself, using real-time image processing to check the overall brightness of the scene and to detect faces. EasySnap [198, 82] provides text-to-speech audio cues to help blind and low-vision photographers take well-framed

[7] `https://www.ellumehealth.com/`

pictures. For pictures involving people, EasySnap uses face detection to continuously monitor the size and position of the subject's face in the photo. For pictures involving objects, the photographer can walk up to a target object so that EasySnap can register its visual features; once the object is registered, the photographer can move away from the object and EasySnap ensures that the object remains in view.

Guided image capture has also been created for specific object categories. One notable example is bank check recognition for online banking[8,9]. To the best of our knowledge, these apps rely on optical character recognition to localize the consistent features of the bank check (e.g., routing number, check ID) and use those regions to infer the remaining contents [78]. Because countries have unique bank check formats, companies often develop one system per country [64, 65], which limits generalizability. Chen et al. [29] created an app called SmartDCap to help people scan paper documents with a smartphone. SmartDCap continuously checks the framing and sharpness of the document, consolidating these metrics into a score that is shown to the end-user along with audio feedback.

RDTScan builds upon these developments with a specific focus on RDTs. RDTScan performs the RDT-specific task of result interpretation while being configurable to a variety of RDT designs. RDTScan also provides real-time, human-readable feedback so that users can capture high-quality photographs of their RDTs and maximize the likelihood of correct interpretation.

3.3 RDTScan Design

In this section, we first introduce the standards and terminology for lateral flow RDTs—the specific subgroup of RDTs that this work addresses. We then use this vocabulary to describe the RDTScan system. Throughout this section, we refer to the HLS (hue-saturation-lightness) color space, which is an alternate image representation to the standard RGB (red-green-blue).

---

[8] `https://play.google.com/store/apps/details?id=com.wf.wellsfargomobile&hl=en_US`

[9] `https://play.google.com/store/apps/details?id=com.infonow.bofa&hl=en_US`

For our purposes, the HLS color space is defined as follows: H $\in$ [0, 179], L $\in$ [0, 255], S $\in$ [0, 255].

3.3.1 Standards and Terminology

In a typical lateral flow test, a liquid biological sample migrates across a strip via capillary flow. The sample first passes over a conjugate pad that holds the particles needed to create the colorimetric output (e.g., colloidal gold, latex) and then over zones that immobilize a target antigen, antibody, or protein. As more of the target is immobilized, visible colored lines appear to the end-user.

Lateral flow RDTs come in a variety of form factors, the two most common being *cassettes* and *dipsticks*. RDT cassettes are activated by putting a liquid biological sample in the sample well and then adding a buffer solution to the buffer well. RDT dipsticks are activated by mixing a biological sample with a buffer solution in a test tube and then dipping the strip into the tube. In this work, we focus our attention on the two RDTs shown in Figure 3.1: (1) AccessBio's CareStart Malaria Pf/Pv test[10], an RDT cassette that analyzes whole blood; and (2) Quidel's QuickVue Influenza A+B test[11], an RDT dipstick that analyzes a nasal swab specimen.

An RDT's results appears in the *result window*—the thin rectangular region where the immunoassay itself is exposed to the user. Whenever an RDT has been activated properly, its *control line* will be visible to the user. Any other lines in its result window, called *test lines*, indicate the presence of a target analyte. The intensity of the test lines are a function of the biological sample's analyte concentration; the higher the load, the more intense the test lines will appear. The control line, on the other hand, is typically intense as long as the user administered the RDT properly.

Lateral flow RDTs can vary by more than just their form factor. RDTs can have one

[10] `http://www.accessbio.net/eng/products/products01_02.asp`

[11] `https://www.quidel.com/immunoassays/rapid-influenza-tests/quickvue-influenza-test`

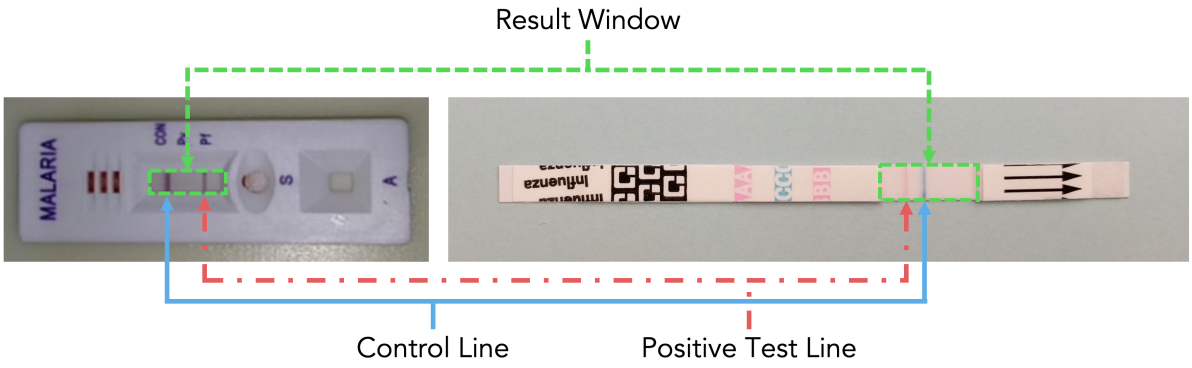


Figure 3.1: The two RDTs evaluated in this work: (left) the CareStart Malaria Pf/Pv RDT cassette, showing a valid positive case of *Plasmodium falciparum*, and (right) the QuickVue Influenza A+B RDT dipstick, showing a valid positive case of influenza A.

or many test lines depending on how many strains of the same pathogen they can detect. Going from left to right in Figure 3.1, the CareStart RDT can show a control line and then two lines indicating the presence of different malaria parasites (*Plasmodium falciparum* and *Plasmodium vivax*). The QuickVue RDT, on the other hand, can show a control line between two test lines for different types of influenza (A and B). The color, position, and order of these lines depend on the manufacturer's specifications.

The diversity of RDT designs, combined with a lack of standardized markings, makes it difficult for a single algorithm to interpret multiple RDT designs. A data-driven algorithm would work if hundreds or thousands of annotated images are available for each new RDT design. However, it would take significant effort and resources to activate and photograph all of those RDTs, creating a barrier for quickly adapting such an algorithm to new RDT designs. These challenges led to the development of RDTScan, which utilizes feature-based template matching to accommodate new RDT designs.

Table 3.1: The metadata required to accommodate a new RDT design with RDTScan.

Data Field	**Description**
Template image	A non-skewed, tightly cropped photo of the RDT
Result window corners	The $(x,\ y)$ pixel coordinates denoting the top-left and bottom-right corners of the general region where the results will appear
Line positions	The (x, y) pixel coordinate denoting the center of the various lines in the RDT
Line meanings	The meaning of the top line (e.g., "control", "influenza A")
Viewfinder scale	The relative height of the viewfinder compared to the screen's height
Fiducial locations (optional)	The $(x,\ y)$ pixel coordinates denoting the top-left and bottom-right corners of variable dark-colored markings that have a fixed location
Line color (optional)	The range of color hue values (H channel) expected for each of the RDT's lines
Line intensity (optional)	The minimum expected difference in brightness (L channel) between the immunoassay and the lines; increasing the threshold requires a more intense line to appear

3.3.2 Required Metadata for New RDTs

To leverage feature-based template matching, RDTScan requires a reference template image and the metadata listed in Table 3.1. The template image is a photograph of an unused, unmodified RDT with as few imperfections as possible (i.e., no shadows or perspective skew). One way to create such an image is by taking a photograph of the target RDT on an

uncluttered background with a free document scanning app like Microsoft's Office Lens[12]. Once the reference template image has been created, the image is loaded in an image-editing software program (e.g., Microsoft Paint, Photoshop) to identify the rough locations of the result window and lines.

RDTScan also accepts three optional pieces of metadata to improve performance for a particular RDT. To aid with detection, RDTScan can take advantage of *fiducials*—dark markers that are consistent in location, but not appearance, across RDTs from the same manufacturer (e.g., bar codes). To aid with interpretation, RDTScan can leverage the expected color hues of the different lines. Finally, RDTScan has a default value for the minimum expected difference in brightness (L channel) between the lines and the background immunoassay for a positive test result, but that parameter can be adjusted to prioritize sensitivity or specificity. The following subsections describe how RDTScan uses the aforementioned metadata.

3.3.3 Detection and Quality Checking

The clearer the image of an RDT, the easier it is for an algorithm to automatically interpret the test results. Because there is an upper limit to how much an image can be post-processed, RDTScan is designed to facilitate high-quality image capture. In our initial work [149], we presented an early instantiation of RDTScan catered towards a single RDT. Participants in that first deployment noted a few issues that hindered their ability to capture a clean photograph of their RDTs, such as instances of glare, shadows, and blood appearing in the result window. Furthermore, we ran into many challenges as we tried to apply that version of RDTScan to different RDT designs like Quidel's QuickVue RDT. These challenges motivated improvements to our RDT detection approach and quality assurance checks. We summarize RDTScan's complete functionality below.

---

[12] `https://www.microsoft.com/en-us/p/office-lens/9wzdncrfj3t8`

Camera Configuration

One way to enforce consistency during image capture is through intelligent hardware configuration. Ambient lighting can affect the appearance of objects, so RDTScan activates the smartphone's flash by default as a dominant illuminator around the RDT. RDTScan also utilizes the operating system's auto-focus, auto-exposure, and auto-white-balance functions to adjust the camera's properties with respect to the target RDT. By default, these functions are designed for global optimization—adjusting the camera's properties with respect to its entire field-of-view. RDTs with a narrow aspect ratio only take up a small part of the overall image frame, so global optimization would cater to the RDT's background in those cases. Instead, RDTScan enforces local optimization at the center of the camera's field-of-view where the RDT is expected to be.

RDT Detection

RDTScan employs feature-based template matching to minimize the bootstrapping effort required to locate an RDT design within an image. RDTScan first extracts unique visual features from both the template image and the camera frame and then identifies matches between feature keypoints via brute-force. RDTScan then uses a least-squares procedure to calculate a 3×3 homography matrix that maps the corners of the template to an irregular quadrilateral that surrounds the RDT in the camera frame. In our prior work [149], we used the BRISK [103] feature extractor because of its efficiency on low-end devices at a slight cost to accuracy. However, getting an accurate homography matrix is particularly critical for dipstick RDTs since their uneven aspect ratio leaves little room for error. In the updated version of RDTScan, we prioritize accuracy over latency by utilizing the SIFT [111] feature extractor. SIFT is designed to be scale-invariant, which provides two benefits: (1) it allows RDTScan to provide more accurate guidance if the size of the RDT is significantly different between the camera frame and the template image, and (2) it allows RDTScan to downsample incoming camera frames by $\times 1/2$ to expedite computation without significantly impacting accuracy.

We further reduce computation time by only extracting SIFT features within the middle of the screen where the viewfinder is located (see details on the user interface in Section 3.3.5). These two optimizations reduce the processing time per frame by a factor of $\times 1/16$ compared to using SIFT to extract feature keypoints from the entire frame. Computational complexity during image capture is an important consideration because RDTScan analyzes video frames in real-time to provide end-users with feedback; as computations become longer, the delay between updated feedback updates increases.

Quality Assurance – Exposure

Images that are dim due to underexposure or washed out due to overexposure make it difficult to see lines in the RDT's result window. To check the camera frame's brightness, RDTScan computes a histogram of the frame's L channel. Frames are considered underexposed if the histogram's maximum value is less than 125, and frames are considered overexposed if at least 20% of the pixels are greater than 255.

Quality Assurance – Sharpness

Because the underlying lateral flow immunoassay enacts changes in line intensity within the result window, image sharpness is the principle factor that determines whether a line will be visible. To check that there is sufficient sharpness, RDTScan computes edge intensity variance of an image I according to the Laplacian operator [151]:

$$\text{var}(\text{Laplacian}(I)) = \text{var}\left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}\right) \ \forall \ (x, y) \in I \tag{3.1}$$

High variance indicates the presence of both smooth regions and intense edges, which is typical of most focused images. However, the ideal amount of edge variance depends on the target RDT design since some are more visually complex than others. RDTScan computes the edge variance of the template image before the app is loaded to serve as a baseline. Because end-users are expected to position their RDT at the center of the frame within a viewfinder,

RDTScan computes the edge variance in that region for incoming frames. A frame passes this check if its edge variance is at least 80% that of the template image.

Quality Assurance – Position, Size, and Orientation

Feature-based template matching can extract any affine transformation between a template and a target image. However, imposing some standards regarding the size, position, and orientation of the RDT in a given camera frame is important for consistent review by both humans and software. To check these properties, the irregular quadrilateral that defines the border of the detected RDT in the camera frame is converted to a rotated rectangle by computing the average length of opposing sides and placing them around the center of the original shape. RDTScan checks the position, size, and orientation of the resulting rectangle as follows:

- **Position**: The distance between the center of the rotated rectangle and the center of the camera frame should be within 10% of the image height.
- **Size**: The area of the rotated rectangle should be close to the ideal size of the RDT in the camera frame as dictated by the viewfinder scale parameter in the metadata. This check has a tolerance of 10% of the camera frame's height.
- **Orientation**: The angle between the rotated rectangle's longer axis and the vertical axis of the camera frame should be between -10° and +10°.

Quality Assurance – Glare

Some lateral flow tests have a clear, glossy membrane to protect the underlying immunoassay from environmental damage. Unfortunately, this membrane can create glare on the result window depending on the relative positions and orientations of the camera, flash, RDT, and ambient lighting during image capture. Just as how RDTScan checks the overall illumination on the entire RDT, it also checks the brightness histogram specifically within the result window for any clipping. As before, the result window is considered underexposed if the

histogram's maximum value is less than 125 and overexposed if at least 20% of the pixels are greater than 255.

Additional Fiducial Detection

Some RDTs, particularly dipsticks, do not have identical markings across tests due to how they are manufactured. Quidel's QuickVue RDTs, for example, are cut from a single sheet at an interval that is different than the underlying pattern. This process creates strips with the same general design but with offset features, posing a challenge for many computer vision-based approaches. To accommodate such designs, RDTScan can leverage fiducials in the target RDT design. RDTScan searches for these fiducials using k-means clustering within a 5D spatial-chromatic space according to RGB color and (x,y)-position. We assume k=5 for the number of clusters since most RDT designs have few unique colors. The position of the darkest cluster is then compared against the fiducial locations specified in the metadata. Incoming camera frames pass this check if the distance between the x-coordinates of the detected and pre-specified fiducial locations is within 50% of the fiducial's width and the corresponding y-coordinates are within 50% of the fiducial's height.

3.3.4 Interpretation

RDTScan analyzes the result windows of captured RDTs for the presence of test and control lines. Our initial version of RDTScan [149] relied strictly upon the L channel to identify lines, making it susceptible to false positives from blood, shadows, or other irregularities in the result window. Below, we describe RDTScan's interpretation algorithm and the improvements that have been made to make it robust to such issues.

Result Window Crop

To isolate the result window from the rest of the image, RDTScan crops the camera frame around the detected RDT's bounding box and rectifies the remainder so it has the same

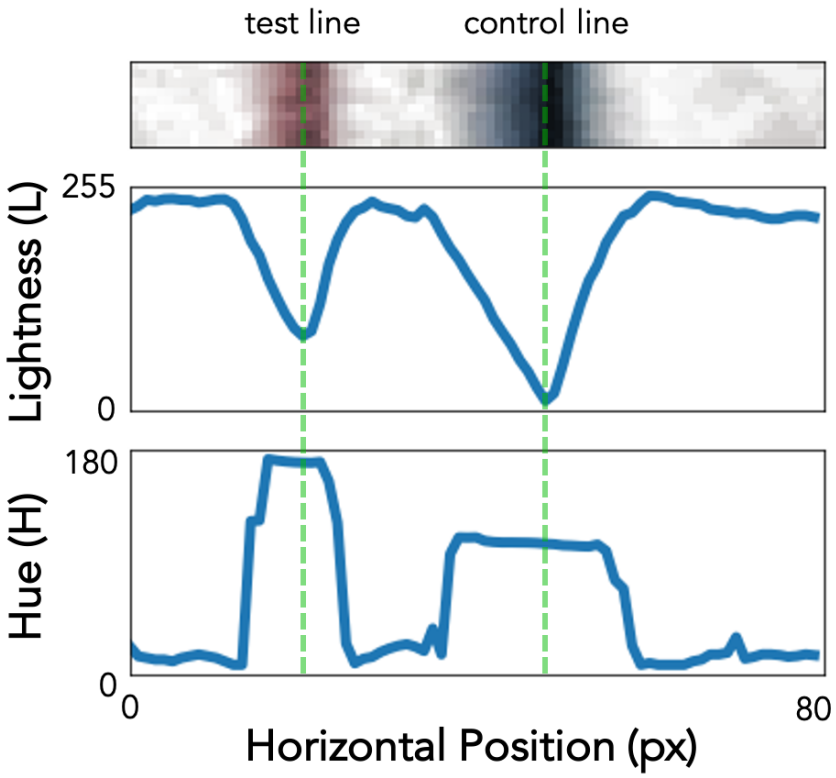


Figure 3.2: An illustration of RDTScan's interpretation algorithm, which leverages both the lightness (L) and hue (H) channels to identify lines within the result window.

dimensions as the template image. The resulting image is cropped a second time according to the coordinates specified in the metadata to highlight the RDT's result window. In many RDT designs, the result window is sunken into the cassette or dipstick, creating shadows along the edges; therefore, a thin border (10% of the result window's height) is cropped from the edges parallel to the control line to ignore those potential false positives.

Result Window Enhancement

RDTScan applies image post-processing to emphasize faint lines. Techniques like brightness histogram equalization typically work best on images that have a variety of dark and bright

pixels. RDTs tend to be mostly white, so applying equalization globally would darken most of the pixels and make faint lines less distinctive. To avoid this issue, RDTScan uses contrast-limited adaptive histogram equalization (CLAHE) [154] to enhance the image's contrast on a tile-by-tile basis.

Presence of Blood

One common issue we saw in our prior work with blood-based RDT cassettes was that CHWs sometimes overused or misplaced blood samples on the RDT to the point where blood would seep into the result window and lines could not be discerned. End-users can easily recognize a bloody RDT without any assistance, but a means of automatically identifying these cases can help rule out algorithmic false positives or expedite manual review by back-end supervisors. To detect blood within the result window, RDTScan applies the following filter to isolate red pixels: $H \in [0, 10] \cup [160, 179]$, $L \in [100, 255]$, $S \in [100, 255]$. An RDT is deemed to have too much blood if over 25% of the result window's area is red.

RDT Result Interpretation

After cropping and enhancing the result window, RDTScan computes the average intensity value (L channel) for each row parallel to the expected orientation of the lines themselves. Those intensity values are plotted against their position along the result window to represent how the intensity varies along the underlying immunoassay (Figure 3.2). Lines are defined as local minima that are at least 60 units deep and cover at least 5% of the result window's main axis. Because the metadata includes the location of the different lines, RDTScan uses that information to both confirm line validity and to assign semantic meaning to them. In some cases, visual noise can manifest at the line locations in a way that creates a trough in the intensity plot. If the metadata includes the expected range of color hue values (H channel) for that line, RDTScan extracts the average hue of the row, and lines with an average hue outside of the expected range are rejected. Otherwise, RDTScan only checks the intensity values for result interpretation.

3.3.5 User Interface

Passing all of the quality assurance checks requires having vigilant end-users who are able to position their camera in the optimal manner during image capture. As shown in Figure 3.3, RDTScan provides an overlay that lies on top of a camera preview while end-users move their smartphone. This overlay includes a viewfinder with the same aspect ratio as the target RDT design and is scaled according to the metadata provided by the developer. RDTScan also displays a list of the quality checks that the most recently processed frame failed. By showing all of the checks simultaneously, end-users can identify regularly occurring or concurrent issues. At the same time, end-users may want to know what steps need to be taken to pass those checks and increase their chance of a successful capture. RDTScan provides human-readable instructions that describe how end-users should adjust their smartphone to capture the best image possible. Only one instruction is provided at a time, and the instructions are prioritized to minimize user effort. The feedback order is as follows:

1. **Image brightness and shadows:** Since end-users may need to change their environment to improve the ambient lighting, these instructions are prioritized first. If the image is overexposed and the flash is on, end-users are instructed to turn off the flash; likewise, end-users are instructed to turn on the flash if it is off and the image is underexposed. If the image is overexposed and the flash is already off, end-users are instructed to decrease the ambient lighting; likewise, end-users are instructed to increase ambient lighting if the flash is already on and the image is underexposed.

2. **Blur:** Image blur is often indicative of a person moving their camera, so it does not make sense to check the other characteristics of the camera frame until the camera is back in focus. If blur is detected, RDTScan simply notifies end-users that they should hold the phone still until the camera can refocus itself.

3. **RDT size:** If the camera is at an improper distance from the RDT, RDTScan may not have enough information to detect an accurate bounding box around the RDT to

correct its other properties. RDTScan instructs end-users to move their camera towards or away from the RDT depending on whether the RDT appears too small or too big relative to the viewfinder.

4. **RDT position and orientation:** Once the camera is at the proper height, translation and rotation suggestions are more likely to be accurate since all of the RDT's corners will be visible and well-spaced. RDTScan provides guidance for camera movements that are parallel to the RDT (i.e., translation and rotation) until the RDT is aligned with the viewfinder.

5. **Glare:** At this point, the position of the camera relative to the RDT should be nearly optimal. However, glare can be created by the flash or ambient lighting due to the angle of the camera relative to those light sources. If glare is detected, RDTScan instructs end-users to slightly tilt their smartphone at an angle until the glare disappears.

Once the current camera frame passes all of the quality checks, RDTScan automatically stores the image and advances so that the user does not have explicitly capture the photograph.

3.4 System Evaluation on Low-End Smartphones

Before deploying RDTScan to CHWs and their supervisors, we first evaluated its performance in terms of processing time, memory footprint, and power consumption.

Processing time: For this analysis, we break down the RDTScan algorithm into two steps: hardware configuration (i.e., starting the camera and setting its parameters) and software (i.e., real-time quality checks, template matching). We measure the processing time for each step on the three devices that were used in our studies: the Tecno WX3, the Tecno W3, and the Tecno POP2. To control other factors that can affect processing time, WiFi and Bluetooth were turned off and all other apps were closed. We ran RDTScan ten times on each device with an RDT cassette in the camera frame to ensure that each quality check would be ran. Table 3.2 summarizes the processing time for camera setup and each of the following

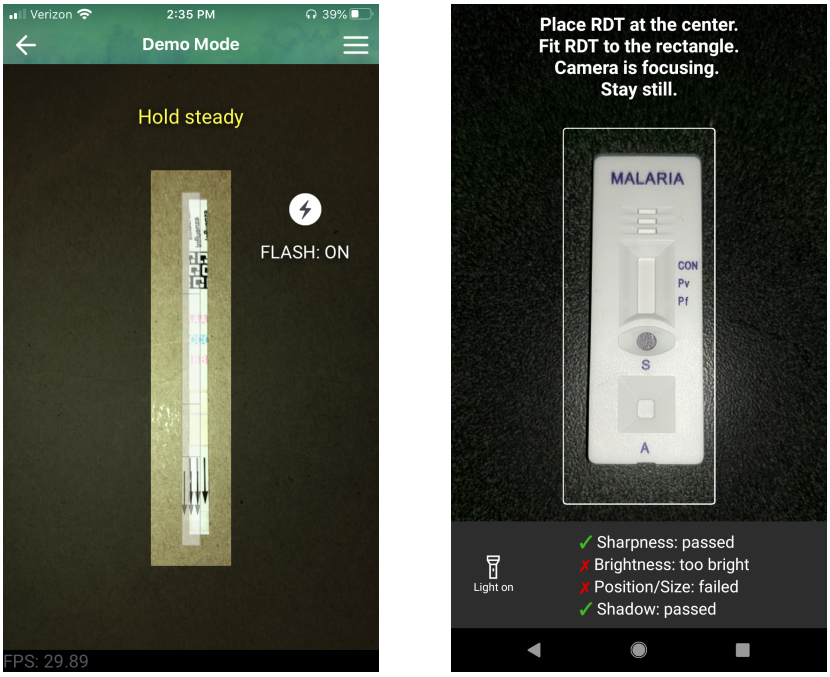


Figure 3.3: Screenshots of the two apps created by our collaborators using RDTScan: (left) flu@home for the Australia Influenza Study and (right) RDT Open Reader for the Kenya Malaria Study.

software procedures: exposure checking, blurriness checking, and feature matching. We found that configuring the camera hardware takes 5 seconds, which is important to consider when analyzing capture time from the user's perspective later in the paper. For the computer vision computations, feature matching took 10 times as long as the exposure and blurriness checks. In aggregate, those operations took 340–520 ms depending on the smartphone, which translates to 2–3 frames per second.

Memory usage: We also evaluated RDTScan's memory usage with Android Studio's Memory Profiler tool. Similarly to our measurements of processing time, we ran the app ten times with all three smartphones. On average, the app used around 110 MB. This memory usage only happens during the capture process, which we limit to 30 seconds to avoid deadlocks in CHWs' workflows. Given that the smartphones we tested have 1 GB RAM

Table 3.2: A summary of system performance for three different low- to middle-end Android smartphones.

	Devices		
	POP2	W3	WX3
Proc. time - Setup	5.12 s	4.46 s	4.37 s
Proc. time - Exposure	15 ms	18 ms	17 ms
Proc. time - Blur	28 ms	36 ms	27 ms
Proc. time - Matching	300 ms	460 ms	310 ms
Memory Footprint	113.8 MB	104.9 MB	107.4 MB
Battery Drain Rate	846.9 mA 21.17%/hr	1033.7 mA 41.35%/hr	990.8 mA 39.63%/hr

and the operating system uses around 500 MB, the app's memory footprint does not put too much overhead on the device.

Battery drain: Because CHWs are out in the field when they administer RDTs, it is important to consider the power consumption that the app incurs. To measure this, we ran the app on each smartphone for an hour, forcing them to run the computer vision operations with the flash on the entire time. All other apps and features (e.g., Bluetooth, Wi-Fi, GPS) were turned off during this test, and the screen brightness was fixed to 50% to emulate screen conditions in a standard lighting environment. We analyzed the resulting battery drain rate using Android's Battery Historian.

As shown in Table 3.2, the Tecno POP2 lost 21% of its charge in one hour, while the Tecno W3 and WX3 lost 41% and 40% of their charge, respectively. This statistic is a bit misleading because Tecno POP2 is a newer smartphone model and thus has more battery capacity than the other smartphones (4,000 mAh vs. 2,500 mAh). The Tecno POP also consumed less current than the others, which can be attributed to improved power management in later builds of Android. Regardless, this evaluation shows that low-end smartphones with

4000 mAh battery capacity can run the app for around 6 hours continuously, which is far longer than what CHWs need for a single day.

3.5 Interpretation Comparison Between Human Expert and Low-End Smartphone

To evaluate RDTScan, we partnered with a non-profit organization in Mali called Muso. Among their other services, Muso employs CHWs who regularly perform malaria testing using RDTs for doorstep care to community members. Before putting RDTScan in the hands of CHWs in the field, we set out to determine whether photographs taken using our app would serve as suitable proxies for physical RDTs. To this end, we carried out a non-inferiority trial with lab technicians to confirm that readings from our app would not be unacceptably worse than direct readings by experts from physical RDTs.

3.5.1 Participants

Two lab technicians with expertise in malaria diagnosis were recruited from the Malaria Research Training Center (MRTC) in Bamako. The first technician was a 47-year-old male with 15 years of experience reading RDTs, while the second technician was a 37-year-old male with 10 years of experience.

3.5.2 Procedure and Apparatus

Before patient recruitment began, the two technicians independently practiced reading malaria RDT cassettes. The technician were guided by a third and more experienced technician to ensure systematic errors in readings were not being made. After the independent training, the technicians came together and discussed any discrepancies between their reading technique to ensure their readings were consistent.

New RDTs were activated by clinicians in Yirimadio, a peri-urban area on the outskirts of Bamako. Patients who presented symptoms of malaria or any signs of severe illness that would normally be tested for malaria had their blood drawn and applied to RDTs. After

the 20-minute wait-time required for RDT activation, one of the technicians recorded their *direct reads* of the RDT cassette in a spreadsheet. Within 15-30 minutes, a research staff member took a picture of the RDT using the RDT capture app installed on one of two Tecno WX3 smartphones. The WX3 has a 5-inch screen, 854×480 pixel display, and 5 MP camera. The smartphones were used in such a way that each lab technician had a smartphone that stored images of the RDTs they interpreted. The research staff ensured that there were no external cues present in the image that could make the RDT uniquely identifiable or reveal its test result to the technicians. Image capture was conducted in a room with similar lighting conditions as the rooms where the lab technicians were reading the physical RDT cassettes.

Each lab technician interpreted the images to produce an *original image read* for each RDT. The images were shuffled to avoid ordering effects and ensure that the sequence did not make specific images recognizable. Two months later, when the enhancement algorithm was developed, the images were post-processed, re-shuffled, and then read again by the same people to produce corresponding *enhanced image reads*.

3.5.3 Analysis

Interpretation by Experts

We assessed RDTScan's performance by measuring the concurrence between direct reads of the RDT cassettes and the various interpretation mechanisms enabled by RDTScan: original image reads, enhanced image reads, and automatic interpretation. We calculated concurrence using sensitivity (true positive rate) and specificity (true negative rate) along with their 95% confidence intervals (CIs). Direct reads were treated as the ground truth since that is the current best practice for many organizations in low- and middle-income settings. RDTs were excluded from our analysis if the test results were inconclusive (e.g., no control line, excessive smudging).

A total of 795 images were captured at the clinic, of which 107 were positive, 668 were negative, and 20 were indeterminate according to direct reading; therefore, our analysis was

Table 3.3: Confusion matrices showing how the interpretation results from (top) original image readings, (middle) enhanced image readings, and (bottom) automatic analysis by an algorithm compare against direct readings of the physical RDT cassettes.

Original Image Reads	**Direct Reads** Pos (+)	Neg (-)	**Total**
Pos (+)	100	1	101
Neg (-)	7	667	674
Total	107	668	775

Enhanced Image Reads	**Direct Reads** Pos (+)	Neg (-)	**Total**
Pos (+)	105	2	107
Neg (-)	2	666	668
Total	107	668	775

Automatic Interpretation	**Direct Reads** Pos (+)	Neg (-)	**Total**
Pos (+)	103	6	109
Neg (-)	4	662	666
Total	107	668	775

conducted on 775 images. The underlying prevalence of malaria in our study population was 13.8%.

The top of Table 3.3 compares original image reads and direct RDT reads. Interpreting the unmodified images taken by RDTScan resulted in a sensitivity of 93.5% (CI: [87.0%, 97.33%]) and a specificity of 99.9% (CI: [99.2%, 100%]). The higher specificity over sensitivity can be attributed to the fact that some of the positive RDTs had faint lines that were even more difficult to read when they were digitized. The middle of Table 3.3 shows the concurrence between the enhanced image reads and the direct reads. This led to a sensitivity of 98.1% (CI: [93.4%, 99.8%]) and a specificity of 99.7% (CI: [98.9%, 100%]). The sensitivity improved by roughly 5% because contrast enhancement made faint lines more obvious and thus reduced false negatives. The false positives resulted from image noise that formed at the test line location after contrast enhancement. The false negatives were due to lines that were faint even after contrast enhancement.

Interpretation by an Algorithm

The bottom of Table 3.3 compares the performance of automatic interpretation against direct reads. Our algorithm was comparable to expert reads, achieving 96.3% (CI: [90.7% to 99.0%]) sensitivity and 99.1% (CI: [98.1% to 99.7%]) specificity. Compared to the enhanced image reads, there were two more false negatives and four more false positives, which could be attributed to two factors: (1) noise generated after the contrast enhancement and (2) blood within the result window. Blood can create false positives or false negatives depending on how it stains the RDT strip. If the stain is throughout the strip, any lines on the strip are rendered indistinguishable from the background; if the stain is localized near the test line, the blood can mimic a positive strip result. Both cases occurred in our dataset, generating a few more false negatives and positives than enhanced image reads.

3.6 Capture Performance in the Wild

Having demonstrated strong concurrence between direct reads of RDT cassettes and images from RDTScan, we deployed our app to CHWs in the field. The purpose of this deployment was to assess RDTScan's usability and efficacy in real-world environments and to collect

feedback for future app iterations.

3.6.1 Participants

We recruited supervisors and CHWs from two program sites in Yirimadio and Tori. Two supervisors and six of their subordinate CHWs were randomly selected from each clinic. The 12 CHWs who participated in our study (1 male, 11 females) were between 23–53 years old and had varying levels of experience reading RDTs (2 months–6 years). These CHWs and their supervisors use smartphone apps built using the Community Health Toolkit[13] in their routine activities, so they were deemed to have a relatively high degree of readiness to adopt new smartphone apps for their work.

3.6.2 Procedure and Apparatus

The deployment study was run for ten weeks between May–August 2019. Before the study started, a half-day training session was held to introduce CHWs to our RDT capture app. The session covered instructions on how to use the app, tips for optimal capture, and study logistics. We had an additional training session for the supervisors on how to best support the CHWs and report app issues to the research team. Each CHW was handed a smartphone pre-installed with RDTScan. Six Tecno POP2 (5.45-inch screen, 960×480 px display, 5 MP camera) and ten Tecno W3 (5-inch screen, 854×480 px display, 5 MP camera) smartphones were used in the study. CHWs were asked to capture every RDT they conducted in the field for the entire study duration. CHWs ran a different number of RDTs and thus produced different numbers of captured images. The images were automatically saved in the smartphone with metadata on the CHW's identity, their smartphone model, a timestamp, and the time taken for RDT capture. To ensure that RDTScan did not significantly disrupt the CHWs' workflow, we implemented a 30-second timeout within the app. CHWs who were able to get a successful automatic capture in that time were taken to a screen that showed the enhanced

[13] `https://communityhealthtoolkit.org`

result window alongside the original image. If an RDT was not successfully recognized before the timeout, an image of whatever was currently in the camera's view was captured. After many discussions with the clinics, we decided to not reveal the automatic interpretation results to the CHWs since doing so could have significantly altered diagnostic outcomes while the app was still under development.

3.6.3 Analysis

For the deployment study, we focused on analyzing the usability of RDTScan both quantitatively and qualitatively. The main quantitative measures were capture time and automatic capture rate. Two researchers reviewed the captured images to assess whether the images had sufficient quality for supervisors to review the RDTs' test results. We also conducted semi-structured interviews to get qualitative feedback from CHWs on RDTScan's usability.

We collected 533 images over the course of the study, with each user capturing 44 images on average (min: 3, max: 79). A subset of images were excluded from our analyses either because the CHWs did not follow our instructions for how to administer the RDT or the RDTs were not correctly captured by RDTScan. That subset includes 145 images where CHWs wrote on the cassettes, 11 images with RDTs that were upside down, 13 images with a smudged or covered camera, and 3 images with contaminated RDTs. We note that RDTScan is robust to writing on the cassette as long as it does not overlap with the RDT's most prominent visual keypoints, but RDTs with writing were excluded unilaterally for consistency. Therefore, we analyzed 361 images were analyzed in total.

Overall Capture Performance

To assess RDTScan's ease of use, we calculated the average capture time and proportion of images captured before the 30-second timeout. Table 3.4 shows the overall capture performance in the deployment study. On average, CHWs took around 20 seconds to capture an RDT image. Automatic capture was triggered 67% of the time before the timeout, and the average capture time was 14.4 seconds in those cases. Our system performance evaluation

Table 3.4: The average capture time and automatic capture rate across all CHWs during the deployment.

	Capture Time		**Automatic**
	w/ Timeout	w/o Timeout	**Capture Rate**
All Trials	20.3 s	14.3 s	67%
Trial 1-10	21.8 s	13.7 s	58%
Trial 11-20	19.8 s	14.9 s	72%
Trial 21-30	18.3 s	14.9 s	78%

revealed that the smartphones took 5 seconds on average to configure the camera's hardware, thus accounting for one-quarter of the average capture time. In other words, CHWs only spent 15 seconds interacting with the app once the camera was ready. Within that time, CHWs usually spent 3–5 seconds positioning the camera relative to the RDT, and the remaining 10 seconds was spent waiting for a frame that would pass all of the quality checks.

Image Quality

All captured images were assessed by two researchers along three criteria: size, position, and brightness. Any disagreements were resolved after by the researchers coming together and discussing the relevant images. In the end, we found that all 354 captured images clearly showed the result window, which is the most critical piece of information during review. However, there were 27 images that did not clearly show the entire RDT; 4 images were over-exposed, 13 images were taken too close, 10 images were taken too far, and 3 images were off-center. Although CHWs exceeded the timeout 33% of the time, the images that were captured at the end of that period typically passed our image quality criteria. We suspect that although the template-matching algorithm may have failed to identify the RDT, the viewfinder interface and real-time guidance helped CHWs position the RDT properly relative

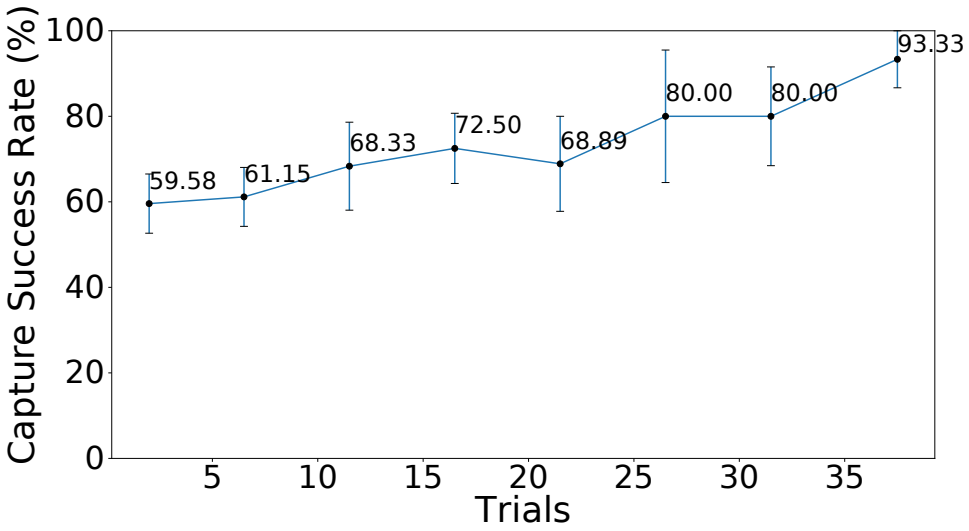


Figure 3.4: The change in automatic capture success rate over time. The error bars show standard error.

to the camera.

Capture Performance Improvement

The deployment study lasted for over 3 months. Since the CHWs were new RDTScan users, we were interested in investigating how quickly they were able to learn how to use the app. We first compared the automatic capture rate and average capture time during their 1^{st}–10^{th}, 11^{th}–21^{st}, and 21^{st}–30^{th} trials. As the bottom two rows in Table 3.4 show, the automatic capture rate improved from 58% in the first 10 trials to 72% to 78% in second and third set of 10 trials, respectively. We found that the average capture time for automatic capture stayed the same, suggesting that the CHWs were able to capture RDTs fairly consistently.

We further analyzed how capture success rate changed over time. Figure 3.4 shows this trend across all CHWs. The graph shows a steady improvement in automatic capture rate as the CHWs gained more experience with RDTScan. After 25 captures, CHWs were successful more than 80% of the time; after 30 trials, that number rose further to 93%. We

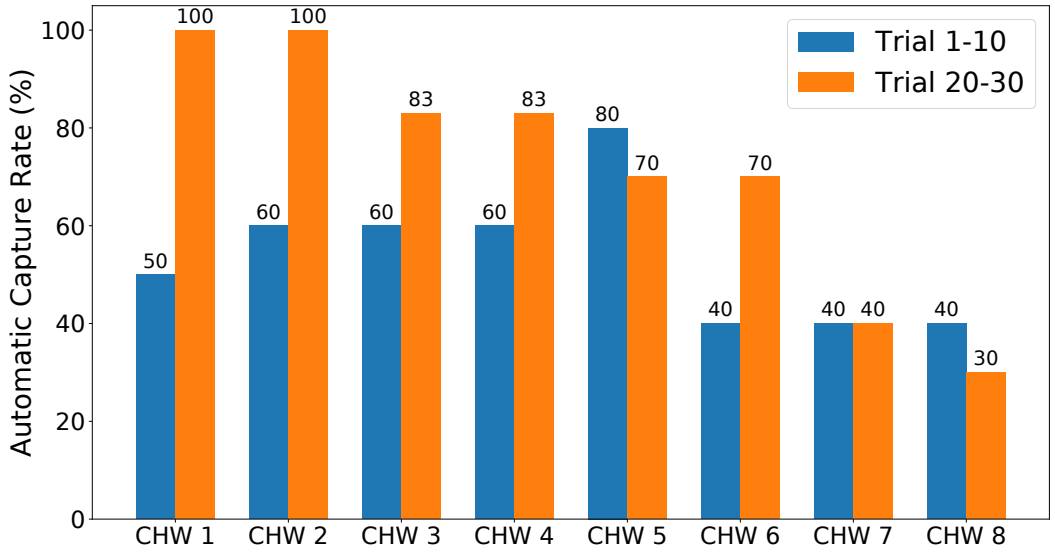


Figure 3.5: The change in automatic capture success rate over time separated by CHWs. Most of the CHWs were more successful using RDTScan over time, with some exceptions being the CHWs who had the least success early on.

further analyzed whether there were differences between different CHWs. For each CHW, we compared their automatic capture rate between their first 10 trials and trials 20–30. One CHW was excluded from this analysis because they only used the app three times. Figure 3.5 summarizes the capture rate for the eight remaining CHWs. Five of the CHWs improved over time, with two people reaching 100% by their 20^{th} trial. Unfortunately, CHWs who started with poor capture rates were more likely to not improve. In our study, supervisors did not provide additional training or support apart from issues like app crashes; in practice, supervisors could provide struggling CHWs with additional guidance to avoid repeated issues.

Qualitative Feedback

We conducted semi-structured interviews with both the CHWs and their supervisors. For the CHWs, the questions were mainly around the usability and utility of RDTScan, along with

any difficulties they experienced during the study. For the supervisors, the questions involved the utility of the app in their RDT workflow. The findings are summarized below.

Improved workflow: The overarching objective of RDTScan is to improve the RDT workflow of CHWs in the field. All of the CHWs agreed that RDTScan made their jobs easier since it voided the need for them to physically return RDT cassettes to clinics for review. Two CHWs pointed out that RDTScan enabled instant feedback from their supervisors when images could be uploaded, which helped them make a prompt decision on whether to treat the patient or readminister the RDT.

Although the CHW supervisors were not direct RDTScan users, the app still brought value to their workflow. One supervisor mentioned that the app eliminated the work he had to do to collect, store, and dispose of all the RDTs returned by his CHWs. The app also reduced the uncertainty that was caused when supervisors received RDTs that may have deteriorated over time. With RDTScan, the images effectively freeze the state of the RDT in time to make it as if the supervisor is with the CHW when they conduct their review. When images can be uploaded immediately after an RDT is administered, supervisors said they could provide quick feedback that would help CHWs decide whether to prescribe an anti-malarial, readminister the RDT, or clear the patient.

Accurately representing RDTs: Aligning with the findings of our first study, all of the CHWs agreed that the captured images were accurate representations of the RDT cassettes. The supervisors also believed there was no substantial difference between the images and the actual RDTs, with one person saying, "[I] have confidence in the RDT image tool as the images are clear and the result readable."

Difficulties: We asked about any difficulties or issues using RDTScan. Most of the CHWs had little experience taking photos with a smartphone, even for their personal use, so the experience of using RDTScan was completely new to them. One CHW pointed out that it was difficult for her to follow the instructions while trying to get all the quality checks passed when she first used the app, but she had more success as she became more experienced with the app. This remark aligns with the previous quantitative analysis that shows improved

automatic capture rates over time. Other CHWs commented that the battery drained quickly over the course of their workday. We believe that the novelty of camera-based smartphones played a small role in this, leading CHWs to underestimate the amount of battery life required to use the smartphone's camera and flash multiple times a day. Nevertheless, battery usage is still a major concern that cannot be ignored when designing for CHW workflows. Most of the images after the 30 second timeout were high-quality, thanks in part to the app's instructions and interface, so we hypothesize that a shorter timeout could avoid long capture times without sacrificing image quality.

3.7 In-Lab Evaluation

Before we released RDTScan, we tested our interpretation algorithm in a controlled lab environment. This evaluation examined our algorithm's limit-of-detection: the lowest analyte concentration level at which a positive result can still be recovered from the RDT. We compared the limit-of-detection across different RDTs, analyte concentrations, smartphones, and lighting conditions according to three interpretation methods: inspection of the RDT directly, inspection of the RDT image, and automatic interpretation.

3.7.1 Procedure

Three researchers with experience in administering RDTs created two dilution series using analyte proxies for the RDTs that would eventually be used in our case studies: malaria *Pv/Pf* protein mixed with deionized water for the Carestart RDTs and extracted influenza A protein for the QuickVue RDTs. The sample-to-buffer ratios for the CareStart series spanned 1:1–1:200, whereas the QuickVue series spanned 1:1–1:1000. These ranges were empirically selected so that a spectrum of line intensities would be created. To account for possible variables in the manufacturing of the immunoassays, the researchers used each analyte concentration to activate two RDTs. The researchers also activated two RDTs without any analyte to serve as negative controls.

The researchers first interpreted the physical tests under white fluorescent ambient

lighting (DIRECT_READ). The researchers then took photos of the RDTs using an app built with RDTScan on two smartphones: a Google Pixel 1 and a Samsung Galaxy S8. The researchers took photos with both smartphones under four lighting sources: fluorescent light, warm white LED (2700 K), cool white LED (5000 K), and ambient light (distant fluorescent + sunlight). The automatic interpretation decisions returned by RDTScan were saved as the algorithm readings (ALGO_READ). The captured images were saved and then pushed through RDTScan's detection and interpretation pipeline up until contrast enhancement to generate an enhanced image of the result window.

Later that day, the same three researchers interpreted the RDT results in the unmodified photographs (IMAGE_READ) so that it could be understood how much information was lost during the image capture process alone. After that, the researchers interpreted the enhanced images that were produced by RDTScan (ENHANCED_IMAGE_READ). For each interpretation method, the RDTs were shuffled and the researchers made their decisions independently to avoid any biases. At the end of this process, the researcher's decisions were aggregated using majority consensus to form a single decision per RDT and per interpretation method.

To summarize, the four interpretation methods in order of increasing digitization and automation are DIRECT_READ, IMAGE_READ, ENHANCED_IMAGE_READ, and ALGO_READ. For each RDT brand, there were 2 DIRECT_READs, 16 IMAGE_READs (2 trials $\times$ 2 smartphones $\times$ 4 lighting conditions), 16 ENHANCED_IMAGE_READs, and 16 ALGO_READs.

3.7.2 Results

Inter-rater reliability was moderate amongst the three researchers according to Fleiss' kappa ($\kappa = 0.69$) [99]. Table 3.5 compares the four different interpretation methods. Note that for all interpretation methods, the same decisions were reached for two RDTs at the same concentration. We consider DIRECT_READ to be the baseline for our analysis since the reported limits-of-detection for both tests are inapplicable to the analyte proxies we used. We compare interpretation methods using McNemar's test [123] when there is a sufficient number of samples to do so (e.g., at least 25 samples with different results from the two methods);

Table 3.5: A comparison of the four different interpretation methods for the (left) QuickVue and (right) QuickVue RDTs. Each box indicates the variance across lighting conditions. The horizontal lines indicate the limit-of-detection for the different interpretation methods.

QuickVue

Dilution	DIRECT_READ	IMAGE_READ	ENHANCED_IMAGE_READ	ALGO_READ
1:1	+	16/16	16/16	16/16
1:10	+	16/16	16/16	16/16
1:100	+	8/16	12/16	16/16
1:200	+	4/16	12/16	16/16
1:400	+	4/16	4/16	16/16
1:600	+	0/16	4/16	16/16
1:800	+	0/16	0/16	9/16
1:1000	–	0/16	0/16	0/16
0:1 (control)	–	0/16	0/16	0/16

CareStart

Dilution	DIRECT_READ	IMAGE_READ	ENHANCED_IMAGE_READ	ALGO_READ
1:1	+	16/16	16/16	16/16
1:5	+	16/16	16/16	16/16
1:10	+	14/16	16/16	16/16
1:20	+	6/16	16/16	16/16
1:40	+	4/16	10/16	16/16
1:60	+	0/16	6/16	16/16
1:80	–	0/16	0/16	8/16
1:100	–	0/16	0/16	0/16
1:200–1:1000	–	0/80	0/80	0/40
0:1 (control)	–	0/16	0/16	0/16

otherwise, we use an exact binomial test. For the QuickVue RDT, `IMAGE_READ` performed worse than `DIRECT_READ` ($\chi^2(1) = 54.018, p < .001$), highlighting the fact that camera sensors can diminish small details during digitization. `ENHANCED_IMAGE_READ` was also worse than `DIRECT_READ` ($\chi^2(1) = 46.021, p < .001$) but outperformed `IMAGE_READ` ($Bin(24, 0.5) = 0.000, p < .001$). Contrast enhancement emphasized subtle details, which lowered the limit-of-detection and increased sensitivity. `ALGO_READ` performed comparably to `DIRECT_READ` ($Bin(8, 0.5) = 0.000, n.s.$), showing that the combination of contrast enhancement and an automated peak-finding algorithm were able to overcome the information that was lost when the RDT was converted into an image.

The different interpretation methods were more similar to one another for the CareStart RDT. As before, `ENHANCED_IMAGE_READ` detected more positive cases than `IMAGE_READ` ($Bin(24, 0.5) = 1.000, p < 0.01$). `ALGO_READ` was able to detect lower concentrations than `DIRECT_READ`, although not to a statistically significant degree ($Bin(8, 0.5) = 0.000, n.s.$). One might argue that the `ALGO_READ` positive results at the 1:80 dilution could have been RDTScan misinterpreting image noise as a line that was not actually present. If that was the case, however, RDTScan would have also produced positive results at even lower dilutions. Therefore, it is more likely that RDTScan was able to detect lines that researchers missed on the CareStart RDTs.

Figure 3.6 shows a box plot of the intensity differences at the RDTs' test line locations. As a reminder, RDTScan detects lines that have a trough depth greater than 60 units. The variance of the boxes captures the different trough depths that were measured across lighting conditions. As expected, the trough depth was large when the analyte concentration was high, and the depth decreased at lower concentration. Contrast enhancement not only significantly increased the trough depths to make faint lines visible, but also increased the resolution between different concentrations. In fact, contrast enhancement increased the trough depth of the negative controls since it intensified visual noise on the surface of the immunoassay. To avoid potentially classifying these cases as diagnostic positives, RDTScan's interpretation algorithm checks trough width (and potentially hue) to distinguish real lines from noise.

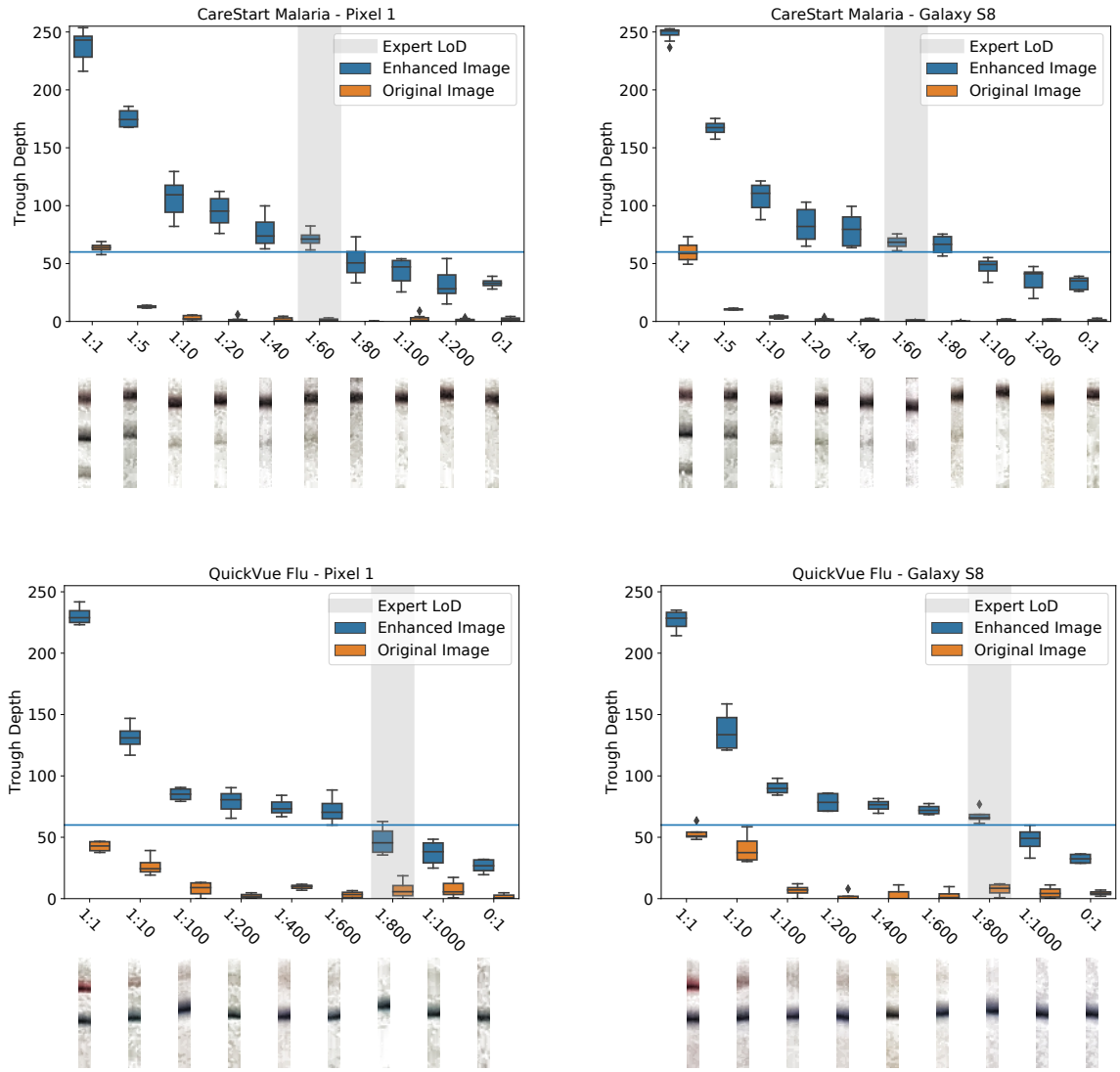


Figure 3.6: The magnitude of the intensity troughs that were measured by RDTScan across the different concentrations. The results are split according to RDT brand (rows) and smartphone models (columns). The variance within each bar aggregates the results across different lighting conditions. Note that the horizontal axes are not to scale given that the respective tests have different sensitivities. Images on the x-axis are examples of contrast enhanced result lines used for interpretation.

There were no statistically significant differences in `ALGO_READ` decisions across the different lighting conditions or RDTs with the same concentration, as illustrated by the low variances in Figure 3.6. This result demonstrates the robustness of RDTScan across environments without the need for accessories or hardware.

3.8 Field Evaluations

Although controlled lab testing was instrumental for characterizing RDTScan's interpretation limits, it only represents the upper limits of RDTScan's performance in the field. RDTScan automates much of the quality control process during image capture, but the fact that researchers with RDT image capture experience handled the smartphones during that study would be assumed to boost the quality of each image. Furthermore, analyte concentrations in a biological sample vary depending on the duration of the infection, the severity of the infection, the site from which the sample was taken, etc. Therefore, deploying RDTScan was the best way to assess its performance in real-world settings.

In this section, we present two field evaluations: at-home influenza testing in Australia and in-clinic malaria testing in Kenya. For each deployment, we describe how our collaborators modified RDTScan for their needs, the demographics and context of the participants, the study procedures, and the performance of RDTScan itself.

3.8.1 Australia Influenza Study

Motivation & Context

Influenza causes seasonal epidemics and occasional pandemics that impact the entire world. As with many other infectious diseases, it is believed that early detection can reduce transmission and improve the outcomes of patients who are at higher risk of complications [28]. At-home influenza testing could reduce the need for patients to travel to a healthcare facility for diagnosis confirmation, enhance the efficiency of healthcare delivery during busy influenza seasons, and even reduce the risk of transmission during a pandemic. In most countries,

there are currently no diagnostic tests for influenza that have regulatory approval for use by untrained individuals outside of clinical settings. We partnered with a group of clinicians and smartphone app developers who were interested in assessing the potential efficacy of Quidel's QuickVue RDT for at-home influenza testing in Australia, where there are an estimated 310,000 general practitioner consultations and 18,000 hospital admissions related to influenza each year [137].

Study App Integration

The existing study protocol included a smartphone app called fluhome (Figure 3.3, left) for an optional home-testing component to a broader study on the spread of influenza in Australia. Our collaborators wanted their app to help participants take a clean photo of their QuickVue RDT so that the test's results could be compared to a clinical gold-standard with minimal reporting errors. There was great interest in the possibility of using an app for automatic interpretation to eventually remove that element of human error; however, there were ethical concerns about showing those results to end-users without regulatory approval. Therefore, their app ran RDTScan's automatic interpretation algorithm in the background and uploaded the resulting decision and image to a back-end server without showing the results to the end-users.

To mitigate participant frustration and ensure that RDTScan did not significantly impact the rest of the study protocol, we added a 30-second timeout for RDTScan within the fluhome app. The app automatically advanced if RDTScan did not successfully detect the participant's RDT during that time, regardless of whether the issue was due to human error or RDTScan's inability to detect the RDT. During pilot testing, we found that RDTScan occasionally had issues detecting the QuickVue RDT due to its thin and inconsistent design. Therefore, we utilized the optional fiducial checking step to ensure proper detection.

Deployment Site and Participants

Study participants were recruited from clinics participating in the Australian Sentinel Practices Research Network (ASPREN)[14]. Clinics within ASPREN share de-identified information on influenza-like illness and other infectious disease conditions as part of national surveillance efforts in Australia each year. Adult patients who visited a clinic participating in ASPREN while exhibiting influenza-like symptoms were invited by their general practitioner (GP) to participate in the study. Details about the study can be found in the protocol published by Lyon et al. [113]. In total, 299 participants enrolled in study (196 female, 103 male) with an average age of 41 $\pm$ 13.9 years.

Procedure

The study was conducted from July 2019 to February 2020; although the typical influenza season in Australia covers these same months, the peak of that particular season happened in July [70]. GPs obtained nasal swabs from participants during clinic visits, which were sent to a reference laboratory for reverse transcription quantitative polymerase chain reaction (RT-qPCR), serving as the gold-standard indicator of influenza status (`PCR_RESULT`). Participants who opted into the home-testing protocol were then sent home with a custom kit that included the materials needed to operate Quidel's QuickVue RDT and a link to download fluhome on their smart-device (Android or iOS smartphone or tablet). Participants were asked by their GP to complete the RDT using the instructions provided in the app. After completing the RDT, participants used the app to take a photograph of the dipstick. In the app, participants completed a survey to indicate if they saw a blue control line on their dipstick, as well as any red or pink test lines on their QuickVue RDT (`DIRECT_READ`). Usability was measured according to capture success rate within the 30-second timeout and the average capture time when participants had a successful capture. The captured images were analyzed in the background using RDTScan (`ALGO_READ`), and they were also later reviewed by an

[14] `http://www.aspren.com.au/`

Table 3.6: A summary of RDTScan's usability during the Australia Influenza Study. Capture success rate quantifies how often participants were able to use RDTScan to get a high-quality photograph of their RDT within the 30-second timeout, while capture time measures the average time it for a successful capture.

Mobile OS	**N**	**Capture Success Rate**	**Capture Time (s)**
iOS	175	88.0%	8.73
Android	124	75.8%	11.52
Age	**N**	**Capture Success Rate**	**Capture Time (s)**
≤ 29	71	90.1%	9.06
30–39	65	96.9%	9.71
40–49	53	81.1%	9.56
50–59	52	72.6%	10.62
≥ 60	34	58.8%	11.00
Total	299*	83.3%	9.77

*A total of 275 participants reported their age, while the rest chose to not disclose that information.

independent researcher who underwent a half-day training session on RDT interpretation to serve as a bronze-standard (`EXPERT_IMAGE_READ`).

Results: Usability

Table 3.6 summarizes the usability metrics across different platforms and age groups. Overall, participants were successful 83.3% of the time in capturing an image of their RDT using RDTScan within the 30-second timeout. When participants had a successful capture, they typically did so in under 10 seconds. Capture success rate was significantly higher amongst iOS users than Android users ($\chi^2(1) = 7.598, p < .01$), and capture time was significantly

lower ($t(245) = -3.161, p < .01$) for iOS users. Although there may be confounding factors due to the typical demographics of Android and iOS users, we hypothesize that this can be partly attributed to the processing power of the operating systems themselves. SIFT operations are roughly 4× slower on Android than on iOS [31], and Android throttles the CPU more aggressively than iOS; together, these differences cause a lower effective frame rate on Android and thus fewer cycles for real-time feedback. Furthermore, there is a broader diversity of Android device specifications, some of which are far inferior to the typical iPhone for the sake of lower prices. There were also noteworthy differences in capture performance between age groups. The capture success rate amongst participants under the age of 40 was greater than 90%, whereas the older participants had far more difficulties. We suspect that younger participants had more success using RDTScan because of their greater familiarity with smartphones and camera-based apps.

Results: Accuracy

A total of 249 images passed all of RDTScan's quality checks and were subsequently interpreted. Five of those images had inconclusive results, meaning that no control line appeared within the result window. For the remaining 244 images, we compared `ALGO_READ` against `EXPERT_IMAGE_READ` and `PCR_RESULT`. Table 3.7 shows the confusion matrices for both comparisons. When compared against the bronze-standard `EXPERT_IMAGE_READ`, RDTScan achieved 97.5% accuracy (sensitivity = 85.7%, specificity = 98.7%). When compared against the gold-standard `PCR_RESULT`, RDTScan achieved 83.6% accuracy (sensitivity = 33.9%, specificity = 98.4%). In evaluating how the other RDT interpretation methods compared to `PCR_RESULT`, `DIRECT_READ` had 76.6% accuracy (sensitivity = 35.2%, specificity = 91.2%), while `EXPERT_IMAGE_READ` had 82.6% accuracy (sensitivity = 32.9%, specificity = 100%). RDTScan demonstrated better accuracy than the novice end-users ($\chi^2(1) = 28.195, p < .001$) and comparable accuracy to the expert ($Bin(14, 0.5) = 6.000, n.s.$) when those readings were compared to `PCR_RESULT`. RDTScan and the other RDT interpretation methods had poor sensitivity when compared to the gold-standard. `PCR_RESULT` was generated from nasal swabs

Table 3.7: Confusion matrices showing the performance of RDTScan against (left) expert interpretation of RDT images and (right) RT-qPCR for the Australia Influenza Study. As points of comparison, `DIRECT_READ` and `EXPERT_IMAGE_READ` achieved 76.6% and 82.6% accuracy when compared to `PCR_RESULT`, respectively.

	`EXPERT_IMAGE_READ`		
`ALGO_READ`	Pos (+)	Neg (–)	
Pos (+)	18	3	PPV = 85.7%
Neg (–)	3	220	NPV = 98.7%
	SNS = 85.7%	SPC = 98.7%	ACC = 97.5%

	`PCR_RESULT`		
`ALGO_READ`	Pos (+)	Neg (–)	
Pos (+)	19	3	PPV = 86.4%
Neg (–)	37	185	NPV = 83.3%
	SNS = 33.9%	SPC = 98.4%	ACC = 83.6%

SNS = sensitivity, SPC = specificity, ACC = accuracy, PPV = positive predictive value, NPV = negative predictive value

collected by trained GPs, whereas the other three readings were generated from self-obtained nasal swabs that were collected by the study participants a few days later. Therefore, the low sensitivity of the QuickVue RDT can be attributed to poor swabbing technique, lower viral load at the time of testing, or improper RDT administration at home. Even when all of the instructions have been properly followed, researchers have witnessed low sensitivity

with the QuickVue RDT. For instance, Agoritsas et al. [4] found that QuickVue RDTs had a sensitivity of 78% in a study where nasal swabs were collected from pediatric patients by research nurses. With our collaborator's study design, we are unable to separate out swabbing and test sensitivity issues.

3.8.2 Kenya Malaria Study

Motivation & Context

The World Health Organization estimates that there were 228 million cases of malaria in 2018 [203]. Malaria is endemic to many regions that have limited resources [61], so populations within those regions rely on community health workers (CHWs) to serve their healthcare needs. CHWs often travel from patient-to-patient, making RDTs a convenient point-of-care solution for malaria diagnosis. CHWs can have competing demands on their time, leading to occasional mistakes and rushed decisions [172]. We partnered with a Kenya-based software company who wanted to support CHWs in the field during their malaria testing duties with the CareStart RDT.

Study App Integration

The protocol for this deployment included a smartphone app called RDT Open Reader, which is designed to provide CHWs with instructions as they administer CareStart RDTs. Similar to the Australia Influenza Study, RDTScan was used to document RDT completion and the automatic interpretation results were hidden from the CHWs. We applied the same 30-second timeout to ensure that RDTScan did not significantly impact CHWs' typical workflow. Pilot testing revealed that our detection algorithm was sufficient for detecting the CareStart RDT without incorporating any of the optional checks (e.g., fiducials, line hues).

Deployment Site and Participants

The study was conducted at a health clinic in Kilifi, Kenya from December 2019 to March 2020. Six CHWs (5 male, 1 female) were responsible for administering RDTs and using the RDT Open Reader app (Figure 3.3, right). All of the CHWs had at least ten years of job experience, which included administering and interpreting RDTs. One CHW was in their 40s, while the rest were in their 30s. The CHWs shared four smartphones that were provided by the clinic: an Oppo A3s, an Infinix Hot 7, a Samsung Galaxy A10s, and a TECNO Pouvior 3 Air.

Procedure

Participating CHWs were trained to use the RDT Open Reader app during two half-day training sessions in a classroom setting, a half-day session in the field, and then a half-day refresher course a few days later. The CHWs were instructed to use the app whenever they provided services to patients presenting with malaria-like symptoms. After administering a CareStart RDT, the CHWs used the app to capture a photograph of it. The CHWs were not consistently supervised to avoid disrupting their duties. The CHWs administered and inspected RDTs in various location in clinic due to the mixed use of facilities, which means that the CHWs interacted with RDTScan in various lighting conditions and background settings. The CHWs were asked to record their interpretation of the physical RDT (`DIRECT_READ`), which was later compared to the results of RDTScan (`ALGO_READ`) and whole blood samples that were later processed through RT-qPCR (`PCR_RESULT`). Similar to the Austrialia Flu Study, usability was measured according to capture success rate within the 30-second timeout and the average capture time for a successful capture.

Results: Usability

Table 3.8 summarizes capture performance across different smartphone models. Since CHWs shared smartphones within the clinic and they were often under considerable time pressure

Table 3.8: A summary of RDTScan's usability during the Kenya Malaria Study. Capture success rate quantifies how often participants were able to use RDTScan to get a high-quality photograph of their RDT within the 30-second timeout, while capture time measures the average time it took for a successful capture.

Smartphone Models	**N**	**Capture Success Rate**	**Capture Time (s)**
Oppo A3s	120	99.2%	8.54
Infinix Hot 7	38	63.2%	10.99
Samsung Galaxy A10s	113	90.3%	6.95
TECNO Pouvoir 3 Air	62	96.8%	9.12
Total	332	90.4%	8.38

to perform their duties, we were unable to separate usability measures between individuals. Overall, the six CHWs were successful 90.4% of the time in capturing an image of their RDT using RDTScan within the 30-second timeout. When the CHWs had a successful capture, they typically did so in around 8.3 seconds. The CHWs had the least success with the Infinix Hot 7 smartphone ($\chi^2(1) = 47.131, p < .001$), which can be attributed to its inferior processing power relative to the other devices (1.3 GHz vs. 1.8–2.0 GHz CPU). In fact, when the smartphone models are sorted according to their processing power, that ordering aligns with the ranking of average capture time. As mentioned before, lower processing power results in a slower effective frame rate, which leads to fewer opportunities for user feedback. The Infinix Hot 7 also does not fully support Android's Camera2 API, but rather legacy support for camera hardware control. This limited support could have affected RDTScan's ability to adjust the exposure and focus while the CHWs captured images using the Infinix Hot 7 phone.

Table 3.9: Confusion matrices showing the performance of RDTScan against (left) expert interpretation of RDT images and (right) RT-qPCR for the Kenya Malaria Study. As a point of comparison, `DIRECT_READ` achieved 84.6% accuracy when compared to `PCR_RESULT`.

	`DIRECT_READ`		
`ALGO_READ`	Pos (+)	Neg (–)	
Pos (+)	212	1	PPV = 99.5%
Neg (–)	10	77	NPV = 88.5%
	SNS = 95.5%	SPC = 98.7%	ACC = 96.3%

	`PCR_RESULT`		
`ALGO_READ`	Pos (+)	Neg (–)	
Pos (+)	156	21	PPV = 73.7%
Neg (–)	12	39	NPV = 88.2%
	SNS = 92.9%	SPC = 65.0%	ACC = 85.5%

SNS = sensitivity, SPC = specificity, ACC = accuracy, PPV = positive predictive value, NPV = negative predictive value

Results: Accuracy

A total of 300 images passed all of RDTScan's quality checks and were subsequently interpreted. Due to mismatches between RDTs that were successfully captured using RDTScan and samples that could be processed by RT-qPCR, 228 images were used for comparison between `ALGO_READ` and `PCR_RESULT`. Table 3.9 shows the confusion matrices for both comparisons. When compared to `DIRECT_READ`, RDTScan achieved 96.3% accuracy (sensitivity = 95.5%,

specificity = 98.7%). When compared against `PCR_RESULT`, RDTScan achieved 85.5% accuracy (sensitivity = 92.9%, specificity = 65.0%). As a point of comparison, `DIRECT_READ` compared against `PCR_RESULT` had 84.6% accuracy (sensitivity = 93.5%, specificity = 60.0%). RDTScan demonstrated comparable accuracy to the CHWs ($Bin(4, 0.5) = 1.000, n.s.$). There were 3 false positive cases when the CHWs saw lines that did not exist according to `PCR_RESULT`. RDTScan was able to correctly interpret these cases as true negatives, showing that RDTScan may be particularly helpful in situations when clinical personnel are asked to administer many RDTs under time pressure. Similar to the QuickVue RDT, the CareStart RDT had significantly lower specificity relative to RT-qPCR. Unlike in the Australia Influenza Study, however, the CHWs had far more experience administering RDTs than the first-time end-users. Instead, the discrepancy may be due to the fact that RT-qPCR is a molecular-based technique that undergoes stages of processing to detect the presence of the pathogen genome, while the RDT mainly works to detect immune reactive elements like antigens. Most of the false positives between `ALGO_READ` and `PCR_RESULT` had clearly visible test lines, so the CareStart RDT itself may have had low specificity. Similar findings have been reported in previous work [133, 184], which found that CareStart RDTs have a specificity around 80% in the field.

3.8.3 Comparison across the Two Case Studies

The two case studies were conducted in vastly different situations, cultures, and environments. In this section, we highlight some of the insights that can be gleaned from comparing and contrasting the studies, and we also talk about some of the protocol decisions that may guide future deployments of RDTScan.

RDT Form Factor

The robustness of our RDT detection approach depends on the appearance of consistent visual keypoints around the result window. Template images with an even aspect ratio (i.e., as wide as they are tall) are less susceptible to localization errors since they are more likely to have keypoints surrounding the result window. As such, the initial version of RDTScan was

able to support the CareStart RDT because of the unique lettering and markings throughout its design. The thin and inconsistent design of the QuickVue RDT, on the other hand, posed greater challenges. We overcame these issues by introducing the notion of fiducial detection in RDTScan (Section 3.3.3), but the lower capture success rate we observed in the Australia Flu Study may still be partly attributable to the QuickVue RDT's design. The contrasting form factors also highlighted the utility of different functions supported by RDTScan. Blood can seep into the result window of the CareStart RDT if it is not administered properly, and glare can appear on the result window because of the glossy film that protects its result window. In the Kenya Malaria Study, RDTScan correctly rejected 15 images because of blood and 6 images because of glare, and none of the remaining images had enough blood or glare to obscure results. Meanwhile, the use of color information to validate control and test lines played a critical role in avoiding false positives since RDT detection was less robust for the QuickVue RDTs.

Usability-Accuracy Tradeoff

The guiding principle of RDTScan is that the higher the quality of the RDT image being analyzed, the more accurate RDTScan can be with automatic interpretation results. In other words, loose quality checks lead to better usability and lower accuracy, while tight quality checks lead to worse usability and higher accuracy. We arrived at the thresholds described in this paper through our prior work [149] and conversations with our collaborators about their slight preference for high accuracy over usability. Both studies used the same settings when possible to avoid further confounds in evaluating RDTScan. In a real-world deployment, however, the prioritization of usability versus accuracy depends on the context in which RDTScan is being used. For example, developers may want to prioritize accuracy when RDTScan is being used as the sole mechanism for diagnosis. On the other hand, developers may want to prioritize usability if RDTScan is being used as a screening tool to identify individuals in a diverse population who need a more expensive clinical test. Space precluded us from presenting a complete grid search of the numerous thresholds and quality checks

within RDTScan, and their impact on the user experience, so we leave that to future work. However, researchers and developers may want to explore adaptive settings for RDTScan's quality checks so that they become less strict over time. The 30-second timeout implemented in both study apps is an extreme case of an adaptive setting, removing all quality checks at the end of the timer for the sake of capturing a picture.

Experience

In our prior work [149], we found that CHWs became more adept at using RDTScan over time. Participants in the Australia Flu Study only used RDTScan once, while the CHWs in the Kenya Malaria Study used RDTScan multiple times; the practice that CHWs' gained from repeatedly using RDTScan may explain the superior usability results in the latter study. Beyond people's familiarity with RDTScan, familiarity with smartphones in general also could have driven the differences we saw in usability metrics. Studies have shown that the elderly face many obstacles when learning new features on their smartphones [122, 15], and our results from the Australia Flu Study reveal that younger participants had better success using RDTScan than older participants.

3.8.4 Comparison to Previous Version of RDTScan

The deployment from our prior work (Mali Malaria Study) was similar to the Kenya Malaria Study. Although the studies involved different RDT brands, both were conducted in sub-Saharan Africa with CHWs as study participants. Furthermore, the CHWs in both studies used low-end smartphones to capture cassette-based RDTs. In light of these similarities, RDTScan's improved RDT detection algorithm led to significantly better usability statistics in the more recent study. The capture success rate rose from 67% to 90.4%, and the average capture time went down from 14.3 seconds to 8.38 seconds. The interpretation accuracy metrics were much more similar across the two studies. The original version of RDTScan had 98.7% accuracy relative to expert readings in the Mali Malaria Study. Meanwhile, the updated version of RDTScan had 96.3% accuracy. The slight reduction in accuracy

could be attributed to the fact that the baseline for the Mali Malaria Study came from the consensus of two lab technicians, while the baseline for the Kenya Malaria Study came from one of six CHWs. For a fairer comparison, we re-ran the original RDTScan interpretation algorithm on the images from the Kenya Malaria Study. We found that the original algorithm would have accepted 21 images that were rejected by the updated version of RDTScan—6 because of glare and 15 because of blood. By letting those images be interpreted, the original version of RDTScan would have been 6.3% less accurate. We also compared the two versions of RDTScan's interpretation algorithm on the images from the Australia Flu Study. The updated version of RDTScan uses color information to validate interpretation results, which led to a 7.4% accuracy improvement from the previous version. This performance boost came from improving RDTScan's ability to localize the QuickVue RDT in 18 cases.

3.9 Discussion

RDTs have already made significant impact on healthcare diagnostics because of their low cost and relevant application to various medical conditions and diseases. As RDTs become approved for use by broader populations, their roles will likely expand even further. We believe that RDTScan will significantly contribute to the expanding role of RDTs, but there is more research to be done to this end. Below, we describe the limitations of our approach and potential areas of future exploration.

3.9.1 Expanding RDTScan to New RDT Designs

We chose to use feature-based template matching for RDT detection because it requires a single template image and some simple metadata for each new RDT design. Although this requirement has significantly less overhead than the large datasets needed to support a machine learning model, our approach is not optimal in all cases. In order to accurately localize an RDT, RDTScan relies on the assumption that unique visual keypoints occur naturally in RDT designs (e.g., logos, labels, and arrows). To assess the validity of this assumption, we conducted an informal study to determine if RDTScan could reliably detect

RDTs from other manufacturers. The images of the RDTs we examined for this informal study can be found in Appendix A.1. We applied RDTScan's detection algorithm to RDTs from two lists curated by major health agencies—one by the World Health Organization for malaria RDTs [141] and one by the Centers for Disease Control for influenza RDTs [49]. Out of the 12 RDTs in those lists, 10 were cassettes and 2 were dipsticks. RDTScan was able to detect 8 of the 10 cassettes, all without the need for fiducial checking; both failure cases were blank white cassettes without any lettering or logos. Thresholding an image based on the cassette's color would be the simplest way of handling such RDT designs, but that would also require end-users to capture their RDT on a distinct background. RDTScan was able to detect both of the dipsticks provided that fiducial checking was enabled. We also examined 3 more pregnancy test dipsticks and found that RDTScan was able to support all of them with fiducial checking. As an alternative to fiducial checking, we considered leveraging multiple templates to cover the range of possible markings; however, that approach would have incurred significantly more computation per frame and thus would have affected the user experience.

A data-driven recognition model may eventually become the ideal approach to RDT detection if RDT designs become more diverse and deep learning on smartphones becomes more computationally efficient. Training such a model will require a large training dataset of RDT images, so we hope that RDTScan can accelerate the creation of such datasets in the future.

3.9.2 Supporting Developers

RDTScan[5] is publicly available as an open-source codebase, and we have witnessed other global health organizations utilize and extend RDTScan to suit their needs since its release. Most notably, we have seen organizations leverage RDTScan for COVID-19 RDTs that are being developed worldwide. Because such RDTs are being actively developed and evaluated, one organization is using RDTScan to support the automatic interpretation of 10 different brands of COVID-19 RDTs. RDTScan's ability to rapidly accommodate new RDT designs

without the need for extensive data collection has been critical to this end.

Beyond the fact that incorporating RDTScan into a smartphone app requires mobile programming experience, using RDTScan also requires familiarity with tools that would enable a person to extract metadata for their target RDT. These tools include a mobile scanner app (e.g., Microsoft Lens) and a photo editing program (e.g., Microsoft Paint), both of which can be downloaded for free. Using these tools does not require programming experience since they have GUIs, and we include detailed instructions on this process for future developers on the RDTScan website. We were able to coach project managers in our case studies to generate this metadata for themselves within a phone call. Nevertheless, we recognize that a smoother process for expanding RDTScan to new RDT designs would make RDTScan even easier for developers to use. We envision a software tool that could help developers more easily extract metadata from their template RDT image, whether through a domain-specific user interface or automatic image analysis. The automatic image analysis approach would share many of the same characteristics as RDTScan's interpretation algorithm, but the tool would be able to take advantage of assumptions about the template to make the analysis easier, such as the fact that all of the lines will be activated.

3.9.3 Standards for RDT Design

Because the concept of RDTs predates that of computer vision, RDTs are not typically designed with image processing in mind. After demonstrating that an unmodified smartphone can be used to improve the effective accuracy of their product in the field, we have begun to collaborate with RDT manufacturers to create a standard RDT design that is optimized for image processing. This prototypical RDT design includes a flat, wide cassette that minimizes shadows and provides sufficient tolerance for RDT localization. To eliminate the need of an RDT template image, our RDT prototype includes consistent fiducials at each corner of the cassette to simplify the detection of the RDT's edges, similar to what is done with the detection of QR codes [40]. In fact, we envision each RDT can be printed with a QR code that includes the same metadata needed to initialize RDTScan, such as the location and

meaning of the different lines along the result window. A thick colored border around the edge of the RDT is included for further correcting any errors in result window localization. Finally, our prototype also includes color references so that the effects of ambient lighting can be removed through color-calibration. RDTScan would still have useful functionality for an RDT with all of these features, such as its result interpretation algorithm and many of its quality assurance checks.

3.9.4 RDT Procedure Controls

RDTScan only checks the end result of the RDT, but that result is the culmination of an entire procedure that often goes unchecked. RDTs are prone to many procedural errors: the use of expired tests, sampling and handling errors of the biological sample, and improper fluid usage, to name a few [57, 69]. Some health organizations have used periodic supervision [128] and training programs [69] to educate their workforce on proper RDT administration, but such measures are only possible for repeated RDT users like CHWs. Instead, we believe that image processing and computer vision can be used as a proxy supervisor of the RDT procedure. For example,

- Optical character recognition can be used to confirm that the expiration date has not passed,

- Color blob detection can be used to check that enough blood has been drawn into a capillary tube, and

- Optical flow can be used to ensure that the user has swirled their nasal swab in a tube for enough time.

Implementing procedural controls will incur additional power consumption on the smartphone, so adding these checks will require a balance of utility and simplicity.

3.9.5 RDTScan's Long-Term Efficacy

We demonstrated that RDTScan provides accurate image capture and interpretation performance when used by people with varying levels of technological and clinical expertise. However, many solutions that rely on mobile apps to address human-centered issues (e.g., health, finance, education) have failed due to the complexity of workflows and relationships, particularly in low- and middle-income settings [186]. Longitudinal deployments in which people are shown the automatic results generated by RDTScan would be needed to assess how RDTScan could be integrated into people's workflows and the level of trust that people are willing to place in an automated system like RDTScan. As a direction of future work, we plan on deploying RDTScan within community health clinics and using both naturalistic observations and interviews to capture people's experience with the system.

3.9.6 Limitations of Smartphones

Smartphones are preferred over other devices like laptops and tablets because of their portability and ability to serve multiple purposes (e.g., phone calls, SMS messaging). Even so, smartphones are not without their limitations. When one lab technician re-visited images of positive RDTs that he had initially determined were negative, he said, "It is clear that the images are positive on the computer screen, but on the smartphone screen it is sometimes hard to be sure because the line is so faint". The preference of viewing images on a laptop instead of smartphones was echoed by other individuals, primarily because the laptop's larger screen size (15 inches vs. 5.5 inches) allowed them to make the images much larger and thus easier to read. Color response, brightness, and resolution can also vary between different screens and impact how people interpret RDT images. If global health organizations are interested in deploying tools like our RDT capture app for on-site image interpretation, they may want to consider the affordances of the devices that display those images. From our experiences, we believe that tablets may provide an optimal balance between cost, portability, and screen size to promote accurate image interpretation.

3.9.7 Expediting the Review Process

Although RDTScan is meant to alleviate the need for CHWs to transport RDTs back to their supervisors in the clinics, supervisors still wanted to review the decisions that the CHWs made. This review process can be burdensome across an entire organization when each supervisor has around 10–20 CHWs for whom they are responsible. Clever grouping and sorting of the RDT photographs could expedite this process in the future. When RDTs are grouped by a particular characteristic, supervisors can assess a representative subset of images within each group. For example, if a supervisor examines a subset of images captured by a particular CHW and concurs with all of their decisions, the supervisor may accept all of the other decisions made by that CHW and turn their attention to one with more disagreements. Our image processing pipeline can also contribute to an intelligent ordering within these groups, sorting images according to line intensity and allowing supervisors to prioritize faint lines over intense or non-existent ones.

3.10 Conclusion

RDTs are becoming an increasingly popular option for point-of-care medical testing. To complement this trend in healthcare, we have created a mobile system that enables RDT detection and interpretation with a smartphone. We demonstrated that template-based feature matching is a feasible approach to RDT interpretation, and we argue that our approach is easier to scale to new RDT designs than a data-driven one. We also showed that RDTScan provides interpretation accuracy levels that are comparable to experts reading the RDTs themselves, opening the door to new workflows for at-home and in-clinic scenarios, particularly in low-resource settings. RDTScan is an open-source library, so we encourage public health organizations to utilize our system for their RDT-related needs. We anticipate that RDTScan can reveal novel insights and support new systems in public health, and we hope that this work inspires engineers and researchers to create software libraries that support the interpretation of other analog devices.

Chapter 4

CAPAPP: SMARTPHONE-BASED CAPILLARY REFILL TIME MEASUREMENT

4.1 Introduction

Capillary refill time (CRT) is a physiological measurement that describes the time it takes for the color to return to tissue that has been blanched by external pressure. CRT assessment is often used to identify cases of circulatory shock, with one of the most common applications being the diagnosis of pediatric dehydration (a form of hypovolemic shock). The most common methodology for measuring CRT entails pressing a person's fingertip until it becomes pale and then visually assessing when the color returns TODO: cite. This procedure is easy to learn and does not require any equipment beyond an optional stopwatch, making it useful for point-of-care assessments and emergency situations. In fact, CRT assessment is one of the many techniques taught during the Pediatric Advanced Life Support course administered by the Red Cross and many other emergency services for treating pediatric patients.

Although CRT assessment requires minimal equipment and training, it is still a highly subjective test that is prone to many errors. Reporting CRT without a stopwatch is imprecise, and reporting CRT with one depends on the examiner's reaction time. Knowing when complete blanching and capillary refill occurs is also highly subjective as it depends on the examiner's color acuity. Lastly, there are no enforceable standards for the optimal amount and duration of force beyond the fact that the capillary bed should be blanched. Given these factors, existing clinical literature typically relies on the pragmatic rule that CRTs longer than 2 seconds should warrant concern. We posit that a more accurate, precise, and repeatable method of measuring CRT that maintains the simplicity and scalability of the test will simultaneously enable new point-of-care assessments and yield new clinical insights.

In recent years, researchers have demonstrated various applications that leverage a smartphone to perform photoplethysmography (PPG) — a technique that optically measures the blood volume changes in a person's finger — by recording the fingertip as it covers both the camera and an illumination source (flash or screen). We use a similar approach to create a smartphone app for CRT assessment called CapApp. To use CapApp, the user presses their finger against the smartphone's front-facing camera. The app analyzes the PPG signal and motion data as the phone vibrates to determine if sufficient pressure has been applied to the fingertip. Once the user's finger has been blanched, the user relaxes their fingertip and the return of blood flow to the fingertip is again measured using the PPG signal.

We evaluate CapApp in a study with 20 healthy adults and 10 adults with Raynaud's phenomenon [75] — a condition in which people experience numbness in their extremities in response to cold temperatures or stress. We manipulated these individuals' CRT using an ice bath to show that CapApp is able to track CRT changes due to temperature changes. We compared the accuracy of our approach against two alternative methods of measuring CRT: visual inspection with a stopwatch (the existing practice) and algorithm-assisted inspection (the gold standard). CRTs estimated by CapApp show a high correlation with temperature ($\rho = 0.74$), which is comparable to the two other methods. The second analysis demonstrates the potential clinical value of CapApp by showing its ability to detect cases of Raynaud's phenomenon. Although Raynaud's phenomenon is not as debilitating as some of the aforementioned ailments, this evaluation shows that CapApp is able to measure clinically significant differences across a population for the sake of diagnosis. From mixed-effects model, we found that CapApp was able to extract meaningful differences between two populations from estimated CRT.

4.2 Related Work

For our summary of related work, we first describe the physiological mechanisms underlying capillary refill and the pathology of elongated CRT. We then enumerate methodologies and applications of smartphone-based photoplethysmography.

4.2.1 Clinical Literature on Capillary Refill Time

Capillary refill time (CRT) is a rapid cardiopulmonary assessment that entails visually inspecting the time it takes for blood to return to the capillaries after they have been emptied by pressure. The primary purpose of CRT assessment is to provide information on peripheral blood perfusion, although it can be impacted by other characteristics of blood flow like blood pressure, the constituents of the blood, and the balance between vasoconstriction and vasodilation (narrowing and widening of blood vessels, respectively). The convenience of CRT assessment makes it particularly suited for low-resource settings, where more expensive medical equipment may not be readily available, or pediatric clinics, where patients may not be compliant or compatible with other assessment tools. Clinical literature suggests that CRTs longer than 2 seconds should warrant further diagnosis [14], but this threshold is debated to this day due to the coarse granularity of the assessment [153].

Along with vital signs like heart rate, respiratory rate, and blood oxygen level, CRT is used as a quick assessment tool for screening various health conditions. Vasoconstriction is triggered in response to circulatory shock or cold temperatures to conserve blood flow and prevent heat loss in vital organs; thus, prolonged CRT can be an early sign of shock or hypothermia. This phenomenon has been particularly studied in pediatric cohorts, where prolonged CRT has been shown to be correlated with septic shock [104, 153]. CRT assessment has also been used to screen children for dehydration, particularly in cases when they are admitted to hospitals with diarrhea. Similar results have been shown for adults, although the findings are less abundant [104]. One condition that has been well characterized in adults using CRT assessment is Raynaud's syndrome [75] — a condition that causes patients to experience numbness or pain in their extremeties due to reduced blood flow in cold environments. Raynaud's syndrome is not life-threatening, but can be the first presenting symptom for conditions ranging from connective tissue disorders to carpal tunnel syndrome or multiple sclerosis [34].

4.2.2 Capillary Refill Time Assessment

The subjective nature of manual CRT assessment makes it susceptible to many errors [153, 7, 94, 158]. Klupp et al [94] demonstrated an interclass coefficient of 0.72 in CRT assessments conducted by five physicians. But, they also found that CRT measurements on the same adult could vary by almost as much as 2 seconds when taken by different clinicians. The capillaries are normally emptied by applying pressure using one's own finger or a clear stiff surface (e.g., plastic, glass), but clinicians can vary in the duration and amount of pressure they exert. Many researchers [181, 171, 63] found that at least three seconds of pressure should be exerted in order to achieve consistent and accurate CRT measurements. Steiner et al. suggested that the pressure should be applied until the capillary bed gets blanched. But, Saavedra et al. [165] also found that lighter pressure leads to a shorter CRT but with greater consistency. CRT assessment can also produce different results depending on where pressure is applied. The most common test sites include the fingertips and fingernails [142], but CRT can also be measured on the chest, forehead, and heels [180, 63]. Prior work has shown that the CRT is shorter when measured at the fingertips compared the heels, and CRT assessments at the forehead and chest tend to be more consistent those at the heel or palm [180, 63]. CRT can vary by people performing the measurement since people have different pressure duration/amount, color acuity, and threshold for capillary refilling.

The lack of a universal standard on the protocol for manual CRT assessment has hampered its clinical utility [104]. As such, there have been many attempts to make this procedure more objective and repeatable using additional hardware. Shavit et al. [173] proposed the use of digital videography for automating CRT assessment. Although they do not describe their algorithm in significant detail, their method resulted in stronger correlations with dehydration status than manual assessment by physicians. PPG sensors [18, 175] can also be used for CRT measurement since it quantifies blood volume changes in the microvascular bed of tissue; the PPG signal flattens when the capillaries are emptied, and then the cardiac signal returns once they are refilled. Despite removing the subjective nature of visual assessment, the

aforementioned works have still relied on physicians to apply and remove pressure from the fingertips. Researchers have therefore built custom hardware to automate the entire process. For example, both Shinozaki et al. [175] and Blaxter et al. [18] proposed systems that used an air pump to apply pressure on finger while measuring CRT with a pulse oximeter. Kerr et al. [91] proposed a robotic arm to for the same purpose. These methods are highly repeatable but come at the cost of additional hardware and poor portability, making them less suitable for point-of-care and rural settings.

In contrast to prior work, CapApp is a purely smartphone-based solution for CRT assessment. Using a technique initially proposed by [190, 60], CapApp is able to give real-time guidance as to whether or not sufficient pressure has been applied to the fingertip. CapApp also builds upon prior work on smartphone-based PPG (described below) to detect blanching and characterize the refill process. Together, these sensing components lead to a novel and accessible CRT assessment tool that we evaluate in an elaborate protocol with both inter- and intra-subject variance.

4.2.3 Smartphone-Based Photoplethysmography

As an optical measurement technique, PPG is often conducted using a photodiode and a light emitter tuned to a particular wavelength that accentuates blood volume changes (typically 530 nm [90]). Many modern-day smartwatches and fitness wristbands include a PPG sensor to enable continuous vital sign tracking, yet people without these devices are excluded from these applications. Additionally, PPG signals at different body sites provide different implications; wristbnads can only extract radial PPG, whereas smartphone can be placed on any parts of body. To bring PPG to wider audiences, researchers have demonstrated the ability to repurpose components of a smartphone as a collective PPG sensor — the camera sensor serving as the photodiode and the flashlight or screen serving as the light emitter [195, 110]. Rather than measuring light emission through the finger, smartphone-based PPG measures the amount of light that is reflected back towards the camera when a person covers both the camera and the light source with the same body part. This technique extends PPG to a

wider class of devices but limits it momentary assessments.

Researchers and engineers have used smartphone-based PPG to measure heart rate and heart rate variability by computing the intervals between the peaks and variability of inter-peak intervals, respectively [110]. Furthermore, researchers [178] have demonstrated that smartphone-based PPG can be used to estimate blood pressure by either analyzing the shape of the pulse signal [121] or measuring the time it takes for a pulse to travel between two locations on the body [195]. Since smartphone cameras are designed to capture light wavelengths across the visible spectrum, smartphones can also be used to perform multi-wavelength PPG. Yan et al. [204] and Jeong et al. [83] argue that multiple wavelengths approach to PPG sensing reduces bias against varied skin tones, increasing heart rate sensing by up to 15%. Wang et al. [194] compute the ratio the PPG signal across color channels to estimate hemoglobin levels, disambiguating the reflected signal from hemoglobin and plasma using machine learning; Bui et al. [**?**] use a similar approach to measuring oxygen saturation.

Our work expands upon this body of literature by applying smartphone-based PPG for CRT measurement. Unlike prior methods, CapApp requires a signal that is not strictly pulsatile. As the user applies pressure, CapApp detects blanching by waiting for a diminished cardiac waveform. Once the user releases the pressure, however, blood rapidly returns to the fingertip and the waveform is dominated by an exponential decay in the reflected signal. Although multi-wavelength PPG has shown promise in various aspects of blood volume sensing, we do not leverage it to CapApp and leave it to future work.

4.3 Design of CapApp

CapApp operationalizes CRT assessment by leveraging the smartphone's camera and flash for PPG signal acquisition and the vibration motor and motion sensor for force sensing. The procedure for using CapApp can be broken into four steps:

1. **Relaxing**: The user gently covers the camera with their finger without exerting any force. CapApp waits until a stable heartbeat is seen through the camera before

advancing.

2. **Pressing**: The smartphone vibrates and the user presses their finger against their phone. CapApp measures how hard the user is pressing the phone by using the accelerometer.

3. **Blanching**: Once the user has applied enough sustained force with their finger, CapApp checks for a diminished PPG signal as a sign of blanching.

4. **Releasing**: The user releases their finger while keeping it on the camera. CapApp measures the time it takes for the blood to refill within the finger.

Extracting the PPG signal from a person's fingertip requires a colocated illumination source and light sensor. CapApp can work with multiple configurations, but we ask users to cover the smartphone's front-facing camera for ergonomic reasons. When people are using CapApp on their own, pressing against the front of the screen is more natural than against the back; when people are using CapApp on someone else, it is easier for them to simultaneously look at the instructions and the other person's finger if they are on the same side of the phone. Furthermore, using the screen as an illumination source is safer than using a flash LED since the latter can become uncomfortably warm when kept on for more than a few seconds. In the following subsections, we explain each component of CapApp in greater detail.

4.3.1 PPG Verification

CapApp leverages smartphone camera-based PPG extraction to continuously measure blood volume in the fingertip [195]. To make this possible, CapApp first instructs the user to place their fingertip over the camera and flash without exerting any force on the smartphone. CapApp continuously measures the amount of light reflected back into the camera by recording a video and computing the average lightness of the pixels within each video frame according to the HLS (hue-lightness-saturation) color space. As blood enters the finger with each pulse, more light is absorbed and less light is reflected back to the camera; conversely, less light is

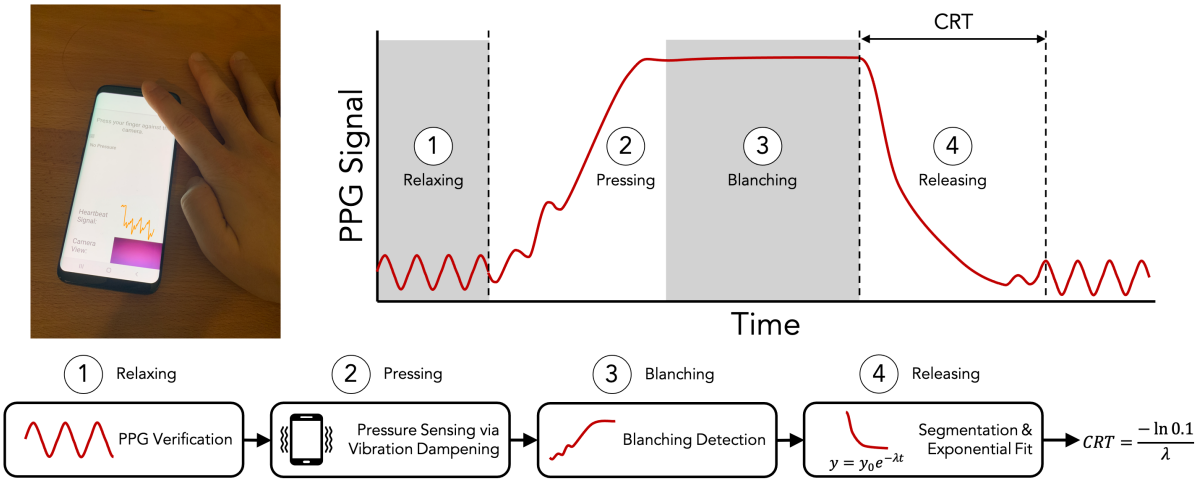


Figure 4.1: (top-left) An image of a person using CapApp to assess their own CRT. (top-right) An illustration of how the PPG signal ideally changes as a person completes a CRT assessment. Lower PPG values imply higher blood volume since the blood absorbs more light. (bottom) The four major components of the CapApp software.

absorbed and more is reflected back when the blood is circulated back to the rest of the body. This ebb-and-flow pattern corresponds to the cardiac cycle.

The quality of this signal is impacted by how the user positions their finger on the camera. If the finger only partially covers the camera, ambient light enters the camera reduces the signal-to-noise ratio of the PPG signal. CapApp checks for this procedural error by checking regular peaks over a 5-second window and measuring the amplitude of the PPG signal. If the user moves their finger during the assessment, the periodicity of the PPG signal is disrupted. CapApp checks for this procedural error by checking for number of peak intervals within the range of typical heart rates (60–100 beats per minute). Once a consistent PPG signal has been seen for 5 seconds, CapApp advances to the next screen.

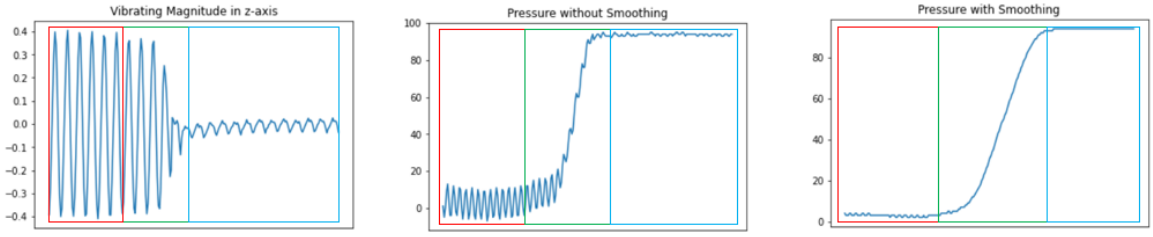


Figure 4.2: An illustration of how the vibration pattern recorded by the accelerometer varies as the user presses the camera with their fingertip. In all graphs, the red box indicates when the fingertip is at rest, while the blue box indicates when pressure is being applied by the fingertip. (left) A graph of the raw acceleration along the z-axis of the phone. (center) The relative pressure magnitude without smoothing. (right) The relative pressure magnitude after smoothing using a moving average.

4.3.2 Pressure Sensing

Measuring CRT requires that the user applies enough force to their fingertip to constrict blood flow. Pressing one's own finger against the front-facing camera of their smartphone can be an unfamiliar sensation to some. Although the amount of force required to cause blanching is not nearly enough to physically damage a phone, some may still be hesitant to exert sufficient force over an adequate period of time. CapApp provides users with real-time feedback to ensure that the amount of force the user exerts meets some minimum threshold.

Prior work has suggested optical methods for estimating the force exerted during smartphone-based PPG. For example, Bui et al. [20] originally proposed that pressure could be estimated by looking for relative changes in average PPG signal intensity, noting that more light is reflected back with increasing force. Later, they proposed an alternative technique that relied on the fact that fingerprint ridges become more dispersed as a person presses with their finger [18]. These methods are sensitive to both the user's skin tone and the blood volume in their fingertip. Instead, we leverage on vibration damping for force estimation [190, 60].

Whenever the vibration motor on a smartphone is turned on, the entire phone shakes and the accelerometer is able to detect those vibrations. As the user presses their finger against their phone, the magnitude of those vibrations decreases since the finger provides structural support and absorbs some of that kinetic energy. Figure 4.2 illustrates this principle from the perspective of the accelerometer, showing how greater forces result in weaker vibrations.

The efficacy of this method for force sensing depends on multiple hardware-specific factors. The most important factor is the relative position of all three smartphone components that are involved in this app: the vibration motor, the accelerometer, and the front-facing camera. The ability of the finger to damp vibrations from the motor decreases as the distance between the camera and the motor increases. Similarly, the ability of the accelerometer to sense the damping decreases as the distance between the accelerometer and the motor increases. The type and orientation of the vibration motor are also important, as damping is most evident when axis of vibration is colinear to the direction of the force being applied. Since most of these hardware specifications are difficult to ascertain from software or online resources, CapApp uses a brief calibration procedure to measure the phone's baseline vibration characteristics. After the user places the phone on a flat surface, CapApp turns on the phone's vibration motor without any other forces acting upon the phone. The resulting motion is measured via the accelerometer along the z-axis since that is the direction along which the user would be exerting most of the force with their finger. The peak frequency in the FFT of that signal is assumed to be the motion caused by the vibration motor, and the magnitude at that frequency is treated as a baseline. When the user exerts force on the phone with their finger, the magnitude at the vibration frequency is compared to the baseline value. CapApp determines that the user is exerting sufficient force when the magnitude at the vibration frequency is 10% of the baseline. To make the force measurement more reliable and stable, CapApp smooths the magnitude values using a moving average with a 0.5-second sliding window.

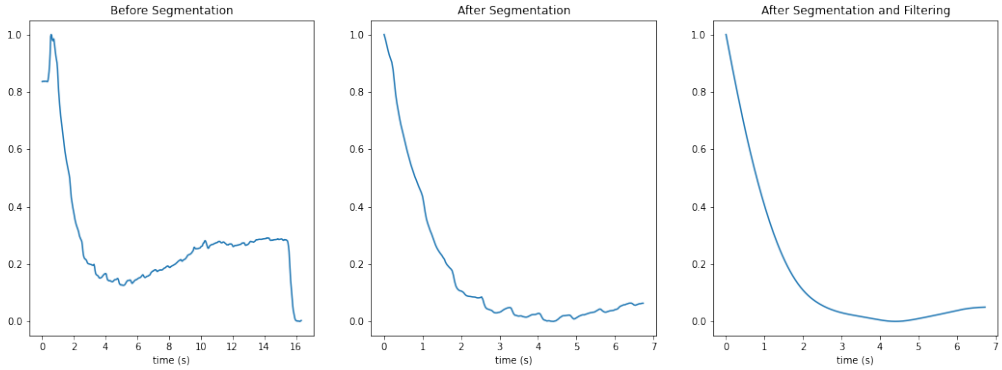


Figure 4.3: Left image is normalized PPG signal before segmentation, Middle is normalized PPG signal after segmentation, Right is normalized PPG signal after applied low-pass filter

4.3.3 Blanching Detection

Perhaps the most crucial step in a CRT assessment is that the measurement site must be completely blanched before the pressure is released. During a typical capillary refill assessment, blanching is either visually determined or assumed, depending on whether the force is exerted in a way that the examiner can see the region of tissue on which the force is being applied (i.e., by pressing with a clear piece of plastic versus their own finger). When visually determining blanching, people are able to reason about whether the finger has reached its final state by comparing the color of the tissue being pressed relative to that of the surrounding tissue. Since the camera can only see the tissue that is directly over the camera, CapApp instead relies on the fact that the core cause of blanching is the lack of blood flow to the tissue. In the PPG signal, this appears as a damping of the pulsatile cardiac waveform. Algorithmically, CapApp determines that blanching has occurred when the average amplitude of lightness values of the PPG signal within a one-second window is less than 5, where the maximum PPG signal value is 100.

4.3.4 CRT Calculation

Segmentation

Once the finger has been sufficiently blanched, CapApp instructs the user to relax their finger. Doing so causes blood to quickly rush back into the fingertip in a nonlinear fashion — quickly at first, and then asymptotically approaching the pulsatile steady state. The corresponding PPG signal appears as an exponentially decaying curve that slowly turns into a cardiac waveform. CapApp requires that the user keeps their finger on the camera for 10 seconds to observe the refill process. Although CRTs can be longer than 10 seconds, CapApp can extrapolate such measurements based on the shape of the curve that is recorded during that period. Furthermore, these occurrences are rare and the clinical utility of distinguishing between CRTs beyond this threshold is unknown.

To calculate the user's CRT from the PPG signal, CapApp first discards all of the data before the user was instructed to relax their fingertip. The data is then normalized between 0 and 1 since the most important characteristic of the data is the shape of the curve. Theoretically, the resulting curve should start with a large negative gradient as blood rushes into the fingertip. However, there is usually a delay since the user needs time to react to the app's instructions. Some some users also apply a small amount of additional force before they relax, resulting in a brief spike in the PPG signal. An example of both of these behaviors is shown in Fig. 4.3. To precisely clip the signal at the start of the refill process, CapApp applies a 0.15-second sliding window after the timestamp with the maximum PPG value to locate the moment at which the slope of the normalized signal reaches -3; the start of the window that meets this criteria is used as the beginning of the signal.

One way to determine the end of the signal would be to identify when the PPG signal reaches its original average value before the CRT assessment; however, this rarely happens due to the sensitivity of smartphone-based PPG and the fact that the finger rarely returns to the exact same position. CapApp instead relies on the return of the pulsatile cardiac waveform. However, Fig. 4.3 also shows that the end of the PPG signal can deviate from

the expected shape mentioned earlier. After the initial rush of blood to the fingertip, many people make subtle adjustments to their fingertip positioning depending on whether they relaxed their fingertip too much or too little. These changes produce gradual increases or decreases, respectively, in the average PPG signal value. More importantly, there may be cases when the user prematurely removes their finger from the camera. The shape of the exponential curve could be inferred by only using the first second of data, but leveraging as much data as possible is advantageous for curve fitting. To precisely clip the signal at the end of the refill process, CapApp first identifies significant peaks or valleys in the PPG signal by moving backwards in time clipping the signal at the start of any window that the height of peak or valley is larger 0.04 or 0.02, respectively. CapApp then applies a 0.5-second sliding window after the timestamp with the minimum PPG value to locate peaks in the PPG signal (ideally due to the cardiac waveform). The PPG signal is clipped when the value of a detected peak deviates by 10% from the value of the first peak.

Curve Fitting

To leverage the latter part of the PPG signal when the cardiac waveform returns, CapApp applies a Butterworth low-pass filter with a cutoff of 0.75 Hz; this cutoff corresponds to a lower bound of 45 beats per minute on the heart rate. The smoothed signal is fit to an exponential decay function of the form $y(t) = Ae^{-\lambda t} + y_0$ using a nonlinear least squares procedure. The fit is bounded such that $A \in [0.9, 1.1]$ and $y_0 \in [-0.01, 0.01]$ since the signal is normalized between 0 and 1. The formula $\frac{-\ln \frac{\delta}{A}}{\lambda}$ specifies how long it takes for the PPG signal to return back to the value δ along the normalized PPG range. Lower values of δ lead to longer CRT estimates and a greater dynamic range for CapApp, but its value does not impact the rank order of CapApp measurements. We set $\delta = 0.1$ so that we may compare our results to those of [175], who use a similar criterion for determining CRT, but we also examine various settings of δ on CapApp's performance in our evaluation.

4.3.5 Implementation

We implemented CapApp on a Samsung Galaxy S8 smartphone. The phone has an 8 MP camera that operates at 30 fps, an accelerometer that records motion data at 100 Hz, and a vibration motor that operates at roughly 50 Hz. Although CapApp could be implemented on a variety of other smartphones, we targeted the Galaxy S8 for two reasons: (1) it has a front-facing camera that is near its screen, and (2) it has a vibration motor that vibrates perpendicularly to the screen (z-axis), which maximizes the signal-to-noise ratio for the pressure sensing component of the app. We maximized the signal-to-noise ratio of the PPG signal by setting the phone's screen to its maximum brightness setting with white pixels around the camera. We further improved the PPG signal quality by adjusting the gains of the camera's red, green, and blue channels to 2.0, 3.0, and 18.0 respectively using Android's Camera2 API[1]. This has been done in prior work in involving smartphone-based PPG [195] to maximize the camera's sensitivity to blue wavelengths since they best differentiate blood from tissue.

4.4 Study Design

Our study was designed to characterize CapApp's ability to track inter- and intra-subject variations in CRT. Inter-subject variation was examined by recruiting people with and without Raynaud's syndrome. We induced intra-subject variation in CRT and tirggered symptoms in participants with Raynaud's syndrome by manipulating their peripheral body temperature. The ethics review policies of the relevant institutions, countries, and funding agencies were followed during the course of this study.

4.4.1 Participants

Participants were recruited via email advertisements sent to mailing lists at and the . The advertisement specifically solicited participation from people who self-identified as having

[1] `https://developer.android.com/training/camera2`

Table 4.1: Demographics for the two cohorts involved in the study protocol.

	Without Raynaud's (N=20)	**With Raynaud's (N=10)**
Gender	Female (10), Male (10)	Female (7), Male (2)
Age	26.15 ± 3.87	31.67 ± 7.00
Skin Tone	Pale (3), Fair (9), Medium (3), Olive (2), Brown (3), Black (0)	Pale (6), Fair (2), Medium (0), Olive (1), Brown (0), Black (0)

Raynaud's syndrome, but we also worked with to send targeted calls to people who were open to participating in research and had Raynaud's syndrome mentioned in their medical history.

The demographics of the cohorts with and without Raynaud's syndrome are shown in Table 4.1. Since skin tone is known to influence the robustness of optical PPG [83, 204], we recorded participants' skin tone according to the Fitzpatrick scale [47]. The demographic discrepancy between the two cohorts can primarily be attributed to the prevalance of Raynaud's syndrome, which is far more common in young Caucasian females [140, 155].

4.4.2 Procedure

Participants had their fingers blanched using two different methods. The first method entailed using CapApp, while the second involved a research coordinator manually exerting and releasing pressure applied to the participant's fingertip using a clear glass slide. The latter procedure is akin to the existing gold standard for CRT assessment; however, we recorded this procedure with a camera mounted on a tripod so that we could apply extract the CRT in multiple ways (see Section 4.4.3). The camera had its exposure and white-balancing levels locked to provide a consistent color when later analyzing this footage. Using both CapApp and the manual approach, pressure was applied to the finger until blanching, and the response was recorded until it either blood flow returned to the fingertip (CapApp) or the

apparent color of the fingertip returned back to normal (visual inspection). Participants were repeatedly asked if they were comfortable with the current pressure levels and the applied pressure level was decreased if the participant experienced any discomfort.

To observe CRT values beyond people's baseline and to trigger symptoms associated with Raynaud's syndrome, we manipulated their peripheral body temperature by asking participants to place their hand in an ice bath until reaching a target temperature. The two cohorts were exposed to different temperature ranges since those with Raynaud's sometimes require more extreme temperatures to trigger their symptoms. The cohort without Raynaud's syndrome had the temperature of their index finger changed from 30°C to 20°C in decreasing increments of 2.5°C. The cohort with Raynaud's syndrome, on the other hand, had the temperature of their index finger changed from 30°C to 20°C in decreasing increments of 2.5°C. The protocol was ended early if participants experienced discomfort due to colder temperatures. At each temperature, participants alternated between using CapApp and having their finger manually blanched five times, resulting in a maximum (5 trials per temperature) $\times$ (5 temperatures) = 25 trials per participant with each measurement method. Fingertip temperatures were adjusted between each trial and measured with a non-contact infrared thermometer. A tolerance of ± 1°C was allowed for declaring when the fingertip reached the target temperature. We recorded the fingertip's temperature before and after each trial to account for thermoregulation and heat transfer from the smartphone when participants used CapApp. We assigned the average of the two temperatures to each measurement, although our findings are generally consistent with alternative handling of the temperature data. Each study session took roughly 70 minutes, and all participants were compensated with $20 USD gift cards for their time.

4.4.3 Ground Truth Calculations

The trials that did not involve CapApp were designed to record the same visual information that would be used for the clinical gold standard. CapApp measures a different visual phenomenon relative to this standard — measuring the change in blood flow throughout the

fingertip rather than observing the external change in coloration. Nevertheless, we expected these measures to be significantly correlated since they both relate to blood circulation. To isolate the potential impact of subjectivity, we processed the video recordings using two techniques: manual inspection by experts and an automated version of the same assessment. We describe the process for generating these assessments below:

Manual Inspection

A team of three experts familiar with CRT assessment were recruited to annotate the video recordings. To simulate what would happen in a real-time clinical scenario, the annotators were instructed to measure the CRT by watching each video at full speed once without any pauses and then timing the phenomenon using a stopwatch. The videos were renamed and shuffled before being sent to the annotators to minimize information leakage between recordings. Most of the videos were equally divided amongst the annotators; however, 106 videos were randomly selected and distributed to all of them to assess inter-rater reliability. The annotators achieved an intraclass correlation coefficient of 0.83, which is considered good reliability according to literature [95]; more measures of variance can be found in our subsequent analyses. Sources of disagreement often stemmed from the variety in the color acuity and reaction times of the annotators. Blanching is also non-uniform across the fingertip, so annotators may have focused on different parts of the fingertip to make their final decision.

Algorithm-Assisted Assessment

As a less subjective version of the clinical gold standard, we created an image processing algorithm similar to the one by Shavit et al. [173] to automate the visual assessment process. Each video was clipped so that the first frame begins after the research coordinator removed the glass slide from the participant's fingertip. The videos were then manually cropped to a small region of interest within the fingertip, roughly 80% of the fingertip's total area, where blanching occurs when it is pressed. Since most of the color variation was in the red channel of the RGB color space, the average red channel value was computed for each frame and then

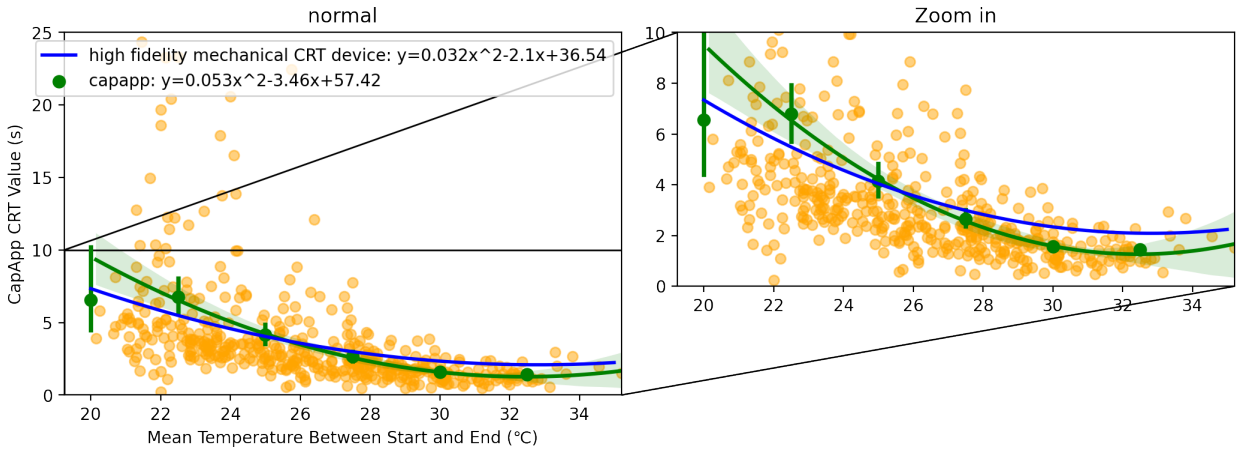


Figure 4.4: The relationship between CRT measurements recovered by CapApp versus temperature. As peripheral body temperature decreases, the average CRT and variance in CRT measurements increases.

plotted over time to produce a curve similar to the one recovered by CapApp. Other color dimensions yielded comparable results, but the red component had the highest signal-to-noise ratio across subjects and skin tones. The CRT was inferred from the curve using the same exponential fitting procedure as the one used in CapApp.

4.5 Results

In this section, we first examine the ability of CapApp to discriminate CRT differences due to changes in peripheral body temperature and Raynaud's syndrome. We then quantify the performance of CapApp relative to the manual and automated versions of the clinical standard. Finally, we combine these analyses in a mixed-effects model that also examines the impact of skin tone on CRT assessment.

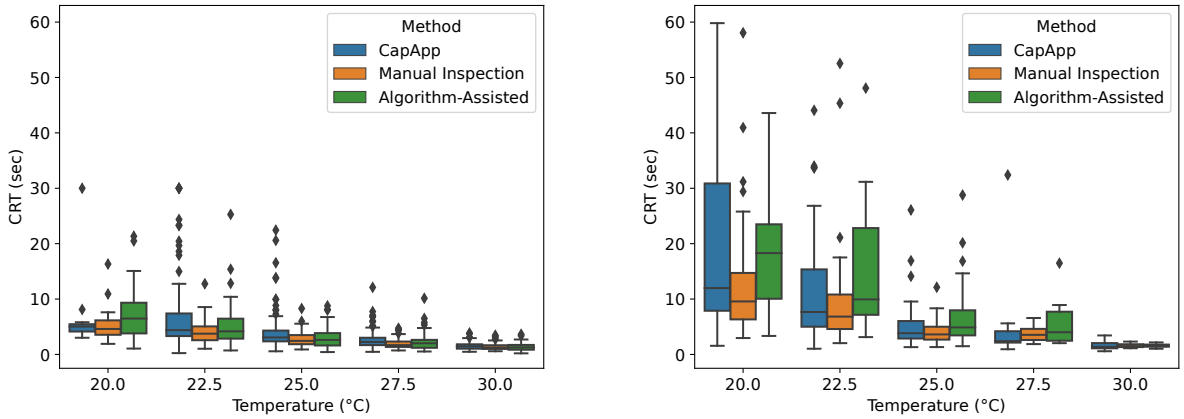
Table 4.2: Comparisons between different techniques, CapApp, manual inspection, and an algorithm-assisted assessment, on correlation of temperature and estimated CRT.

	Spearman Correlation (ρ)	
	All participants	**Average of each participant**
CapApp	-0.74	-0.79 ± 0.21
Manual Inspection	-0.75	-0.81 ± 0.11
Algorithm-Assisted	-0.72	-0.75 ± 0.18

4.5.1 Relationship to Temperature

Fig. 4.4 shows the relationship between peripheral body temperature and CRT as captured by CapApp. Prior literature has shown that the relationship between the two follows a quadratic relationship of which vertex lies around normal hand temperature of 32°C. As hand temperature decreases, CRT shows quadratic increase because the peripheral perfusion becomes more restricted. At the same time variance of CRT also increases with temperature decrease because the peripheral perfusion becomes irregular.

To further quantify the nonlinear relationship between temperature and CRT, we measured the correlation between the two across all data samples using Spearman's rank correlation coefficient. We also computed the correlation between temperature and CRT for each individual's data and then report the aggregate of those values since there are physiological differences that can lead to different trends (e.g., hydration, body mass). The results of this analysis are presented in Table 4.2. CapApp achieved a correlation of -0.72 across all participants and an average personal correlation of -0.83 ± 0.12. In other words, CapApp is able to accurately estimate CRTs according to human physiology, capturing longer CRT for lower hand temperature and vice versa.



(a) Healthy Participants

(b) Raynaud's Syndrome Participants

Figure 4.5: Boxplot with all three methods comparing CRT against time.

4.5.2 Comparison Between Measurement Techniques

Fig. 4.5 illustrates how all three measurement techniques tracked changes in CRT. We emphasize that CapApp does not capture the exact same phenomenon as the clinical gold standard. CapApp measures CRT using PPG signals, whereas other methods use color changes on the skin. Furthermore, CapApp' and other methods' CRT measurements are measured separately, whereas manual inspection and algorithm assisted measure CRTs using the same video recording. So, we did not expect CapApp's measurements to align with those from the other techniques. Nevertheless, we found that CapApp and manual inspection performed comparably to one another according to the correlation between their CRT measurements and peripheral body temperature. Respectively, these methods achieve correlations of -0.74 and -0.75 across all participants and average personal correlations of -0.79 $\pm$ 0.21 and -0.81 $\pm$ 0.11. Table 4.3 shows that the these two methods are highly correlated (ρ=0.90). Since CapApp and manual inspection measurements are independent and infeasible to pair to the corresponding measurements, we take the average CRT in each temperature buckets (e.g.,

Table 4.3: Performance comparisons between three different techniques for measuring CRT: CapApp, manual inspection, and an algorithm-assisted assessment. Note that the latter two measure a phenomenon that is correlated but different from the one being measured by CapApp.

Candidate Method	**"Ground Truth" Method**	**Correlation (ρ)**	**Mean Difference (s)**	**Mean Absolute Difference (s)**
Algorithm-Assisted	Manual Inspection	0.90	1.85 ± 2.61	2.23 ± 2.37
CapApp	Manual Inspection	0.85	1.59 ± 3.58	2.81 ± 3.23
CapApp	Algorithm-Assisted	0.79	0.50 ± 3.41	2.64 ± 2.50

20°C – 30°C) to compute the correlation.

The algorithm-assisted approach performed worse than the aforementioned techniques, achieving an overall correlation of -0.72 across all participants and an average personal correlation of -0.75 ± 0.18. Despite our best attempts to control the ambient lighting and exposure of the camera, we suspect that even the smallest motion artefacts and shadows negatively impacted the CRT estimates from this technique. We hypothesize that the experts annotators were able to ignore these issues via the phenomenon of color constancy [174]. Meanwhile, CapApp derived its measurements from blood flow underneath the tissue, rendering it less sensitive to ambient lighting. Furthermore, capillary does not refill uniformly within the blanched area; white spots in the finger indicates areas that show slow capillary refill. Although the algorithm-assisted method takes an average value of center region of the finger, the non-uniform capillary refill is not completely captured. Additionally, we found that it is highly correlated with the manual inspection; we suspect it is mainly due to the fact that both measurements are using the same video.

We further report the agreement between measurement techniques across all temperature conditions in Table 4.3. Since manual inspection and the algorithm-assisted assessment both rely on the visible blanching of the outward-facing tissue from the same video recordings, those two techniques resulted in the highest concurrence. The correlation coefficient between

Table 4.4: CapApp's performance comparisons against manual inspection and algorithm-assisted assessment for different CapApp's CRT thresholds.

CapApp Threshold (δ)	**"Ground Truth" Method**	**Correlation (ρ)**	**Mean Difference (s)**	**Mean Absolute Difference (s)**
10%	Manual Inspection	0.85	1.59 ± 3.58	2.81 ± 3.23
20%	Manual Inspection	0.85	0.17 ± 2.85	2.03 ± 2.30
25%	Manual Inspection	0.85	-0.45 ± 2.64	1.86 ± 2.11
30%	Manual Inspection	0.85	-0.95 ± 2.53	1.86 ± 2.08
10%	Algorithm-Assisted	0.79	0.50 ± 3.41	2.64 ± 2.50
20%	Algorithm-Assisted	0.79	-1.41 ± 3.16	2.67 ± 2.41
25%	Algorithm-Assisted	0.79	-2.03 ± 3.23	2.87 ± 2.61
30%	Algorithm-Assisted	0.79	-2.53 ± 3.34	3.11 ± 2.86

them was 0.90 and the mean absolute difference between them was 2.23 ± 2.37. Manual inspection slightly underestimated CRT relative to the algorithm-assisted assessment, which could be attributed to the fact that the human's color acuity worse than camera sensor. The CRT measurements from CapApp were also highly correlated with those from manual inspection ($\rho = 0.85$), although the mean absolute difference was higher (2.81 ± 3.23 seconds). CapApp slightly overestimated CRT relative to the manual inspection, which could be due to differences in detecting color changes under and over the skin. To our surprise, CapApp had a worse correlation with the algorithm-assisted assessment ($\rho = 0.79$), but showing similar difference of 2.64 ± 2.50 with other comparisons. We suspect the low correlation is caused by (1) CRT measurements are computed from independent trials and (2) low correlation between temperature and the algorithm-assisted method.

4.5.3 Effects of CapApp CRT Threshold

CapApp uses exponential decay function to estimate CRT by calculate the time taken to reach 10% of horizontal asymptote. The return threshold of 10% is used in the previous work [175] for the digitally-assisted method and we used the same threshold. However, in

Table 4.3, we found that CapApp overestimated CRT than human, showing mean difference of 1.59 seconds. We hypothesize the difference is caused by color acuity of camera sensor and human eyes that 10% return threshold has high sensitivity to color changes compared to human annotators. We explored different return threshold levels to adjust CapApp to have similar color acuity of manual inspection. Table 4.4 shows correlation and difference between manual inspection and CapApp. When threshold increases, CapApp is less sensitive to subtle color changes, mean difference decreases as well as mean absolute difference. However, when the threshold is at 30% CapApp is too insensitive to the color changes that it underestimates CRT by around 1 second compared to manual inspection. We found that return threshold of 20%–25% shows similar CRT estimation with reasonable difference. We also evaluated the same comparison with the algorithm-assisted assessment. We found that high thresholds lead to CapApp underestimating CRTs and larger absolute differences. It is attributed to the fact that both methods rely on smartphone camera for detecting color changes and the algorithm-assisted method's return threshold is also fixed at 10%.

4.5.4 Mixed-Effects Model

We further analyzed the effect of temperature on CRT by constructing mixed-effects models. These models also included subject-specific information (skin tone and Raynaud's status) to identify potential biases in the various measurement techniques. To facilitate statistical comparisons and visualization, we aggregate temperature into seven equally sized buckets and treat temperature as categorical variable. We fit one model per measurement technique with temperature and skin tone as fixed effects. We also included a random slope and intercept for each participant to account for other user-specific biases like age and body composition that affect baseline and changes in CRT due to temperature. The equation for the mixed-effects model was as follows:

$$\begin{aligned} CRT \sim & (temperature * raynauds\_status) + skin\_tone \\ & + (temperature * raynauds\_status \mid participant\_ID) \end{aligned}$$

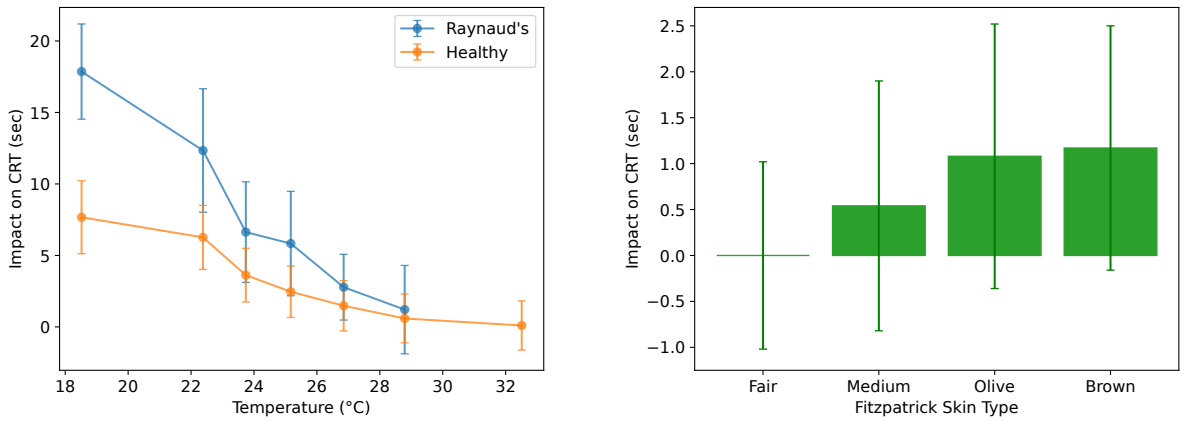


Figure 4.6: Effect of Different Fields from Mixed Effects Model

Fig. 4.6 specifically illustrates the model fit for CapApp. The results are normalized such that the residual with the lowest value has a contribution of 0. Confirming our previous correlational analyses, the left graph of Fig. 4.6 shows that colder temperatures lead to higher contributions in CRT. Each data point represents the center of seven equally sized bin in terms of the number of the data point present in the bin. We see that, for the healthy participants, the CRT value constantly increases, by about 5.67 seconds, as the mean temperature decreases from 28.8°C to 22.4°C; the standard error also increases from 1.71 to 2.24 seconds for the corresponding the temperature change. On the other hand, for the participants with Raynaud's syndrome, the CRT value more drastically increases, by about 9.56 seconds; the stand error also increases from 3.09 to 3.33 seconds. These findings align with human physiology. CRTs tend to get more inconsistent and longer as hand temperature decreases [175]. And, blood vessels are more constricted for Raynaud's syndrome patients as temperature decreases, showing significantly slower and inconsistent CRTs than healthy people [48]

The right graph of Fig. 4.6 shows the impact of skin tone on CapApp's estimates. CapApp tends to show higher CRT values for participants with darker skin tone; CRTs estimated

Table 4.5: Estimates and standard error estimated by a mixed-effects model.

Variable	Value	CapApp (s) ($R^2 = 0.60$)		Manual Inspection (s) ($R^2 = 0.68$)		Algorithm-Assisted (s) ($R^2 = 0.78$)	
		Normal	Raynaud's	Normal	Raynaud's	Normal	Raynaud's
Temperature	18.5°C	7.67 ± 2.55	17.86 ± 3.33	3.78 ± 0.94	11.56 ± 2.61	5.17 ± 1.37	17.74 ± 3.13
	22.4°C	6.26 ± 2.24	12.34 ± 4.32	2.89 ± 0.91	9.26 ± 2.80	3.78 ± 1.23	12.27 ± 2.78
	23.7°C	3.62 ± 1.88	6.63 ± 3.52	2.32 ± 0.92	4.42 ± 1.38	2.58 ± 1.24	8.52 ± 2.51
	25.2°C	2.46 ± 1.80	5.83 ± 3.65	1.47 ± 0.89	2.73 ± 1.09	1.54 ± 1.16	5.35 ± 1.80
	26.9°C	1.48 ± 1.80	2.78 ± 2.30	0.97 ± 0.88	2.60 ± 1.13	1.26 ± 1.15	3.65 ± 1.58
	28.8°C	0.59 ± 1.71	1.21 ± 3.09	0.59 ± 0.87	0.00 ± 1.64	0.45 ± 1.14	3.48 ± 2.02
	32.5°C	0.10 ± 1.72		0.16 ± 0.87		0.00 ± 1.13	
Skin Tone	Fair	0.00 ± 1.02		0.26 ± 0.49		0.32 ± 0.68	
	Medium	0.54 ± 1.36		0.70 ± 0.65		0.46 ± 0.88	
	Olive	1.08 ± 1.44		1.16 ± 0.67		1.19 ± 0.93	
	Brown	1.17 ± 1.33		0.55 ± 0.65		0.49 ± 0.88	

from "Brown" skin tone are 1.19 seconds longer compared to "Pale" or "Fair" skin tone. The standard error for the CRT estimation also increases from 1.01 to 1.33 seconds. Since CRT is measured based on skin color changes, it is found to be affected by skin thickness and skin tone [138]; at the same time, the degree of differences is not well-studied. This finding is an important findings that research community in both medicine and engineering to take account of when building automated CRT assessment or other skin color-based physiological signals.

Table 4.5 compares the model parameters for all three measurement techniques. The three models' fits achieved coefficients of determination (R^2) [135, 134, 85] between 0.60 and 0.78, indicating moderate goodness of fit; highest and lowest are respectively the algorithm-assisted and CapApp. Figure 4.7 specifically shows the impact of temperature on CRT estimation of CapApp, manual inspection, and algorithm-assisted assessment. CRT estimated by CapApp is most sensitive to temperature changes for both healthy and Raynaud's syndrome participants, respectively showing CRT changes of 7.57 and 16.65 seconds from highest to lowest temperature. The manual inspection shows 3.62 and 11.56 seconds of

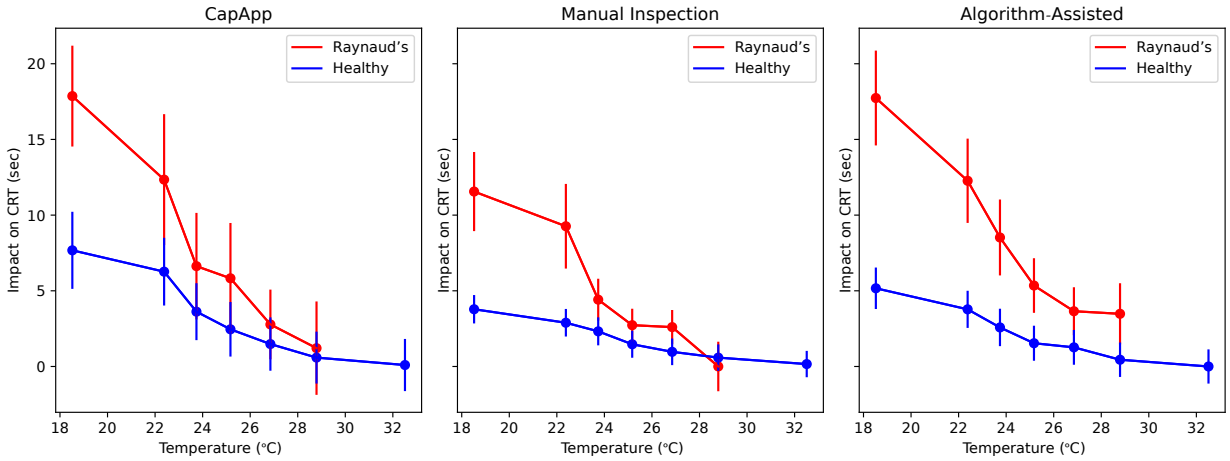


Figure 4.7: Effects of temperature on CRT for different techniques

CRT changes, while the algorithm-assisted method shows 5.17 and 14.26 seconds. Although the algorithm-assisted method shows distinguishable gap between healthy and Raynaud's syndrome participants, it is attributed its lower sensitivity to temperature changes for healthy participants. Additionally, the algorithm-assisted method generally shows higher CRT estimates for Raynaud's syndrome participants at normal hand temperature of 28.8°C. This finding does not reflect physiology of the Raynaud's syndrome, where patients often show normal blood circulation at warmer hand temperature, implying that the algorithm-assisted method generally overestimates CRTs for Raynaud's syndrome participants. In general, CapApp shows higher standard error. We suspect that detecting blood volume changes under the skin using PPG signal is not only sensitive to subtle change but also more variable. Although the other methods rely on less sensitive and more consistent signal, CapApp can achieve comparable performance in CRT estimation. For skin tone, all three methods show similar impacts of 1.16 – 1.19 seconds of gap between lighter and darker skin tones.

4.6 Discussion

CRT is a crucial vital sign for assessing dehydration and shock, yet the lack of a convenient and standardized method for measuring it has hindered its utility [14, 153]. We address this gap by leveraging a combination of mobile sensing techniques to create a smartphone-based CRT measurement tool. In the following section, we discuss the limitations of our approach, opportunities for improving CapApp, and potential areas for future exploration.

4.6.1 CapApp Versus Manual Assessment

Clinical literature currently suggests that CRTs exceeding 2 seconds warrant further investigation [14], but this heuristic has also been questioned due to the subjectivity and inconsistency inherent to manual assessment [153, 7, 94, 158]. Although prior literature has proposed dedicated hardware for CRT assessment, such devices have yet to see the wide-scale adoption that would be required for clinical trials. CapApp leverages the sensors that already exist within smartphones, making it easier to give people access to repeatable and objective CRT assessment. Therefore, we are hopeful that CapApp could enable explorations into the clinical utility of this vital sign.

We anticipate that the existing heuristics for CRT may not be applicable to measurements taken by CapApp. Whereas normal CRT assessment relies on identifying visible color changes in the skin, CapApp measures changes in the underlying blood volume to estimate CRT. We explored the potential of an automatic CRT assessment process that quantifies skin color change, but found that it was difficult to account for the non-uniform change in skin color and subtle changes in lighting when the position of the finger moved relative to the camera. Although we observed moderate correlation between measurements captured by CapApp versus manual methods, CRTs measured by CapApp tended to be longer. We hypothesize the difference can be attributed to two factors. First, the capillaries are the first part of the finger that get refilled, and the amount of blood required to refill the entire fingertip is significantly more than that which is required to refill the capillaries. Second, CapApp

measures CRT based on when the normalized PPG signal returns to 10% of its maximum value. This threshold was selected based on prior work [175] but could be adjusted to achieve stronger agreement with existing measures. Changing the threshold would also impact the sensitivity and specificity of CapApp, allowing for it to be adapted depending on the intended use case and the costs associated with incorrect decisions.

4.6.2 Confounding Factors for CapApp

PPG Signal Acquisition

Although we designed CapApp to reduce some of the variance associated with the typical CRT assessment process, it is still susceptible to many confounds. Smartphone-based PPG assessment requires that the user completely covers the camera and flash with their finger. Failure to do so can result in additional light from the ambient environment reaching the camera, contributing noise to the PPG signal. Since CapApp requires that the user adjusts their fingertip in order to apply and release pressure, that subtle motion can also influence the PPG signal beyond changes in blood volume. Changes in the ambient lighting itself can degrade signal quality as well, although this confound should not be significant since most of the light received at the camera comes from internal reflection at the fingertip.

Our analyses revealed that all three CRT assessment techniques were susceptible to biases according to people's skin tone, with darker tones resulting in longer CRT measurements. This result was not surprising since prior work has identified similar issues with traditional PPG sensors like the heart rate monitors built into smartwatches [204, 83]. Although we were able to use a mixed-effects model to quantify the average impact of skin tone bias in CapApp, future work could implement techniques for combatting this undesirable confound. One way to do this would be to separately examine the color response across multiple wavelengths as suggested by [204, 83].

Pressure Sensing

The pressure sensing component to CapApp could be susceptible to its own set of confounds. As noted by Tung et al. [190] and Goel et al. [60], smartphone manufacturers employ different kinds of vibration motors, exhibiting either linear or rotational vibration patterns. Smartphones with a linear vibration pattern that is orthogonal to the optical axis of the camera are not conducive to CapApp since pressure by the finger will not sufficiently dampen the vibration. Smartphone models can also differ in their relative layout of the vibration motor, camera, and accelerometer. The closer the accelerometer is to the vibration motor, the stronger the vibrations the accelerometer is able to detect. The closer the camera is to the vibration motor, the more damping that the finger can exert. Both of these factors affect the dynamic range and sensitivity of the force estimation calculation, respectively. The vibration signal measured at the accelerometer can also be negatively impacted if the user adjusts the smartphone's position as they perform the test. CapApp can detect obvious motion artefacts, yet subtle adjustments like changing one's grip while holding the phone can reduce the degree of dampening by the fingertip itself.

Study Design

Our protocol had unique characteristics that may have impacted the significance of our findings. We induced changes in CRT by adjusting participants' peripheral body temperature. CapApp requires that the screen is turned on at its maximum brightness to increase the signal-to-noise ratio of the PPG signal, which results in warming of the smartphone over long periods of time. Therefore, we saw larger changes in peripheral body temperature when participants used CapApp (2.00°C $\pm$ 1.36°C) versus when they manually exerted pressure on their own fingertip (1.29°C $\pm$ 1.13°C). We reported all of our results using the mean temperature as a compromise between other alternatives.

4.6.3 Augmenting the Smartphone to Improve CapApp

Although our goal was to enable convenient and repeatable CRT assessment without the need for additional hardware, one way to address the aforementioned confounds would be to incorporate an attachment to the smartphone. Shinozaki et al. [175] and Blaxter et al. [18] created dedicated hardware for CRT assessment, utilizing an air pump to apply pressure on the finger and measuring CRT with a pulse oximeter. The smartphone could take the place of the pulse oximeter in their design, but the air pump would still need to be powered in order to maintain just the right amount of pressure on the fingertip. Instead, we envision a completely passive attachment that relies on a class II lever (similar to a nail clipper with a spring between the bar and the face of the smartphone) to assist in applying and maintaining consistent pressure. The attachment could also help cover the fingertip around the camera, reducing the potential impact of ambient lighting. Such an attachment could easily be fabricated in bulk using 3D printing or other digital fabrication methods.

4.7 Conclusion

In this work, we proposed CapApp, a smartphone application that leverages existing device components for CRT assessment. CapApp simultaneously operationalizes the visual assessment of blanching and capillary refill while guiding users to follow the correct procedure via real-time feedback for how the user applies and releases pressure from their fingertip. We evaluated CapApp with 30 participants showing that CRTs estimated by CapApp align with physiological changes due to temperature changes. Furthermore, we demonstrated that CapApp can extract meaningful CRT differences between people with and without Raynaud's syndrome. Beyond showing the feasibility of using smartphones for point-of-care CRT assessment, we hope that our work represents the first steps towards new clinical discoveries with respect to screening cases of circulatory shock.

Chapter 5

RELIABLE AND TRUSTWORTHY MACHINE LEARNING FOR HEALTH USING DATASET SHIFT DETECTION

5.1 Introduction

Advances in artificial intelligence and machine learning have made medical diagnostic and screening tools more accurate and accessible. AI-powered diagnostic tools [8, 35] are intended to assist medical personnel by making unbiased decision based on thousands of examples. In recent years, these models [36, 108, 118, 79] are even becoming available to consumers through the growth of mobile health with the intention of expediting diagnoses through increasingly frequent testing. Moreover, mobile health [5, 120] aims to improve access to medical expertise for those who are uninsured or live far away from hospitals.

Despite the potential benefits of health AI systems, there are concerns about their performance in real-world settings. Data-driven models learn from examples, making them heavily reliant on the data upon which they have been trained. However, datasets often fail to get complete coverage over a domain, particularly for emerging datasets; when new pulmonary diseases (e.g., MERS and COVID-19) emerge, a pulmonary classifier trained on the existing lung sounds would not be able to interpret sound of the new diseases. Previous work [106, 209] has found that machine learning models behave unpredictably on the unseen data. This problem [5, 183] is especially critical for medical diagnostic and screening tools since there are significant repercussions for mistakes.

Researchers have proposed methods to estimate the uncertainty of a machine learning models' predictions based on the input [72, 105, 102, 169, 107]. Out-of-distribution detection methods can distinguish whether the input lies within the distribution of the training dataset, with out-of-distribution data leading to less reliable prediction results. However, such important information has not been widely explored in the context of health applications.

When health applications are put into the hands of consumers with limited understanding of the underlying algorithms, they may upload poor quality data that lies outside the distribution of the data that was collected by experts. For example, consumers who are using a health application that involves image processing may take photographs in poor lighting conditions or framing of the target object. Even when the data is high-quality, it may be captured with a smartphone that has different hardware specifications than the devices that were used to collect the training dataset. Unless the models are explicitly designed or trained to detect invalid data, the models will incorrectly produce a diagnostically meaningless result.

In this work, we explore the utility of out-of-distribution detection for improving model performance and user-perceived trustworthiness of health-related models. We first benchmark our approach using publicly available deep learning models relating to various medical challenges and sensing domains — images for skin lesion classification, motion data for Parkinson's disease severity, and audio for lung sound classification. After demonstrating that these models are susceptible to dataset shift, we demonstrate that the state-of-the-art out-of-distribution detectors can effectively exclude such data with over 95% detection accuracy in most cases. We then explore the implications of this detector on user-perceived trustworthiness of the health models. After translating the out-of-distribution score into a human-interpretable metric, CONFIDENCE SCORE, we found that showing this information to end-users improved the user-perceived trustworthiness of the models. Furthermore, participants stated that they were more willing to make medical decisions based on models when they were shown the certainty metric. Our contributions in this work are as follows:

- We identify and quantify the limitations of current health deep learning models when encountered with unseen data,

- We evaluate the utility of out-of-distribution detection on various data types (e.g., image, audio, motion) for medical screening and diagnosis, and

- We evaluate the impact that dataset shift information has on user-perceived trustwor-

thiness of health diagnostic results.

5.2 Related Work

5.2.1 Machine Learning-based Health Screening and Diagnosis

In recent years, machine learning has been widely used for medical diagnosis and screening tool to help doctors diagnosis patients easier, faster, and more accurately. Machine learning models that learn from large-scale medical datasets are able to detect various symptoms and conditions, including mental health [59, 182], retinal disease [35], lung cancer [8]. With the increasing ubiquity of smartphone and advances in its computing power, machine learning-based health screening can be done on mobile devices. Various machine learning-based mobile health applications have been proposed to detect health conditions (e.g., traumatic brain injury [119], pancreatic cancer [118], jaundice [36]) and vital signals (e.g., heart rate [108], respiratory rate [108], heart rate variability [79], blood pressure [170], SpO2 [76]). Such mobile health applications can benefit nurses, health workers, and the general population for easier medical screening.

While the health machine learning models show high accuracy on their own test datasets, their performance is questionable in real-world settings where the input data can vary drastically, resulting in unreliable prediction results [185, 162]. Researchers have investigated the dataset shift problem for medical imaging (e.g., x-ray [26, 24], fundus eye images [26], CT scans [192], dermatology [162, 146]), focusing on developing and evaluating out-of-distribution detection methods for specific domains. However, as more consumer-facing health applications are available in the market, this issue can lead the users to make medical decision based on incorrect results. In this work, we aim to explore ways to leverage dataset shift information to make the health machine learning models more reliable and trustworthy to the users.

5.2.2 Dataset Shift Detection

Recently researchers have proposed various methods to estimate the models' uncertainty due to dataset shift. The proposed methods leverage the output of the models to effectively detect *out-of-distribution* input that are different from the known distribution, *in-distribution*. Softmax confidence [72] has been the baseline for the out-of-distribution detection. Several work has been proposed for out-of-distribution detection using deep ensemble [98], Mahalanobis distance [102], Gram matrices [169], energy score [107], temperature scaling [105, 169], input perturbation [105, 102], mean and variance of channels activations [156]. Alternate training strategies [73, 101, 116, 127] have been proposed to enable model to detect out-of-distribution. Generative models [136, 159, 129, 208] are proposed to detect out-of-distribution examples. However, many approaches require re-training and re-designing of the models and prior knowledge of out-of-distribution datasets; it is not realistic to apply these methods to the existing models. In this work, we explore Mahalanobis distance- [102], Gram matrices- [169], and energy-based [107] out-of-distribution detection methods for reliable and trustworthy machine learning for health since these methods show reasonable out-of-distribution detection performance, do not require retraining or prior knowledge of out-of-distribution datasets, and work on pre-trained discrimitive classifiers.

Trustworthy AI Machine learning systems are deployed in real-world settings to billions of users, making significant impacts on high-stake decision making such as healthcare, policy, economy, and transportation. Failures in machine learning systems can cause fatal consequences and building trustworthy AI is one of the most important problems in machine learning community. In recent years, there are active and ongoing efforts aimed at making machine learning systems causal [9, 139], explainable [2, 93, 112, 84, 89], fair [3, 157, 43], robust [44, 50, 71, 45, 188], and privacy-preserving [1, 148, 124, 187]. This work contributes to trustworthy AI by improving reliability and user-perceived trustworthiness of machine learning for health using estimated uncertainty. Bhatt et al. [16] proposed to leverage uncertainty for users making decision and placing trust in machine learning models. This

work explores similar approach where we adopt out-of-distribution detection as a method to measure uncertainty. We took a step further to investigate and quantify its effect in improving reliability and trustworthiness in the context of machine learning for health.

5.3 Background: Dataset Shift Detection Methods

We aim to leverage state-of-the-art out-of-distribution detection methods [102, 169, 107] in the health domain for users to safely use health deep learning models. We selected three out-of-distribution methods that show high accuracy on different out-of-distribution datasets, do not require re-training and prior knowledge of out-of-distribution datasets, and work on pre-trained discrimitive classifiers. These characteristics are important to help developers or other stakeholders (e.g., regulators, auditors, platforms) easily adopt out-of-distribution detectors to any pre-trained models. In this section, we provide background on each out-of-distribution detection methods.

5.3.1 Mahalanobis Distance-Based Out-of-Distribution Detection

Mahalanobis distance is used to measure the proximity of a point to a certain Gaussian distribution. In the Mahalanobis distance-based out-of-distribution detector [102], this property is used to represent each class's samples at each layer of a network as a class conditional Gaussian distribution with mean $\hat{\mu}_{cl}$ and co-variance $\hat{\Sigma}_{cl}$, where c indicates the class and l indicates the layer in the model. Given a sample input x to a neural network, for each layer, it computes the minimum layer-wise class conditional Mahalanobis distances for x. That is, for each layer, it finds the Mahalanobis distance associated with the closest class to x. In other words, this is equivalent to $M(x) = \max_c -(f(x) - \mu_c)^T \Sigma^{-1} (f(x) - \mu_c)$. The authors have demonstrated that adding small noise to the input can help better distinguish between in-distribution and out-of-distribution data. As the authors suggested, for the real-world setting where the out-of-distribution datasets are generally not available, we obtain the input noise magnitude by generating adversarial samples generated by FGSM [62].

5.3.2 Gram Matrices-Based Out-of-Distribution Detection

Gram matrices are used to compute pairwise feature correlations and encode stylistic attributes. For out-of-distribution detection [169], higher order Gram matrices are used to compute class-conditional bounds of feature correlations at all hidden layers of the network as higher order shows better out-of-distribution detection performance. Higher order Gram matrices is expressed as $G_l^P = (F_l^P F_l^{P^T})^{\frac{1}{P}}$, where F_l is feature map at l-th layer and P is order. All elements of Gram matrices of an input at each layer are compared against the prepossessed minimum and maximum Gram matrices element values from in-distribution dataset to obtain deviation. If the input data is predicted as a certain class, the minimum and maximum values of the corresponding class will be used for comparison. The comparison is done for each layer to obtain layerwise deviations. Then, the deviations are used to get a total deviation, which is defined by the normalized sum of layerwise deviations. Whether the input data is from out-of-distribution is determined with a threshold which is defined as 95% percentile of the total deviations of in-distribution energy score distribution.

5.3.3 Energy-Based Out-of-Distribution Detection

The energy-based out-of-distribution detector [107] seeks to provide an alternative scoring function to the softmax function that is less susceptible to over-confidence and therefore can better distinguish between in and out-of-distribution inputs. It takes a discriminative classifier $f(c)$ that maps input $x \in R^D$ to logits, which are traditionally used to derive a categorical confidence score using a softmax function. It defines the energy function on the classifier as $E(x; f) = -T \cdot log \sum_i^K e^{f_i(x)/T}$, where K is the number of classes in the model's output space and T is a temperature parameter that can be used to alter the shape of the energy score distribution. Energy score threshold that distinguishes between in- and out-of-distribution samples is defined at the 95% percentile of the in-distribution samples.

5.4 Performance Evaluation

In this section, we demonstrate the performance degradation of the existing health models when encountered with out-of-distribution datasets highlighting that the existing models are vulnerable to dataset shift. We then evaluate the performance of state-of-the-art out-of-distribution detectors for distinguishing between in- and out-of-distribution examples. We have selected out-of-distribution datasets that consist of both near- and far-from-distribution samples that represent realistic use cases in the real-world settings. For mobile health applications that use mobile sensors for health screening, non-expert users are expected to input data collected by themselves. Unlike clinicians, who may either receive training on how to operate these mobile apps or may already understand what must be done to generate high-quality, non-expert consumers may collect relevant but low-quality data due to environmental factors or totally irrelevant data by mistake or a lack of understanding. To reflect these scenarios, we include out-of-distribution datasets caused by covariate shift, label shift, and open-set recognition. The covariate and label shifts aim to evaluate a model's performance when tested on data pertaining to the same classification task but from different data sources and environment. Open-set recognition evaluates the model's performance on new classes not included in the training set. In Table 5.2, we indicate dataset shift type for each out-of-distribution dataset.

5.4.1 Models and Datasets

Skin lesion A DenseNet-121 based skin lesion classifier [146] was used in this work. The model aims to classify an image into seven different skin lesions: actinic keratoses, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, melanocytic nevi and vascular lesions. The following datasets are used for training and evaluation:

- (**In-distribution**) HAM10000 [189, 32]: (CC BY-NC 4.0) A dataset containing 10,000 samples of dermatoscopic skin tumor images taken using different devices and from different populations. These tumors are part of 7 classes: actinic keratoses, basal cell

carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.

- (**Out-of-distribution**) ISIC 2017 [33]: (CC BY-NC 4.0) A previous version of the HAM100000 dataset which contains 2000 dermatoscopic skin tumor images labelled for binary classification. A tumor is labelled malignant if it corresponds to melanoma to benign if it corresponds to nevus or seborrheic keratosis.

- (**Out-of-distribution**) Face [37]: (CC BY 4.0) A dataset containing frontal view face images of 102 adults without making a neutral facial expression. Face images are personally identifiable information. But, all individuals gave signed consent for their images to be "used in lab-based and web-based studies in their original or altered forms and to illustrate research (e.g., in scientific journals, news media or presentations)."

- (**Out-of-distribution**) CIFAR-10 [96]: (MIT License) A common image classification benchmark with 10 non-medical classes (airplane, car, cat, dog, horse, bird, deer, ship, frog, truck) which contains 6,000 images per class.

Lung Sound A lung sound classification model [52] classifies normal lung sound, wheeze, and crackle from an audio sample. This model is based on ResNet-34 and uses spectograms of audio samples as inputs and outputs 4 lung sound classes (normal, wheezing, crackle, and wheezing + crackle).

- (**In-distribution**) ICBHI 2017 Respiratory Challenge [160]: A dataset collected using multiple microphones and stethoscopes containing 6898 samples normal lung sound, wheeze, and crackle audio

- (**Out-of-distribution**) Stethoscope [51]: (CC BY 4.0) A dataset containing stethoscope respitory sounds with 336 samples of normal breathing, wheeze, and crackle audio sounds. The dataset was collected using a 3M Littmann Electronic Stethoscope.

- (**Out-of-distribution**) AudioSet [55]: (CC BY 4.0) A large dataset of millions of sound labelled YouTube audio of which a portion of the dataset contains breathing, cough, and wheezing samples which we use to create a suitable out-of-distribution dataset for this model.

Parkinson's Disease This is a binary classification model [207] that showed highest performance in Parkinson's disease digital biomarker DREAM challenge [176]. The model uses accelerometer signals to detect tremors in a person's movement and outputs whether a participant has Parkinson's. This model consists of 5 1D-convolutional layers and a single output.

- (**In-distribution**) mPower [19]: (CC BY 4.0) A dataset contains 30-second accelerometer readings from 3,100 participants at rest for both healthy and Parkinson's patients. The dataset was used in Parkinson's disease digital biomarker DREAM challenge [176].

- (**Out-of-distribution**) Kaggle Parkinson's disease [58]: (CC0 1.0) A dataset with accelerometer readings from healthy participants simulate movements of Parkinson's patients.

- (**Out-of-distribution**) MotionSense [115]: (MIT License) A dataset contains accelerometer readings from 24 participants performing various activities (e.g., walking, jogging, sitting, standing, etc).

- (**Out-of-distribution**) MHEALTH [13]: An activity classification dataset which contains accelerometer readings from 10 participants executing various activities (e.g., standing, sitting, walking, cycling, etc).

5.4.2 Performance Impact by Dataset Shift

In evaluating the model's performance on the out-of-distribution dataset, we used pre-trained models from the previous work[1] [52] when the authors make it available. Otherwise, we trained the model in the same way specified in their work[2] [146, 207]. We trained skin lesion model [146] for 150 epochs using Adam optimizer with a learning rate of 0.0001 and weight decay of 0.2. For Parkinson's disease model [207], we trained for 50 epochs using Adam optimizer with a learning rate of 0.0005. The pre-trained lung sound model [52] is trained for 200 epochs using SGD optimizer with a learning rate of 0.001 and momentum of 0.9. For all of these models, we used an 80/20 split and applied the same preprocessing for train and test sets. All training and testing is done in a server (Intel Xeon 2.1GHz, 64GB, GeForce RTX 2080 Ti) from an internal cluster. We then ran inference on each dataset and calculated the classification accuracy for the datasets that have corresponding labels. For the datasets that do not have the same labels from the in-distribution, the accuracy could not be computed. Table 5.1 summarizes the classification accuracy for the models on in- and out-of-distribution datasets.

We generally observed a significant performance drop for all health machine learning models that are tested with out-of-distribution datasets. The models output unreasonable and arbitrary predictions on datasets that are not related health conditions. For example, skin lesion classifier predicts all face images as vascular lesions and CIFAR10 images as various types skin lesions. Similarly, Parkinson's disease classifier predicts significant portion of physical activities by health participants as tremor caused Parkinson's disease. For lung sound classification, ordinary sound events (e.g., speech, walking, laughing) are classified as a certain type of lung sounds (e.g., crackle, wheezing). When the models are evaluated on out-of-distribution datasets that have similar data characteristics to the in-distribution data(i.e., near-distribution datasets), all health models exhibit a performance decrease that

[1] `https://github.com/microsoft/RespireNet`

[2] `https://github.com/GuanLab/PDDB`

Health ML Models	**In-Distribution**	**Out-of-Distribution**		
Skin Lesion	HAM10000	ISIC 2017	Face	CIFAR
(DenseNet-121)	92.05%	74.00%	N/A	N/A
Lung Sound	ICBHI 2017	Stethoscope	AudioSet	
(ResNet-34)	78.50%	2.10%	N/A	
Parkinson's Disease	mPower	Kaggle Parkinson's	MotionSense	MHEALTH
(5×1D-Conv)	82.01%	26.67%	45.83%	10.00%

Table 5.1: Accuracy of health deep learning models on in-distribution and out-of-distribution dataset. Accuracy is not available (N/A) for out-of-distribution datasets that do not have corresponding labels.

ranges from 18% to 76%. This implies that the models are also sensitive to small dataset shift, such as datasets collected with different devices and in different environments. All of these failure scenarios can occur in real-world deployment of health machine learning applications. Users can input a face image to skin lesion classifier, improperly record lung sound and input ambient sound to the lung sound classifier, or input motion data when they are not at rest to the Parkinson's disease classifier. Furthermore, diverse sensors and devices used in real-world deployment can cause significant performance drop. This evaluation demonstrates that users are exposed to the health machine learning applications that can provide unreliable diagnostic results.

Out-of-Distribution Detection Performance

The previous evaluation implies that it is crucial to determine whether the input data belongs to in- or out-of-distribution to avoid failures caused by dataset shift. In this section, we investigate the feasibility of using state-of-the-art out-of-distribution detection methods in the context of machine learning for health. We evaluate Mahalanobis distance- [102], Gram

matrices- [169], and energy-based [107] methods, which work on any pre-trained discrimitive classifiers and do not need re-training and prior knowledge of out-of-distribution datasets, in detecting out-of-distribution data for different health models.

5.4.3 Experimental Setup

For Mahalanobis distance-based method[3], we extracted Mahalanobis distance-based scores from the output dense and residual block found in DenseNet and ResNet respectively. For the Parkinson's model which does not contain dense and residual blocks, we extracted the scores at the end of each convolutional layer. Then, we optimized the input noise magnitude using in-distribution samples and corresponding adversarial samples generated by FGSM [62]. The noise magnitude obtained is 0.0 for skin lesion classifier, 0.0005 for lung sound classifier, and 0.0 for Parkinson's disease classifier. For Gram matrices-based method[4], we extracted class-specific minimum and maximum correlation values for all orders of Gram matrices for all feature pairs. Total deviation values, which are used for out-of-distribution detection threshold, are computed with multiple sets of random samples from in-distribution datasets. For energy-based method[5], we use their method that does not require fine-tuning to avoid re-training of the network. We use the default temperature scaling ($T = 1$) as suggested in [107]. All evaluations are repeated for 5 trials and we report the mean (Table 5.2) and standard deviation (Table A.2) of all metrics.

5.4.4 Evaluation Metrics

For out-of-distribution detection, it is common to use true negative rate (TNR) at 95% true positive rate (TPR), AUROC, and detection accuracy to evaluate the performance of a detector. Particularly, as the out-of-distribution problem is a binary classification problem, we

[3] `https://github.com/pokaxpoka/deep_Mahalanobis_detector`

[4] `https://github.com/VectorInstitute/gram-ood-detection`

[5] `https://github.com/wetliu/energy_ood`

consider out-of-distribution samples as negative and in-distribution samples as positive. TNR at TPR 95% is defined as the percentage of correctly detected out-of-distribution samples, when 95% of in-distribution samples are correctly detected. The AUROC metric measures the area under the TPR vs FPR curve. The detection accuracy measures the maximum possible classification accuracy over all possible thresholds in distinguishing between in-distribution and out-of-distribution examples. Detailed explanations on the metrics are available in Appendix A.2.1.

5.4.5 Results

Table 5.2 shows out-of-distribution detection performance for different methods across different health machine learning models and datasets. Overall, Mahalanobis distance- and Gram matrices-based out-of-distribution detection methods consistently show outstanding performance across different neural networks and different out-of-distribution datasets, showing TNR @ TPR95 of 95% or above. These methods show lower performance in distinguishing near-distribution datasets (e.g., ISIC 2017, Stethoscope, Kaggle Parkinson's), which aligns with the results from previous out-of-distribution work [169]. On the other hand, the energy-based method did not show reasonable performance in detecting out-of-distribution samples. We found that the energy scores of out-of-distribution samples were not able to effectively discriminated from in-distribution samples as shown in Appendix A.2.2. Note that we used the energy scores without fine-tuning; however, the authors of energy score-based method [107] have demonstrated that a classifier that is fine-tuned using the energy score in place of the softmax score shows significant improvement in out-of-distribution detection performance. This evaluation implies that state-of-the-art out-of-distribution detectors can be applied to health machine learning applications to provide reliable diagnostic results to the users.

				Validation on OOD Samples (TNR @ TPR95/AUROC/Detection Accuracy)		
Health ML Models	**In-Distribution**	**Out-of-Distribution**	**Distribution Shift**	**Mahalanobis**	**Gram**	**Energy**
Skin Lesion	HAM10000	ISIC 2017	Covariate/label shift	10.13 / 58.21 / 59.28	25.90 / 81.14 / 74.98	14.28 / 76.20 / 70.76
(DenseNet-121)		Face	Open-set recognition	100.00 / 99.98 / 99.96	95.01 / 98.20 / 96.34	0.00 / 80.45 / 84.81
		CIFAR10	Open-set recognition	99.83 / 99.90 / 99.61	95.14 / 98.66 / 96.90	5.06 / 58.33 / 57.89
Lung Sound	ICBHI	AudioSet	Open-set recognition	97.96 / 99.47 / 97.34	96.55 / 99.18 / 95.97	8.12 / 56.79 / 57.13
(ResNet-34)		Stethoscope	Covariate/label shift	45.60 / 86.27 / 80.57	41.77 / 83.75 / 76.05	7.29 / 60.98 / 58.94
Parkinson's Disease	mPower	MotionSense	Open-set recognition	100.00 / 99.86 / 99.89	100.00 / 99.94 / 99.60	0.00 / 58.71 / 64.96
(5×1D-Conv)		mHealth	Open-set recognition	100.00 / 100.00 / 100.00	100.0 / 99.99 / 99.99	0.00 / 41.41 / 59.44
		Kaggle Parkinson's	Covariate/label shift	98.00 / 99.89 / 99.47	98.96 / 99.96 / 99.67	70.00 / 95.91 / 93.34

Table 5.2: Out-of-distribution detection performance across different networks and datasets.

5.5 User Study

According to the trustworthy AI literature [46], providing users with interpretable information can enhance the trustworthiness of the result and potentially impact users' decisions. We therefore conducted an online survey-based user study to validate this effect and the impact that our approach has on model trustworthiness. We first defined CONFIDENCE SCORE as how confident the model is in interpreting an input. In other words, in-distribution input would have a high CONFIDENCE SCORE, whereas out-of-distribution input would have a low CONFIDENCE SCORE. We compute CONFIDENCE SCORE by scaling raw out-of-distribution scores from out-of-distribution detectors [102, 169] to 0–100, where 0 is most likely to be an out-of-distribution example and 100 is most likely to be an in-distribution example. Scaling is done in a piecewise manner. When out-of-distribution scores are within an in-distribution threshold, which is set to include 95% of in-distribution examples, we compute min-max scaling that ranges from 90 to 100. In this way, we ensure that most of the in-distribution examples have confidence scores of 90 or above. When out-of-distribution scores outside of an in-distribution threshold, we compute min-max scaling from 0 to 90, where the same denominator is used as above since out-of-distribution examples might not be available in practice and any negative values are clipped to 0. We then investigate the effect of CONFIDENCE SCORE on user-perceived trustworthiness and its impact on medical decisions. Additionally, we also quantify potential learning that can be gained when it comes to distinguishing between in- and out-of-distribution input samples. Specifically, we aim to answer the following research questions:

RQ. 1 How does the CONFIDENCE SCORE affect the perceived trustworthiness of diagnostic results?

RQ. 2 How does the CONFIDENCE SCORE affect medical decisions based on diagnostic results?

RQ. 3 Is there a potential learning effect from CONFIDENCE SCORE when it comes to distinguishing between input data with high and low CONFIDENCE SCORE?

5.5.1 Study Procedure and Participants

The overview of the online user study is illustrated in Figure 5.1 and a list of the user study interfaces is detailed in Appendix A.2.4. In short, the interface displays simulated results from the health screening models used in Section 5.4 (i.e., models for skin cancer, lung sound, and Parkinson's disease). We made the input data as human-readable as possible to maximize interpretability. Images were shown as is, while audio was included in an audio player so that participants could play, pause, and stop the track. We presented the motion data as a time-series plot of accelerometer signals from x-, y-, and z-axis. Since time-series data can be particularly challenging for people with a limited experience in sensor signals, we explain that high-amplitude signals are associated with fast motion while while low-amplitude signals are associated with slow motion. The interface explained the model's purpose and accuracy, which was fixed to 90% to remove potential bias. For each model, the interface presents prediction results in two different conditions: (baseline) input and result, and (confidence score) input, result, and CONFIDENCE SCORE. For each result, we asked participants how much they trust the model's prediction and whether they would be willing to make a medical decision based on that result. Participants saw a total of 24 scenarios (3 data types (image, audio, motion) $\times$ 2 conditions (baseline vs. CONFIDENCE SCORE $\times$ 2 CONFIDENCE SCORE (high vs. low) $\times$ 2 results (positive vs. negative). To provide realistic experience, we provide different skin tone images for skin lesion samples based on the reported skin tone. With the exception of the data type, the scenarios were shuffled across all other factors to avoid any ordering effects. After participants saw all of the scenarios for a given data type, we presented them with five data examples and asked them to pick the ones that the model would be confident in processing according to CONFIDENCE SCORE. We added these questions to assess whether people were able to learn about how the CONFIDENCE SCORE was being generated after seeing a series of examples.

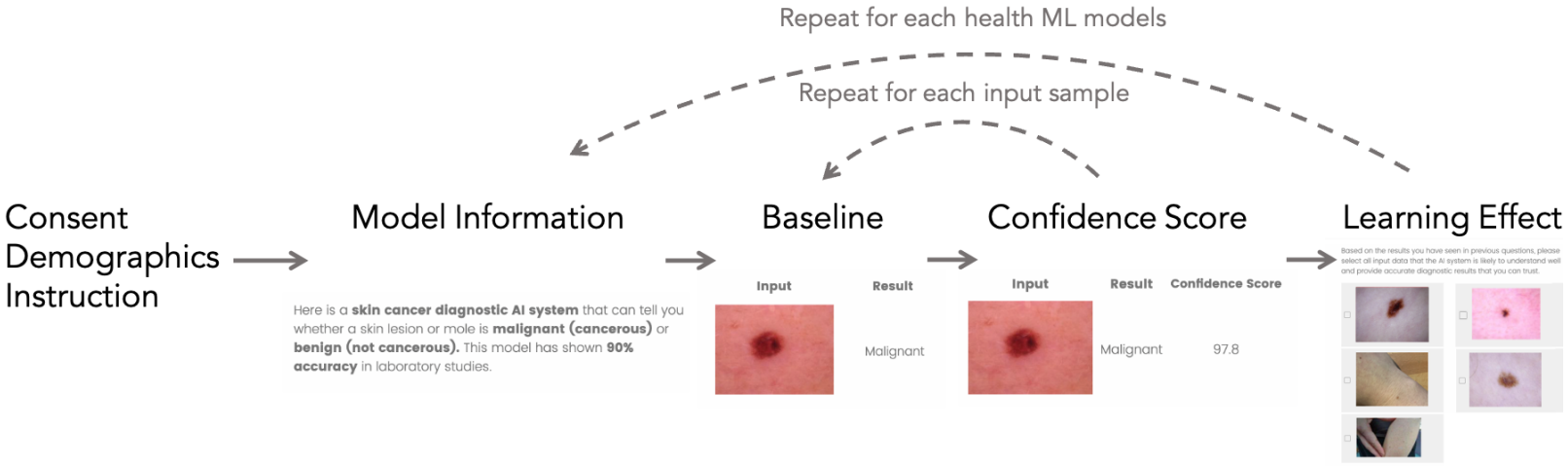


Figure 5.1: User study overview. The participants are first asked to give consent, read instruction, and provide demographics. Then, they report perceived trustworthiness and willingness to make a medical decision after seeing input samples that consist of different data types, diagnostic results, and CONFIDENCE SCORE for baseline and CONFIDENCE SCORE condition. Screenshots of the user study interface are demonstrated in Appendix A.2.4.

We intend to target ordinary, non-expert consumers at random rather than expert clinicians for our study. Our research is primarily directed toward the boom in consumer-facing mobile health applications, where non-expert users are expected to collect input data themselves. We believe this is where models are most susceptible to out-of-distribution inputs, providing unreliable predictions to the users. For the safe use of AI-powered health applications, the users would need support via automated uncertainty measures. To this end, We recruited participants from Amazon Mechanical Turk and compensated with $3 USD for a 15-min online study. In total, 192 participants (155 male, 67 female) completed the online study with an average age of 42.7 ± 9.1 years. The study was approved by Institutional Review Boards at the University of Washington.

5.5.2 Results

We analyzed the responses using the Wilcoxon signed-rank test [200] to compute a pairwise comparison of the categorical responses between the baseline and CONFIDENCE SCORE conditions. Table 5.3 summarizes these statistical results along with the Rosenthal correlation coefficient [161] (r) for effect size.

RQ. 1: User-Perceived Trustworthiness In general, we found that user-perceived trustworthiness ($p < 0.001$) was higher in the CONFIDENCE SCORE condition with medium effect size ($r = 0.393$). In other words, the dataset shift information helped the participants decide when to trust or not to trust the output of the models. Higher CONFIDENCE SCORES led to increasing trustworthiness; high scores had a large effect size ($r = 0.475$), while low scores had a medium effect size ($r = 0.317$). The impact on trustworthiness was similar for positive and negative diagnostic results. The effect sizes varied for the different input data types, with the images having the largest effect size and motion having the smallest. We suspect that the effect size was correlated with the intuitiveness of the data types, with images being more intuitive than motion data.

	User-perceived trustworthiness			Impact on making medical decisions		
	Wilcoxon Test (W)	p	Effect Size (r)	Wilcoxon Test (W)	p	Effect Size (r)
All	529,950.0	$< 0.001^{***}$	0.393	223,227.0	$< 0.001^{***}$	0.178
In-Distribution	138,778.5	$< 0.001^{***}$	0.475	52,056.0	$< 0.001^{***}$	0.200
Out-of-distribution	126,814.0	$< 0.001^{***}$	0.317	59,790.5	0.001^{***}	0.158
Negative result	131,910.0	$< 0.001^{***}$	0.393	51,197.5	0.026^{*}	0.100
Positive result	133,301.0	$< 0.001^{***}$	0.394	60,751.5	$< 0.001^{***}$	0.258
Image	55,890.0	$< 0.001^{***}$	0.436	20,884.0	$< 0.001^{***}$	0.225
Audio	64,440.0	$< 0.001^{***}$	0.384	25,848.0	0.002^{**}	0.173
Motion	56,767.5	$< 0.001^{***}$	0.361	28,084.5	0.019^{*}	0.133

Table 5.3: Results of Wilcoxon test in comparing baseline and CONFIDENCE SCORE conditions for the perceived trustworthiness and impact on decision making. All comparisons show statistically significant results. (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

RQ. 2: Impact on Making Medical Decisions When we examined the impact of CONFIDENCE SCORE on making medical decisions, we found that there was a statistically significant difference ($p < 0.001$) between the baseline and CONFIDENCE SCORE conditions. In other words, participants were more willing to make medical decisions when positive results were presented with high CONFIDENCE SCORE and vice versa. Similar to the results for user-perceived trustworthiness, the effect of CONFIDENCE SCORE was larger on input data with high scores than low scores and highest for images compared to audio and motion data.

RQ. 3: Learning Effect on Distinguishing In- and Out-of-Distribution Input Data We found that the participants were able to learn from their interaction with CONFIDENCE SCORE. The average Jaccard index when it came to selecting high CONFIDENCE SCORE input data was 0.75, 0.66, and 0.64 for image, audio, and motion data, respectively, which is a moderately high similarity. As with our other results, the Jaccard index was highest on images and lowest on the motion data, implying that ability to understand the input data also has impact on learning effect. This implies that the dataset shift information can make users better understand input data that the machine learning models can interpret for the future interaction.

5.6 Discussion

Dataset Shift Information for Health Application Users

Based on the results from our performance evaluation and user study, we can imagine two potential use cases of the dataset shift information to improve safety and trust in mHealth applications. First, mHealth applications with machine learning models can exclude out-of-distribution samples to avoid making inferences and suggestions that are likely to be inaccurate and unreasonable. Second, our user study shows that the dataset shift information can enhance their interaction with the health machine learning applications. It was found to be particularly effective in improving trustworthiness for in-distribution data and leading the users to make the right medical decision. As the users interact with the health applications longer, they would have better understanding of importance of data quality for future

interactions.

Dataset Shift Information for Health Application Developers

Our dataset shift information not only improves the user experience, but also yields potential benefits for model developers. If a user correctly captures data but the model rejects it as being out-of-distribution, then there likely exists intrinsic problems or biases with the model. For example, if a skin lesion classifier is only trained on data from people with pale skin and a user with darker skin submits an image of their own, the out-of-distribution detector be triggered due to the incompleteness of the training dataset. The same issues may occur when training dataset is only collected from a specific set of sensors (e.g., camera, microphone, IMU) with particular specifications.

Limitations and Future Work

Detecting near-distribution samples (e.g., ISIC 2017, Stethoscope, Kaggle Parkinson's) was a difficult problem for all the out-of-distribution detectors we evaluated. For the near-distribution datasets, we evaluated model's accuracy on the data that are distinguished as in-distribution. This issue is actively investigated by researchers and the improved near-distribution detection method would benefit this work.

Our user study was limited in the fact that it dealt with hypothetical scenarios. There were no repercussions for users decisions, so they may not have spent as much time making their decisions as they would in real life. There are also many other factors that impact people's health-related decision making, such as the perceived severity of the medical condition and the perceived benefits of taking action [81, 179]. We tried to make some of the data more realistic by aligning data with the user's demographic information (e.g., we displayed skin lesion images based on their reported skin tone); nevertheless, participants were aware that the data was not their own. Additionally, increased trust might be affected by participants' own understanding and interpretation of the input data. Although we observed increased and decreased trust in examples with high and low CONFIDENCE SCORE, respectively, randomizing

the CONFIDENCE SCORE of input data could further quantify impact of CONFIDENCE SCORE on user trust of health predictions.

In future work, we would like to (1) investigate the best way to present this information for the users, (2) leverage the dataset shift information for finding potential biases in the train dataset and inherent problems with the model, (3) investigate an out-of-distribution method for better near-distribution detection performance.

5.6.1 Conclusion

In this work, we investigated the utility of dataset shift information for improving reliability and trustworthiness of machine learning-based health applications. Using publicly available health deep learning models and datasets, we first demonstrated that the models fail when encountered with unseen data. We then evaluated the out-of-distribution detection performance of state-of-the-art methods, showing high accuracy in distinguishing between in- and out-of-distribution datasets for different input data types (e.g., image, audio, motion data). We conducted an online user study to investigate the effect of dataset shift information on potential users. We found that the participants trusted prediction results with high CONFIDENCE SCORE and are more willing to make a right medical decision, while they considered prediction results with low CONFIDENCE SCORE less trustworthy and are less willing to make medical decision. This work shows that the dataset shift is a meaningful piece of information for building consumer-facing trustworthy AI applications for high-stake decision making.

Chapter 6
IMPLICATIONS AND CONCLUSION

6.1 Summary

In this dissertation, I propose approaches to design feedback for *non-expert* users, which enables them to use mHealth applications accurately and reliably. Through three projects, I provide supporting evidence for the following thesis statement:

> *Real-time and post-hoc feedback can enable non-expert users to acquire high-quality data while using sensor-based mHealth applications, improving the accuracy and reliability of health screening algorithms.*

I presented real-time, sensor-driven feedback for *non-expert* users to improve the quality of the acquired data. In Chapter 3, I described a strategy to leverage real-time image processing techniques to ensure users capture the target images that meet the predefined image quality criteria. Clear instructions and a user interface were also presented to guide users to meet the criteria. In Chapter 4, I described feedback designs for data acquisition that involves complex procedures. The mHealth application closely evaluated user-acquired data and their interaction with the app to ensure each step was correctly performed. Failures in any step could result in invalid data. When such failures are detected, users are asked to reattempt the step.

I also presented feedback for *non-expert* users to get reliable and trustworthy health predictions from "black-box" machine learning algorithms by leveraging state-of-the-art methods for uncertainty estimation of machine learning models. In Chapter 5, I proposed to use out-of-distribution detection in the health machine learning pipeline for (1) dismissing any results with high uncertainty for reliability and (2) augmenting uncertainty score with

the prediction for user trust. It is still possible that, even with the highest quality input data, a machine learning model can fail to provide accurate, reliable results if the data do not lie within the model's training dataset. By protecting users from such failures, model uncertainty is a critical piece of information that should be provided to the users.

In summary, I argue that these different feedback strategies should be considered to be adopted to any mHealth applications that rely on user's input data and machine learning algorithms.

6.2 Implications and Future Directions

mHealth applications [11, 42] that measure vital signs or provide health screening are already available to the general public. There is no doubt that mHealth will be further integrated into people's daily lives and improve access to healthcare outside of clinical settings. It is in the foreseeable future that mHealth applications will become available to more users, leverage more sensors in both existing and emerging mobile devices, and rely on the underlying algorithms that are more sophisticated and complicated. As this trend continues, there are new research directions and potential problems that need to be addressed. In this section, I discuss the implications of this thesis and recommendations for future research in this domain.

6.2.1 Quality Assurance for Heterogeneous Hardware

Projects presented in this thesis are evaluated with a handful of smartphones that are available in the market. In RDTScan, we witnessed performance differences within these smartphones in terms of capture duration and interpretation accuracy. We analyzed that differences in both computation resources and sensor sensitivity play a role in such disparity. In CapApp, we mainly worked with a specific smartphone model to empirically define the thresholds for high-quality sensor signals. The decision is made after running a preliminary experiment with several smartphones and finding significant differences in accelerometer sensitivity and vibration motor placement and power. Such differences are extremely difficult to know in advance; smartphone manufacturers rarely provide details of all the sensors and the datasheet

alone does not provide sufficient information to understand the differences. Even when the quality assurance algorithms work well in the existing smartphones, it is still unknown whether they would show similar performance for new smartphones. The problem would get worse as the applications are deployed in smartwatches or AR/VR headsets.

One possible solution would be that manufacturers provide sample sensor signals with a few predefined benchmark scenarios for sensors. The health applications developers can better understand the differences of sensors from their test and potential target devices to adjust the quality thresholds for diverse mobile devices. This could be also a promising direction to leverage domain adaptation techniques if quality assurance algorithms are built with a data-driven machine learning model. The ML model can be trained with test device data and further fine-tuned and adapted to target devices with a handful of sample sensor data.

6.2.2 Digital Literacy and Inclusiveness

mHealth applications aim to enable people to frequently and easily measure their vital signs and screen for health conditions with mobile devices without requiring clinic visits. I believe people who need frequent medical check-ups and/or have limited access to healthcare would benefit most from mHealth applications. However, in Chapter 3, RDTScan is found to show longer capture duration and lower accuracy when used by older populations than younger populations. RDTScan also showed a similar trend for low-end smartphones and users in low- and mid-income countries than those in high-income countries using high-end smartphones. I believe such a gap is caused by the digital literacy of these populations; they are less familiar with interacting with different sensors in the smartphone. There should be targeted efforts to make mHealth applications more inclusive to people who are less experienced with mobile devices. From RDTScan project, I found that in-person tutorials and training sessions always help get people on board and familiar with mHealth applications. However, it is not a scalable solution if mHealth applications target the general public. An effective onboarding process is proposed in the previous work [23]. In-depth studies [201, 25] show interesting insights on improving usability for older adults and people in low- and mid-income countries.

Incorporating such strategies for people with a limited mobile device or mHealth application experience would support a diverse mHealth application user population. Additionally, the computation-efficient approach would also close the gap independently from usability to ensure the performance is high across diverse smartphones, improving the inclusiveness of mHealth applications for low- and mid-income countries.

6.2.3 Fairness and Bias in Machine Learning for Health

An important failure mode of health applications, which is not addressed in this thesis, is failing to provide unbiased health predictions across a diverse population. It is found that pulse oximeter and PPG signal shows varying signal quality across people with different skin tones, resulting in varied accuracy in heart rate estimation [83, 204]. It is also reported that public x-ray datasets have a very high risk of inducing bias in machine learning models because the datasets contain imbalanced demographics [56, 167]. Such failure could result in a disparity in providing accurate health screening results between different groups of people, further widening the existing healthcare disparities.

One way to mitigate the bias in ML models is to create a balanced dataset. This would ensure that the models have an equal chance of learning features from diverse groups. However, we found that there is an inherent bias within the public dataset that tries to mitigate ML bias. For example, a face image dataset [88] that has a balanced number of images across different races and genders uses binary "gender" classification (e.g., female and male) as a bias evaluation task, where it does not include other non-binary genders. Furthermore, the balanced datasets alone do not address the issue because the rate at which machine learning models learn and generalize varies across different demographics. Unbiased training techniques [6] are also necessary to keep the model learning in an unbiased and fair manner.

The models are often evaluated in terms of classification accuracy or regression error among different demographics to detect and identify potential biases. However, I believe accuracy is just one of numerous factors to evaluate ML models. In Chapter 5, I investigated that the models provide unreliable and unpredictable results when encountered with unseen

data. In other words, the model prediction results would be uncertain for the groups that were less present in the training dataset. If the model is particularly more uncertain about the data from a specific demographic, the model's predictions are likely to be more unreliable and unpredictable for the group, resulting in unfair reliability compared to other groups. I believe it would be an interesting research direction to leverage the estimated uncertainty for evaluating and mitigating bias in ML, which provides another dimension in understanding ML bias.

6.2.4 Toward Building Reliable mHealth Application

This thesis focuses on user aspects of mHealth applications; when an algorithm works well, intelligent feedback can guide users to get the expected accurate output. In the mHealth application pipeline, both data acquisition and algorithm play a crucial role in providing reliable and accurate results. Data acquisition often involves users and the strategies proposed in this work would help users acquire high-quality data. However, when the criteria for high-quality data are too strict that the users could never satisfy it, it not only degrades the usability of the mHealth applications but also the robustness of the algorithm should be questioned. Additionally, when mHealth applications fail even with high-quality data, failure could be caused by the underlying algorithm. To address this scenario, I also explored providing the algorithm's estimated uncertainty to inform the users of any unexpected failures from the algorithms to protect users from trusting unreliable health predictions. However, this does not address the root cause of the failure of the underlying algorithm.

I argue that it is the developers' responsibility to monitor failures and improve the algorithm. Developers should employ a thorough failure analysis pipeline to monitor and address failures due to data quality and algorithmic limitations. For example, with an increasing failure rate in satisfying a specific data quality trait, the developers should consider modifying the algorithms to allow more lenient criteria for the trait. With an increasing failure rate with certain groups of users or certain regions, the algorithm could be biased and should be immediately addressed. A specific set of devices could show inferior performances,

which should be addressed with the aforementioned measures for device heterogeneity. The frequency of updates can be determined by the severity, importance, and complexity of addressing the failures.

The machine learning community has investigated a vast amount of efforts to address failures of ML algorithms. To build ML models that are inherently more robust, researchers and engineers often train generative models that learn an actual distribution of data rather than decision boundaries. To make the existing models adapt to new devices or groups of users, numerous domain adaption techniques (e.g., transfer learning, few-shot learning, meta-learning, federated learning) have been proposed. GAN-based synthetic data generation can be also leveraged to build more robust datasets.

As the robustness of the algorithm or the machine learning model improves, there is less burden on users to collect high-quality data. And, the quality assurance criteria can become more lenient for users, improving the usability of the mHealth applications. However, this does not mean that the feedback strategies become meaningless if robust models are deployed. The users would still need to be informed of what data the algorithm expects and how to acquire such data. Even extremely robust models would not have 100% accuracy and complete coverage of all possible data in the world; unexpected failures from the algorithms should be informed to the users. In the end, the feedback for the users should not just reside in between the users and mHealth applications. The failures from mHealth application interactions should be used as feedback to the developers to improve the algorithms. In return, the users will be benefited from easier and more accurate use of mHealth applications. Although this thesis mostly focuses on the users' interaction with mHealth applications, I believe efforts for both algorithm and usability aspects would truly achieve accurate, reliable, and trustworthy mHealth applications in the future.

Appendix A

APPENDIX

A.1 Appendix – RDTScan

A.1.1 Configurability Test with Other RDTs

To assess the extent to which RDTScan can be configured to new RDT designs, we examined whether RDTScan could reliably detect a range of RDT products not covered in our case studies. We generated the list of RDTs for this study by combining two lists curated by major health agencies—one by the World Health Organization for malaria RDTs [141] and one by the Centers for Disease Control for influenza RDTs [49]. After finding that the lists had very few dipstick RDTs (9 cassettes vs. 2 dipsticks), we added 3 additional dipsticks that are used for pregnancy testing. The results of this study are shown below. RDTScan was able to detect 8 of the 10 cassettes, all without the need for fiducial checking; both failure cases were blank white cassettes without any lettering or logos. RDTScan was able to detect all of the dipsticks provided that fiducial checking was enabled.

RDT Name	**Form Factor**	**Compatibility with RDTScan**
Binax Now Influenza A&B Card (Alere)	Card	Yes
QuickVue Influenza A+B (Quidel)	Dipstick	Yes (w/ fiducial detection)
LifeSign LLC Status Flu A&B (Princeton BioMeditech)	Cassette	No (blank cassette)

XPECT Flu A&B (Remel/Thermo Fisher)	Cassette	Yes
OSOM Ultra Plus Flu A&B Test (Sekisui Diagnostics)	Dipstick	Yes (w/ fiducial detection)
CareStart Flu A&B Plus (Access Bio)	Cassette	Yes
CareStart Malaria Pf/Pv (Access Bio)	Cassette	Yes
SD Bioline Malaria Ag Pf/Pv (Alere)	Cassette	Yes
PALUTOP +4 optima (All. Diag)	Cassette	Yes
One Step Malaria HRP2/pLDH (Wondfo)	Cassette	No (blank cassette)
AllTest Malaria P.f./Pan (AllTest)	Cassette	Yes
Asan Easy Test Malaria Pf/Pan Ag (ASAN Pharm.)	Cassette	Yes
Clearview hCG II Dipsticks (Clearview)	Dipstick	Yes (w/ fiducial detection)
Pregnancy Urine Dip-Strips (CLIAwaived, Inc)	Dipstick	Yes (w/ fiducial detection)
OSOM hCG DipStick Urine Test (Sekisui Diagnostics)	Dipstick	Yes (w/ fiducial detection)

A.2 Appendix – Reliable and Trustworthy ML for Health

A.2.1 Performance Metrics

In out-of-distribution performance evaluation in Section 5.4.2, we use the following metrics that has been used in previous out-of-distribution work [102, 169]:

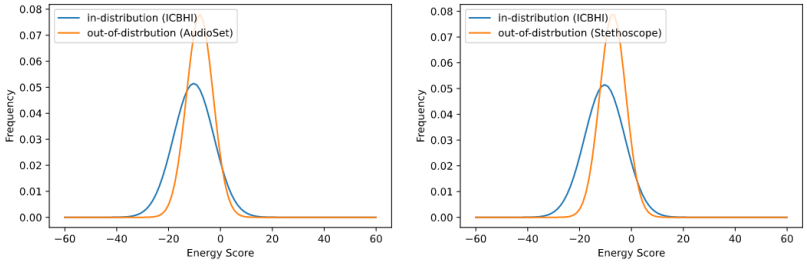
- **True negative rate (TNR) at 95% true positive rate** (TPR) is defined as the percentage of correctly detected out-of-distribution samples, when 95% of in-distribution samples are correctly detected. TNR is calculated $TNR = TN/(FP + TN)$ and $TPR = TP/(TP + TN)$, where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

- **Area under the receiver operating curve (AUROC)** is defined as the area under the plot of true positive rate (TPR) versue false positive rate (FPR), where $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$.

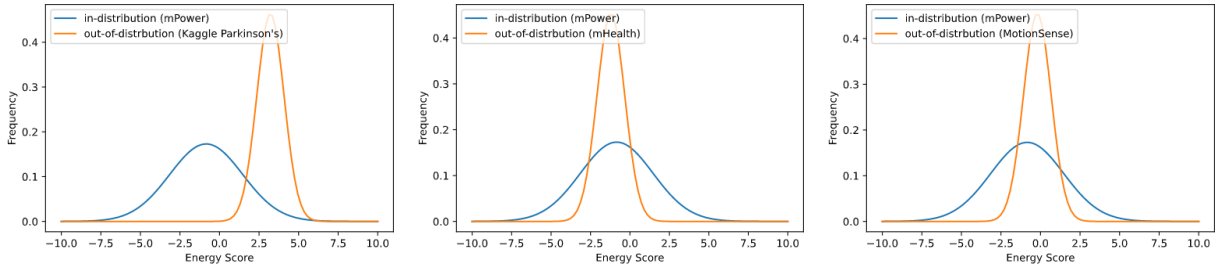
- **Detection accuracy** is defined as the maximum classification accuracy over all possible thresholds in classifying in- and out-of-distribution data.

A.2.2 Energy-Based OOD Detection Analysis

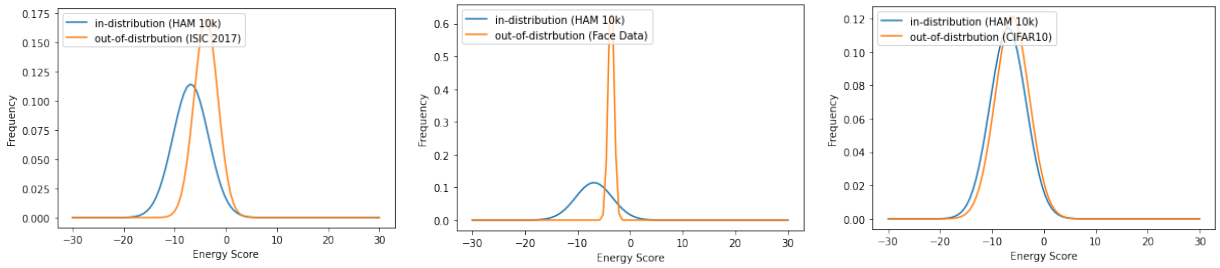
In Section 5.4.2, energy-based out-of-distribution detection method does not show comparable performance to methods using Mahalanobis distance and Gram matrices. We further analyze the method by comparing the distribution of energy score between in- and out-of-distribution as shown Figure A.1. In most cases, the distribution of the energy scores are overlapped, making it difficulty to detect out-of-distribution samples using energy score. In this work, we use energy-based method without fine-tuning, which is suitable for adopting the method to any pre-trained models. However, as the authors have demonstrated in their paper [107], fine-tuned energy-based method that requires re-training of a classifier, shows significant improvement in detecting out-of-distribution samples.



(a) TNR@TPR95: 8.12 (b) TNR@TPR95: 7.29



(c) TNR@TPR95: 70.00 (d) TNR@TPR95: 0.00 (e) TNR@TPR95: 0.00



(f) TNR@TPR95: 14.28 (g) TNR@TPR95: 0.00 (h) TNR@TPR95: 5.06

Figure A.1: Energy score distribution across different in- and out-of-distribution datasets.

A.2.3 Out-of-Distribution Performance with Confidence Interval

				Validation on OOD Samples (TNR @ TPR95/AUROC/Detection Accuracy)		
Health ML Models	**In-Distribution**	**Out-of-Distribution**	**Distribution Shift**	**Mahalanobis**	**Gram**	**Energy**
Skin Lesion	HAM10000	ISIC 2017	Covariate/label shift	10.13 / 58.21 / 59.28	25.90 / 81.14 / 74.98	14.28 / 76.20 / 70.76
(DenseNet-121)				±2.61/ ±3.30 / ±2.38	±1.22 / ±1.89 / ±1.12	±0.49 / ±0.18 / ±0.16
		Face	Open-set recognition	100.00 / 99.98 / 99.96	95.01 / 98.20 / 96.34	0.00 / 80.45 / 84.81
				±0.00 / ±0.02 / ±0.04	±1.48 / ±0.41 / ±0.63	±0.00 / ±0.14 / ±0.25
		CIFAR10	Open-set recognition	99.83 / 99.90 / 99.61	95.14 / 98.66 / 96.90	5.06 / 58.33 / 57.89
				±0.18 / ±0.10 / ±0.39	±1.43 / ±1.37 / ±1.94	±0.26 / ±0.92 / ±0.67
Lung Sound	ICBHI	AudioSet	Open-set recognition	97.96 / 99.47 / 97.34	96.55 / 99.18 / 95.97	8.12 / 56.79 / 57.13
(ResNet-34)				±0.73 / ±0.26 / ±0.45	±1.67 / ±0.30 / ±0.62	±0.24 / ±0.15 / ±0.14
		Stethoscope	Covariate/label shift	45.60 / 86.27 / 80.57	41.77 / 83.75 / 76.05	7.29 / 60.98 / 58.94
				±4.95 / ±1.42 / ±1.55	±1.62 / ±0.63 / ±0.38	±1.22 / ±0.74 / ±0.63
Parkinson's Disease	mPower	MotionSense	Open-set recognition	100.00 / 99.86 / 99.89	100.00 / 99.94 / 99.60	0.00 / 58.71 / 64.96
(5×1D-Conv)				±0.00 / ± 0.13 / ±0.10	±0.00 / ±0.24 / ±0.14	±0.00 / ±0.59 / ±0.32
		mHealth	Open-set recognition	100.00 / 100.00 / 100.00	100.0 / 99.99 / 99.99	0.00 / 41.41 / 59.44
				± 0.00 / ±0.00 / ±0.00	±0.00 / ±0.02 / ±0.01	±0.00 / ±1.09 / ±1.10
		Kaggle Parkinson's	Covariate/label shift	98.00 / 99.89 / 99.47	98.96 / 99.96 / 99.67	70.00 / 95.91 / 93.34
				±2.45 / ±0.14 / ±1.25	±0.00 / ±0.02 / ±0.03	±4.68 / ±0.30 / ±0.32

Table A.2: Out-of-Distribution Detection Performance Across Multiple Tasks. Evaluation is repeated for 5 times. Mean and standard deviation of metrics are reported.

A.2.4 User Study Interface

In this section, we provide screenshots and list of examples that were used in the user study.

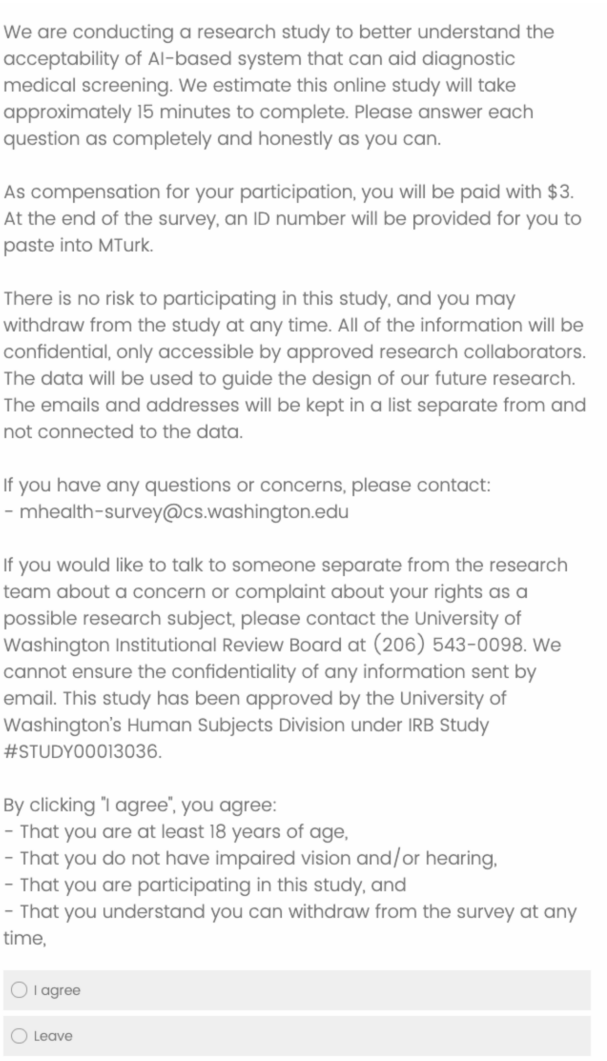


Figure A.2: User study consent form. Note that the name of the institution is redacted for the review.

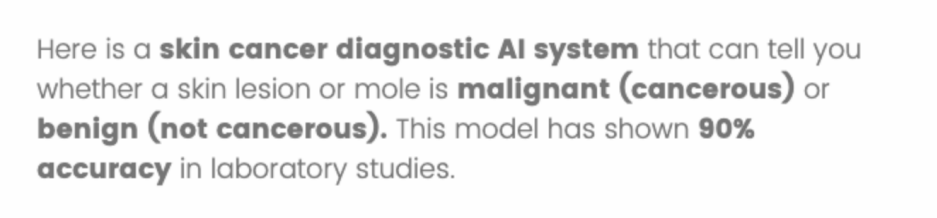


Figure A.3: Interface that shows information about a health machine learning model. It shows target health condition, possible prediction results, and its accuracy.

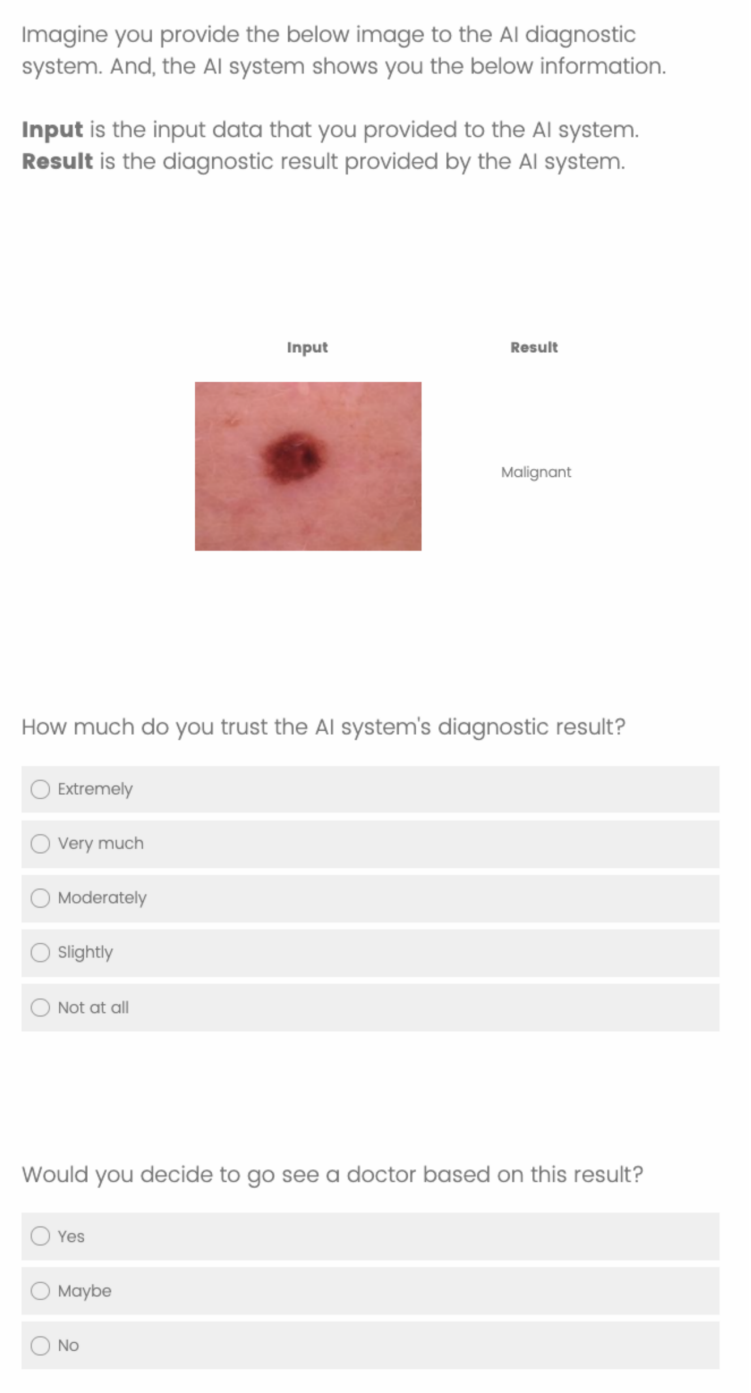


Figure A.4: Interface that shows baseline condition. This condition only presents input data and prediction results and asks questions on user-perceived trustworthiness and impact on making medical decisions.

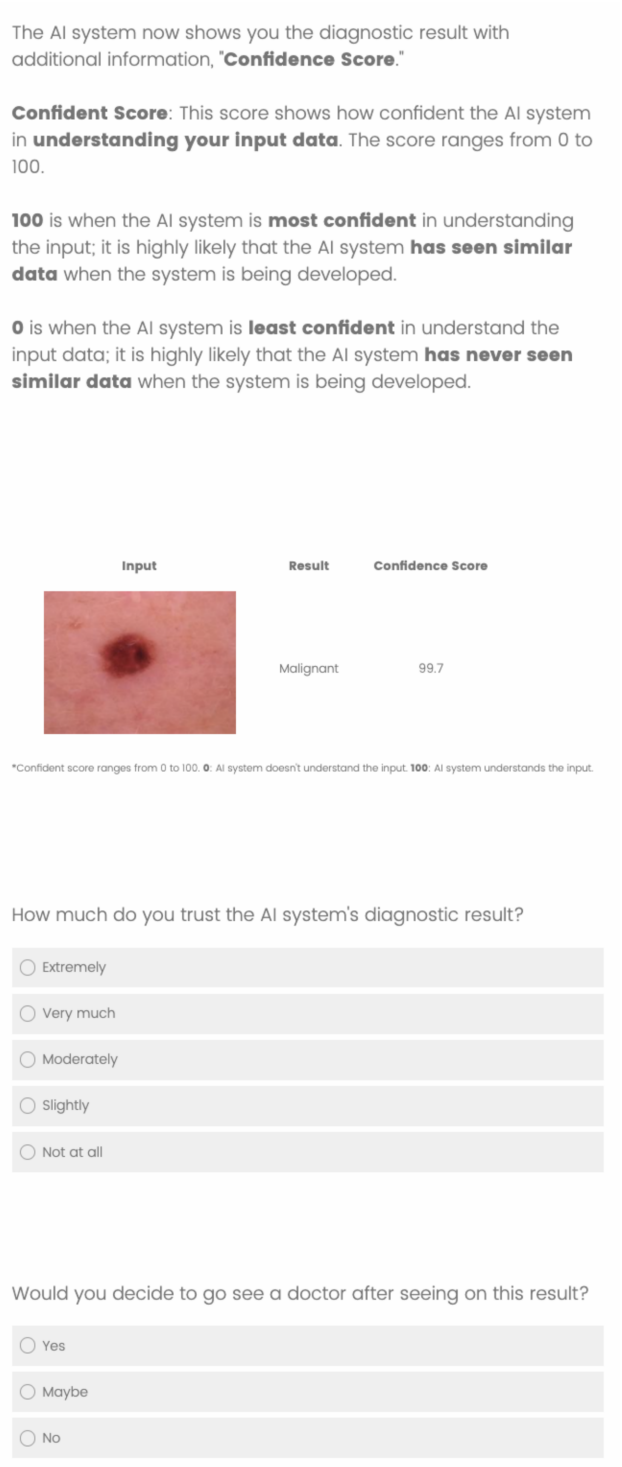


Figure A.5: Interface that shows CONFIDENCE SCORE condition. This condition only presents input data, prediction results, and CONFIDENCE SCORE.

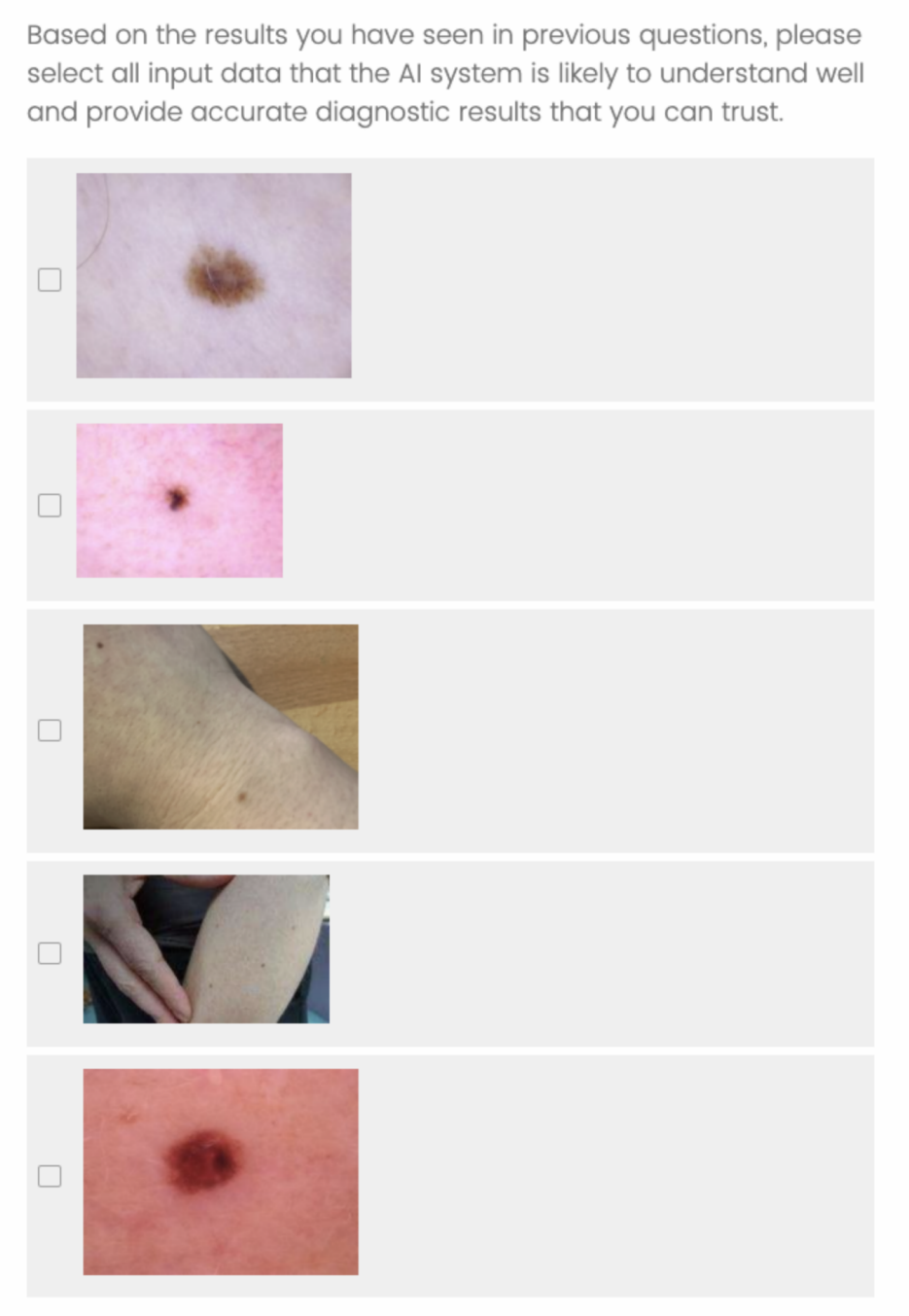
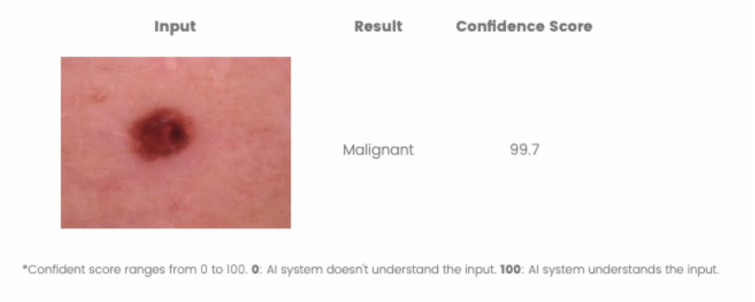
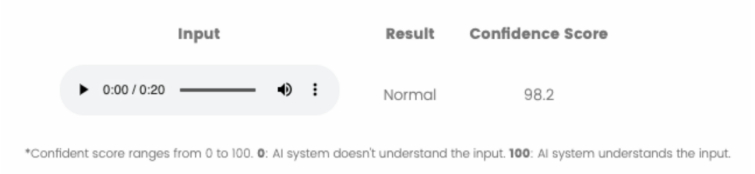


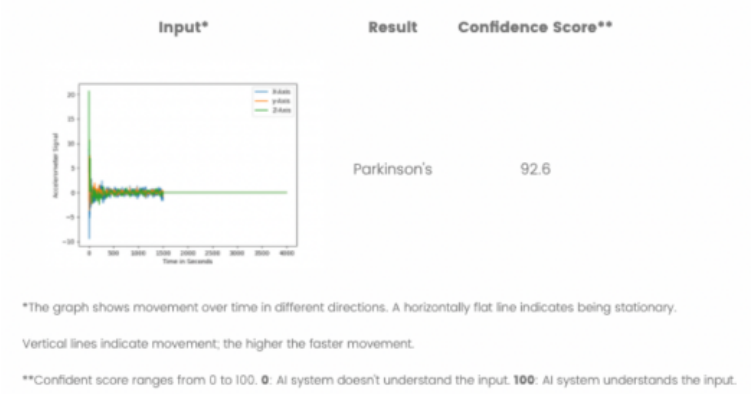
Figure A.6: Interface that asks users to select input data that would have high CONFIDENCE SCORE to explore potential learning effect through their interaction with CONFIDENCE SCORE.



(a) Image input is shown in a visible size.



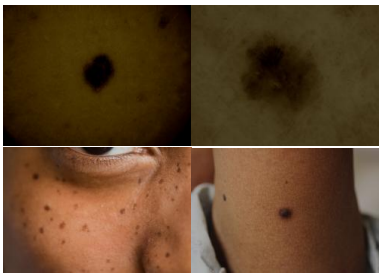
(b) Audio player is embedded for the participants to listen to the input data.



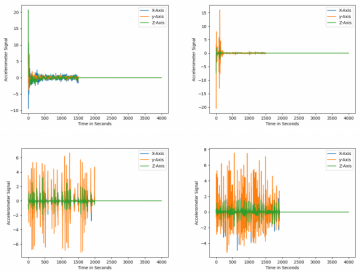
(c) Motion data is shown as a time-series plot of accelerometer signal. We provide additional explanation about how to interpret the signal.

Figure A.7: Interface to display different input data types.

(a) Input examples for skin cancer classifier for the participants who self-report to have light-colored skin tone.



(b) Input examples for skin cancer classifier for the participants who self-report to have dark-colored skin tone.



(c) Input examples for Parkinson's disease classifier.

Figure A.8: List of input examples used in the user study. For each input type, top row shows in-distribution inputs and bottom row shows out-of-distribution inputs. Left column shows positive diagnostic results and right column shows negative diagnostic results. Note that audio samples are not included due to its difficulty to visualize.

BIBLIOGRAPHY

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pages 9505–9515, 2018.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *36th International Conference on Machine Learning, ICML 2019*, pages 120–129, 2019.

[4] Konstantinos Agoritsas, Kathy Mack, Bema K Bonsu, Debbie Goodman, Douglas Salamon, and Mario J Marcon. Evaluation of the quidel quickvue test for detection of influenza a and b viruses in the pediatric emergency medicine setting by use of three specimen collection methods. *Journal of clinical microbiology*, 44(7):2638–2641, 2006.

[5] Saba Akbar, Enrico Coiera, and Farah Magrabi. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *Journal of the American Medical Informatics Association*, 27(2):330–340, 2020.

[6] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[7] Bronwen Anderson, Anne-Maree Kelly, Debra Kerr, and Damien Jolley. Capillary refill time in adults has poor inter-observer agreement. *hong kong Journal of Emergency Medicine*, 15(2):71–74, 2008.

[8] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[9] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[10] Chukwuemeka CA Azikiwe, C C Ifezulike, Iyeopu M Siminialayi, Louis U Amazu, J C Enye, and O E Nwakwunite. A comparative laboratory diagnosis of malaria: Microscopy versus rapid diagnostic test kits. *Asian Pacific Journal of Tropical Biomedicine*, 2(4):307–310, apr 2012.

[11] Sean Bae, Silviu Borac, Yunus Emre, Jonathan Wang, Jiang Wu, Mehr Kashyap, Si-Hyuck Kang, Liwen Chen, Melissa Moran, Julie Cannon, et al. Prospective validation of smartphone-based heart rate and respiratory rate measurement algorithms. *Communications medicine*, 2(1):1–10, 2022.

[12] BankMyCell. How Many Smartphones Are In The World?, 2020.

[13] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, pages 91–98. Springer, 2014.

[14] LARRY J BARAFF. Capillary refill: is it a useful clinical sign? *Pediatrics*, 92(5):723–724, 1993.

[15] Anabela Berenguer, Jorge Goncalves, Simo Hosio, Denzil Ferreira, Theodoros Anagnostopoulos, and Vassilis Kostakos. Are Smartphones Ubiquitous?: An in-depth survey of smartphone adoption by seniors. *IEEE Consumer Electronics Magazine*, 6(1):104–110, jan 2017.

[16] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586*, 2020.

[17] L. Bissonnette and M. G. Bergeron. Diagnosing infections–current and anticipated technologies for point-of-care diagnostics and home-based testing, 2010.

[18] LL Blaxter, David E Morris, John A Crowe, C Henry, S Hill, Don Sharkey, H Vyas, and Barrie R Hayes-Gill. An automated quasi-continuous capillary refill timing device. *Physiological measurement*, 37(1):83, 2015.

[19] Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Ray Dorsey, et al.

The mpower study, parkinson disease mobile data collected using researchkit. *Scientific data*, 3(1):1–9, 2016.

[20] Nam Bui, Anh Nguyen, Phuc Nguyen, Hoang Truong, Ashwin Ashok, Thang Dinh, Robin Deterding, and Tam Vu. Pho2: Smartphone based blood oxygen level measurement systems using near-ir and red wave-guided light. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, pages 1–14, 2017.

[21] Grace Burleson, Mustafa Naseem, and Kentaro Toyama. An exploration of african-american pregnant women's information-seeking behavior in detroit. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, pages 1–12, 2020.

[22] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[23] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[24] Erdi Çallı, Keelin Murphy, Ecem Sogancioglu, and Bram Van Ginneken. Frodo: Free rejection of out-of-distribution samples: application to chest x-ray analysis. *arXiv preprint arXiv:1907.01253*, 2019.

[25] Clara Calvert, Andrea Kolkenbeck-Ruh, Simone H Crouch, Larske M Soepnel, and Lisa J Ware. Reliability, usability and identified need for home-based cardiometabolic health self-assessment during the covid-19 pandemic in soweto, south africa. *Scientific reports*, 12(1):1–9, 2022.

[26] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[27] Scott Carter, John Adcock, John Doherty, and Stacy Branham. NudgeCam: Toward targeted, higher quality media capture. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 615–618, 2010.

[28] Centers for Disease Control and Prevention. Influenza Signs and Symptoms and the Role of Laboratory Diagnostics, 2016.

[29] Francine Chen, Scott Carter, Laurent Denoue, and Jayant Kumar. SmartDCap: Semi-automatic capture of higher quality document images from a smartphone. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 287–296, mar 2013.

[30] F Chiadmi, J Schlatter, B Mounkassa, P Ovetchkine, and N Vermerie. Fast diagnostic tests in the management of group A beta-heamolytic streptococcal pharyngitis. *Annales de biologie clinique*, 62(5):573–577, sep 2004.

[31] Claudiu Cobârzan, Marco A. Hudelist, Klaus Schoeffmann, and Manfred Jürgen Primus. Mobile image analysis: Android vs. ios. In Xiangjian He, Suhuai Luo, Dacheng Tao, Changsheng Xu, Jie Yang, and Muhammad Abul Hasan, editors, *MultiMedia Modeling*, pages 99–110, Cham, 2015. Springer International Publishing.

[32] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[33] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[34] BORIS AFONSO Cruz, ED Queiroz, SV Nunes, ACHILES Cruz Filho, GILBERTO BELISARIO Campos, EL Monteiro, and HUMBERTO Crivellari. Severe raynaud's phenomenon associated with interferon-beta therapy for multiple sclerosis: case report. *Arquivos de Neuro-psiquiatria*, 58(2B):556–559, 2000.

[35] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[36] Lilian De Greef, Mayank Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 331–342, 2014.

[37] Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 4 2021.

[38] Nicola Dell and Gaetano Borriello. Mobile tools for point-of-care diagnostics in the developing world. In *Proc. DEV '13*, pages 1–10, New York, New York, USA, 2013. ACM Press.

[39] Nicola Dell, Ian Francis, Haynes Sheppard, Raiva Simbi, and Gaetano Borriello. Field evaluation of a camera-based mobile health system in low-resource settings. In *Proc. MobileHCI '14*, pages 33–42, 2014.

[40] Denso ADC. QR Code Essentials. Technical report, 2011.

[41] Brian DeRenzi, Neal Lesh, Tapan Parikh, Clayton Sims, Werner Maokla, Mwajuma Chemba, Yuna Hamisi, David S hellenberg, Marc Mitchell, and Gaetano Borriello. E-imci: Improving pediatric health care in low-income countries. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 753–762, 2008.

[42] Michael Dixon, Logan Schneider, Jeffrey Yu, Jonathan Hsu, Anupam Pathak, D Shin, Reena Singhal Lee, Mark Rajan Malhotra, Ken Mixter, Mike McConnell, et al. Sleep-wake detection with a contactless, bedside radar sleep sensing system. 2021.

[43] Miro Dudík, William Chen, Solon Barocas, Mario Inchiosa, Nick Lewins, Miruna Oprescu, Joy Qiao, Mehrnoosh Sameki, Mario Schlener, Jason Tuo, and Hanna Wallach. Assessing and mitigating unfairness in credit models with the fairlearn toolkit. 2020.

[44] Ivan Evtimov, Weidong Cui, Ece Kamar, Emre Kiciman, Tadayoshi Kohno, and Jerry Li. Security and machine learning in the real world. *arXiv preprint arXiv:2007.07205*, 2020.

[45] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[46] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns:. *Big Data & Society*, 6(1):1–14, 2019.

[47] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[48] Dustin E Fleck and Mark F Hoeltzel. Hand and foot color change: diagnosis and management. *Pediatrics In Review*, 38(11):511–519, 2017.

[49] Centers for Disease Control and Prevention. Rapid influenza diagnostic tests (ridts).

[50] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

[51] Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021.

[52] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting, 2020.

[53] Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, Philip J Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, 293(10):1223–1238, 2005.

[54] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognition*, 84:345–356, 2018.

[55] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[56] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 2022.

[57] Philippe Gillet, Marcella Mori, Jef Van Den Ende, and Jan Jacobs. Buffer substitution in malaria rapid diagnostic tests causes false-positive results. *Malaria Journal*, 9(1):215, jul 2010.

[58] Giorgia. Simulation of parkinson movement disorders – kaggle, May 2018.

[59] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11, 2017.

[60] Mayank Goel, Jacob Wobbrock, and Shwetak Patel. Gripsense: using built-in sensors to detect hand posture and pressure on commodity mobile phones. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 545–554, 2012.

[61] Douglas Gollin and Christian Zimmermann. Malaria: Disease Impacts and Long-Run Income Differences. *Department of Economics Working Paper Series*, (2997):33, 2007.

[62] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[63] Marc H Gorelick, Kathy N Shaw, and M Douglas Baker. Effect of ambient temperature on capillary refill in healthy children. *Pediatrics*, 92(5):699–702, 1993.

[64] N. Gorski, V. Anisimov, E. Augustin, O. Baret, D. Price, and J. C. Simon. A2iA Check Reader: A family of bank check recognition systems. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pages 527–530. IEEE Computer Society, 1999.

[65] Nikolai Gorski, Valery Anisimov, Emmanuel Augustin, Olivier Baret, and Sergey Maximov. Industrial bank check processing: The A2iA CheckReader™. *International Journal on Document Analysis and Recognition*, 3(4):196–206, 2001.

[66] Robert A Greenes. *Clinical decision support: the road ahead*. Elsevier, 2011.

[67] Domenico Grimaldi, Yuriy Kurylyak, Francesco Lamonaca, and Alfonso Nastro. Photoplethysmography detection by smartphone's videocamera. In *Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, volume 1, pages 488–491. IEEE, 2011.

[68] GSMA. The Mobile Economy Sub-Saharan Africa. Technical report, 2020.

[69] Steven A Harvey, Larissa Jennings, Masela Chinyama, Fred Masaninga, Kurt Mulholland, and David R Bell. Improving community health worker use of malaria rapid diagnostic tests in Zambia: package instructions, job aid and job aid-plus-training. *Malaria Journal*, 7(1):160, dec 2008.

[70] Nigel Hawkes. Flu: Australia sees early start to season. *BMJ (Clinical research ed.)*, 366:l4603, jul 2019.

[71] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[72] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR (Poster)*, 2016.

[73] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[74] Sócrates Herrera, Andrés F Vallejo, Juan P Quintero, Myriam Arévalo-Herrera, Marcela Cancino, and Santiago Ferro. Field evaluation of an automated RDT reader and data management device for Plasmodium falciparum/Plasmodium vivax malaria in endemic areas of Colombia. *Malaria Journal*, 13(1):87, mar 2014.

[75] AL Herrick. Pathogenesis of raynaud's phenomenon. *Rheumatology*, 44(5):587–596, 2005.

[76] Jason S Hoffman, Varun Viswanath, Xinyi Ding, Matthew J Thompson, Eric C Larson, Shwetak N Patel, and Edward Wang. Smartphone camera oximetry in an induced hypoxemia study. *arXiv preprint arXiv:2104.00038*, 2021.

[77] Shichu Huang, Koji Abe, Steven Bennett, Tinny Liang, Paula D. Ladd, Lindsay Yokobe, Caitlin E. Anderson, Kamal Shah, Josh Bishop, Mike Purfield, Peter C. Kauffman, Sai Paul, AnneMarie E. Welch, Bonnie Strelitz, Kristin Follmer, Kelsey Pullar, Luis Sanchez-Erebia, Emily Gerth-Guyette, Gonzalo Domingo, Eileen Klein, Janet A. Englund, Elain Fu, and Paul Yager. Disposable Autonomous Device for Swab-to-Result Diagnosis of Influenza. *Analytical Chemistry*, 89(11):5776–5783, jun 2017.

[78] L. Huette, P. Barbosa-Pereira, O. Bougeois, J. V. Moreau, B. Plessis, P. Courtellemont, and Y. LeCourtier. Multi-Bank Check Recognition System: Consideration on the Numeral Amount Recognition Module. pages 133–156. dec 1997.

[79] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. Vitamon: Measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 1–14, 2019.

[80] Jan Jacobs, Barbara Barbé, Philippe Gillet, Michael Aidoo, Elisa Serra-Casas, Jan Van Erps, Joelle Daviaud, Sandra Incardona, Jane Cunningham, and Theodoor Visser. Harmonization of malaria rapid diagnostic tests: Best practices in labelling including instructions for use. *Malaria Journal*, 13(1):505, dec 2014.

[81] Nancy K Janz and Marshall H Becker. The health belief model: A decade later. *Health education quarterly*, 11(1):1–47, 1984.

[82] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. Supporting blind photography. In *ASSETS'11: Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 203–210, 2011.

[83] In Choel Jeong, Hyungro Yoon, Hyunjeong Kang, and Hojun Yeom. Effects of skin surface temperature on photoplethysmograph. *Journal of healthcare engineering*, 5(4):429–438, 2014.

[84] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, pages 5541–5552, 2018.

[85] Paul CD Johnson. Extension of nakagawa & schielzeth's r2glmm to random slopes models. *Methods in ecology and evolution*, 5(9):944–946, 2014.

[86] Alinune N Kabaghe, Benjamin J Visser, Rene Spijker, Kamija S Phiri, Martin P Grobusch, and Michèle Van Vugt. Health workers' compliance to rapid diagnostic tests (RDTs) to guide malaria treatment: A systematic review and meta-Analysis. *Malaria Journal*, 15(1):163, dec 2016.

[87] Eleni Kakalou, Vasileios Papastamopoulos, Panagiotis Ioannidis, Kostas Papanikolaou, Ourania Georgiou, and Athanasios Skoutelis. Early HIV diagnosis through use of rapid diagnosis test (RDT) in the community and direct link to HIV care: a pilot project for vulnerable populations in Athens, Greece. *Journal of the International AIDS Society*, 17:19619, nov 2014.

[88] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[89] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[90] A Reşit Kavsaoğlu, Kemal Polat, and Muthusamy Hariharan. Non-invasive prediction of hemoglobin level using machine learning techniques with the ppg signal's characteristics features. *Applied Soft Computing*, 37:983–991, 2015.

[91] Emmett Kerr, Sonya Coleman, Martin McGinnity, and Andrea Shepherd. Measurement of capillary refill time (crt) in healthy subjects using a robotic hand. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1291–1298, 2018.

[92] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics*, 6(2):e24, 2018.

[93] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[94] Nerida L Klupp and Anne-Maree Keenan. An evaluation of the reliability and validity of capillary refill time test. *The Foot*, 17(1):15–20, 2007.

[95] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[96] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[97] Setor Kunutsor, John Walley, Elly Katabira, Simon Muchuro, Hudson Balidawa, Elizabeth Namagala, and Eric Ikoona. Using mobile phones to improve clinic attendance amongst an antiretroviral treatment cohort in rural uganda: a cross-sectional and prospective study. *AIDS and behavior*, 14(6):1347–1352, 2010.

[98] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413, 2017.

[99] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[100] Eric C Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N Patel. Spirosmart: using a microphone to measure lung function on a mobile phone. In *Proceedings of the 2012 ACM Conference on ubiquitous computing*, pages 280–289, 2012.

[101] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[102] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach,

H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[103] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. Ieee, 2011.

[104] J Lewin and I Maconochie. Capillary refill time in adults. *Emergency Medicine Journal*, 25(6):325–326, 2008.

[105] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[106] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[107] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

[108] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[109] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. *MetaPhys: Few-Shot Adaptation for Non-Contact Physiological Measurement*, page 154–163. Association for Computing Machinery, New York, NY, USA, 2021.

[110] Giulio Lovisotto, Henry Turner, Simon Eberz, and Ivan Martinovic. Seeing red: Ppg biometrics using smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 818–819, 2020.

[111] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, nov 2004.

[112] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 4768–4777, 2017.

[113] Victoria Lyon, Monica Zigman Suchsland, Monique Chilver, Nigel Stocks, Barry Lutz, Philip Su, Shawna Cooper, Chunjong Park, Libby Rose Lavitt, Alex Mariakakis, et al. Diagnostic accuracy of an app-guided, self-administered test for influenza among individuals presenting to general practice with influenza-like illness: study protocol. *BMJ open*, 10(11):e036298, 2020.

[114] David Mabey, Rosanna W. Peeling, Andrew Ustianowski, and Mark D. Perkins. Diagnostics for the developing world, mar 2004.

[115] Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, IoTDI '19, pages 49–58, New York, NY, USA, 2019. ACM.

[116] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

[117] Christine Manyando, Eric M Njunju, Justin Chileshe, Seter Siziya, and Clive Shiff. Rapid diagnostic tests for malaria and health workers' adherence to test results at health facilities in Zambia. *Malaria Journal*, 13(1):166, may 2014.

[118] Alex Mariakakis, Megan A Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N Patel. Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–26, 2017.

[119] Alex Mariakakis, Jacob Baudin, Eric Whitmire, Vardhman Mehta, Megan A Banks, Anthony Law, Lynn Mcgrath, and Shwetak N Patel. Pupilscreen: using smartphones to assess traumatic brain injury. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–27, 2017.

[120] Alex Mariakakis, Edward Wang, Shwetak Patel, and Mayank Goel. Challenges in realizing smartphone-based health sensing. *IEEE Pervasive Computing*, 18(2):76–84, apr 2019.

[121] Gloria Martínez, Newton Howard, Derek Abbott, Kenneth Lim, Rabab Ward, and Mohamed Elgendi. Can photoplethysmography replace arterial blood pressure in the assessment of blood pressure? *Journal of clinical medicine*, 7(10):316, 2018.

[122] Ronald E. McGaughey, Steven M. Zeltmann, and Mark E. McMurtrey. Motivations and obstacles to smartphone use by the elderly: Developing a research framework. *International Journal of Electronic Finance*, 7(3-4):177–195, 2013.

[123] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, jun 1947.

[124] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706, 2019.

[125] B Middleton, DF Sittig, and A Wright. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearbook of medical informatics*, (Suppl 1):S103, 2016.

[126] Marc Mitchell, Neal Lesh, Hilarie Cranmer, Hamish Fraser, Irina Haivas, and Kate Wolf. Improving care–improving access: the use of electronic decision support with aids patients in south africa. *International Journal of Healthcare Technology and Management*, 10(3):156–168, 2009.

[127] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020.

[128] Pablo J Montoya, Sheila A Lukehart, Paula E Brentlinger, Ana J Blanco, Florencia Floriano, Josefa Sairosse, and Stephen Gloyd. Comparison of the diagnostic accuracy of a rapid immunochromatographic test and the rapid plasma reagin test for antenatal syphilis screening in Mozambique. *Bulletin of the World Health Organization*, 84(2):97–104, 2006.

[129] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021.

[130] Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE reviews in biomedical engineering*, 8:98–113, 2014.

[131] Onur Mudanyali, Stoyan Dimitrov, Uzair Sikora, Swati Padmanabhan, Isa Navruz, and Aydogan Ozcan. Integrated rapid-diagnostic-test reader platform on a cellphone. *Lab on a Chip*, 12(15):2678–2686, aug 2012.

[132] David Mukanga, James K. Tibenderana, Juliet Kiguli, George W. Pariyo, Peter Waiswa, Francis Bajunirwe, Brian Mutamba, Helen Counihan, Godfrey Ojiambo, and Karin Kallander. Community acceptability of use of rapid diagnostic tests for malaria by community health workers in Uganda. *Malaria Journal*, 9(1), 2010.

[133] Marshal M. Mweu, Juliana Wambua, Fixtan Njuga, Philip Bejon, and Daniel Mwanga. Bayesian evaluation of the performance of three diagnostic tests for Plasmodium falciparum infection in a low-transmission setting in Kilifi County, Kenya. *Wellcome Open Research*, 4:67, oct 2019.

[134] Shinichi Nakagawa, Paul CD Johnson, and Holger Schielzeth. The coefficient of determination r 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213, 2017.

[135] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142, 2013.

[136] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 5:5, 2019.

[137] Anthony T Newall, Paul A Scuffham, and Brent Hodgkinson. Economic Report into the Cost of Influenza to the Australian Health System Report to the Influenza Specialist Group Executive summary. Technical Report March, 2007.

[138] Amanda Nickel, Shen Jiang, Natalie Napolitano, Kota Saeki, Hideaki Hirahara, Vinay Nadkarni, and Akira Nishisaki. Impact of skin color on accuracy of capillary refill time measurement by pulse oximeter. *Circulation*, 138(Suppl_2):A276–A276, 2018.

[139] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2020.

[140] Niels Olsen and Steen Levin Nielsen. Prevalence of primary raynaud phenomena in young females. *Scandinavian journal of clinical and laboratory investigation*, 38(8):761–764, 1978.

[141] World Health Organization. Final product list – who malaria rdt product testing round 7.

[142] World Health Organization, World Health Organization. Department of Child, Adolescent Health, and UNICEF. *Management of the child with a serious infection or severe malnutrition: guidelines for care at the first-referral level in developing countries*. World Health Organization, 2000.

[143] Akaninyene Otu, Bassey Ebenso, Okey Okuzu, and Egbe Osifo-Dawodu. Using a mhealth tutorial application to change knowledge and attitude of frontline health workers to ebola virus disease in nigeria: a before-and-after study. *Human Resources for Health*, 14(1):1–9, 2016.

[144] Aydogan Ozcan. Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools. *Lab on a Chip*, 14(17):3187–3194, 2014.

[145] Haydar Ozkan and Osman Semih Kayhan. A Novel Automatic Rapid Diagnostic Test Reader Platform. *Computational and Mathematical Methods in Medicine*, 2016:1–10, apr 2016.

[146] Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020.

[147] Nitika Pant Pai, Caroline Vadnais, Claudia Denkinger, Nora Engel, and Madhukar Pai. Point-of-Care Testing for Infectious Diseases: Diversity, Complexity, and Barriers in Low- And Middle-Income Countries. *PLoS Medicine*, 9(9), sep 2012.

[148] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

[149] Chunjong Park, Alex Mariakakis, Jane Yang, Diego Lassala, Yasamba Djiguiba, Youssouf Keita, Hawa Diarra, Beatrice Wasunna, Fatou Fall, Marème Soda Gaye, et al. Supporting smartphone-based image capture of rapid diagnostic tests in low-resource settings. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, pages 1–11, 2020.

[150] Chunjong Park, Hung Ngo, Libby Rose Lavitt, Vincent Karuri, Shiven Bhatt, Peter Lubell-Doughtie, Anuraj H Shankar, Leonard Ndwiga, Victor Osoti, Juliana K Wambua, et al. The design and evaluation of a mobile system for rapid diagnostic test interpretation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–26, 2021.

[151] J L Pech-Pacheco, G Crist, J Chamorro-Mart Nez, and & J Fern Andez-Valdivia. Diatom autofocusing in brightteld microscopy: a comparative study. Technical report.

[152] Trevor Perrier, Nicola Dell, Brian DeRenzi, Richard Anderson, John Kinuthia, Jennifer Unger, and Grace John-Stewart. Engaging pregnant women in kenya with a hybrid

computer-human sms communication system. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1429–1438, 2015.

[153] Amelia Pickard, Walter Karlen, and J Mark Ansermino. Capillary refill time: is it still a useful clinical sign? *Anesthesia & Analgesia*, 113(1):120–123, 2011.

[154] Stephen M Pizer, E. Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[155] Janet Elizabeth Pope. Primary raynaud phenomenon. *American Family Physician*, 90(6):403–404, 2014.

[156] Igor M. Quintanilha, Roberto de M. E. Filho, José Lezama, Mauricio Delbracio, and Leonardo O. Nunes. Detecting out-of-distribution samples using low-order deep features statistics. 2018.

[157] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[158] N Venkata Raju, M Jeffrey Maisels, Elizabeth Kring, and Laura Schwarz-Warner. Capillary refill time in the hands and feet of normal newborn infants. *Clinical pediatrics*, 38(3):139–144, 1999.

[159] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

[160] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001, 2019.

[161] R. Rosenthal, H. Rosenthal, and inc Sage Publications. *Meta-Analytic Procedures for Social Research*. Applied Social Research Methods. SAGE Publications, 1991.

[162] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *arXiv preprint arXiv:2104.03829*, 2021.

[163] Esmée Ruizendaal, Susan Dierickx, Koen Peeters Grietens, Henk DFH Schallig, Franco Pagnoni, and Petra F Mens. Success or failure of critical steps in community case management of malaria with rapid diagnostic tests: A systematic review, jun 2014.

[164] S. W. Ryu, J. H. Lee, J. Kim, M. A. Jang, J. H. Nam, M. S. Byoun, and C. S. Lim. Comparison of two new generation influenza rapid diagnostic tests with instrument-based digital readout systems for influenza virus detection. *British Journal of Biomedical Science*, 73(3):115–120, sep 2016.

[165] Jose M Saavedra, Glenn D Harris, Song Li, and Laurence Finberg. Capillary refilling (skin turgor) in the assessment of dehydration. *American journal of diseases of children*, 145(3):296–298, 1991.

[166] Elliot Saba. *Techniques for Cough Sound Analysis*. PhD thesis, 2018.

[167] Beatriz Garcia Santa Cruz, Matías Nicolás Bossa, Jan Sölter, and Andreas Dominik Husch. Public covid-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem. *Medical image analysis*, 74:102225, 2021.

[168] Devesh Sarda, Chunjong Park, Hung Ngo, Shwetak Patel, and Alex Mariakakis. Rdtcheck: A smartphone app for monitoring rapid diagnostic test administration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[169] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 8491–8501, 2020.

[170] Patrick Schoettker, Jean Degott, Gregory Hofmann, Martin Proença, Guillaume Bonnier, Alia Lemkaddem, Mathieu Lemay, Raoul Schorer, Urvan Christen, Jean-François Knebel, et al. Blood pressure measurements with the optibp smartphone app validated against reference auscultatory measurements. *Scientific Reports*, 10(1):1–12, 2020.

[171] David L Schriger and Larry Baraff. Defining normal capillary refill: variation with age, sex, and temperature. *Annals of emergency medicine*, 17(9):932–935, 1988.

[172] Osama M.E. Seidahmed, Muneir M.N. Mohamedein, Afrah A. Elsir, Fayez T. Ali, El Fatih M. Malik, and Eldirdieri S. Ahmed. End-user errors in applying two malaria rapid diagnostic tests in a remote area of Sudan. *Tropical Medicine & International Health*, 13(3):406–409, feb 2008.

[173] Itai Shavit, Rollin Brant, Cheri Nijssen-Jordan, Roger Galbraith, and David W Johnson. A novel imaging technique to measure capillary-refill time: improving diagnostic accuracy for dehydration in young children with gastroenteritis. *Pediatrics*, 118(6):2402–2408, 2006.

[174] Steven K. Shevell. 4 - color appearance. In Steven K. Shevell, editor, *The Science of Color (Second Edition)*, pages 149–190. Elsevier Science Ltd, Amsterdam, second edition edition, 2003.

[175] Koichiro Shinozaki, Kota Saeki, Lee S Jacobson, Julianne M Falotico, Timmy Li, Hideaki Hirahara, Katsuyuki Horie, Naoki Kobayashi, Steve Weisner, Joshua W Lampe, et al. Evaluation of accuracy of capillary refill index with pneumatic fingertip compression. *Journal of Clinical Monitoring and Computing*, 35(1):135–145, 2021.

[176] Solveig K Sieberts, Jennifer Schaff, Marlena Duda, Bálint Ármin Pataki, Ming Sun, Phil Snyder, Jean-Francois Daneault, Federico Parisi, Gianluca Costante, Udi Rubin, et al. Crowdsourcing digital health measures to predict parkinson's disease severity: the parkinson's disease digital biomarker dream challenge. *NPJ digital medicine*, 4(1):1–12, 2021.

[177] Shri Prakash Singh, Siddhivinayak Hirve, M. Mamun Huda, Megha Raj Banjara, Narendra Kumar, Dinesh Mondal, Shyam Sundar, Pradeep Das, Chitra Kumar Gurung, Suman Rijal, C. P. Thakur, Beena Varghese, and Axel Kroeger. Options for active case detection of visceral leishmaniasis in endemic districts of India, Nepal and Bangladesh, comparing yield, feasibility and costs. *PLoS Neglected Tropical Diseases*, 5(2), 2011.

[178] Joe Steinman, Andrew Barszczyk, Hong-Shuo Sun, Kang Lee, and Zhong-Ping Feng. Smartphones and video cameras: Future methods for blood pressure measurement. *Frontiers in Digital Health*, 3, 2021.

[179] Victor J Strecher and Irwin M Rosenstock. The health belief model. *Cambridge handbook of psychology, health and medicine*, 113:117, 1997.

[180] Krzysztof S Strozik, Clarissa H Pieper, and Jacques Roller. Capillary refilling time in newborn babies: normal values. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 76(3):F193–F196, 1997.

[181] KS Strozik, CH Pieper, and F Cools1. Capillary refilling time in newborns–optimal pressing time, sites of testing and normal values. *Acta paediatrica*, 87(3):310–312, 1998.

[182] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1):1–26, 2020.

[183] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[184] Rene Ning Teh, Irene Ule Ngole Sumbele, Gillian Asoba Nkeudem, Derick Ndelle Meduke, Samuel Takang Ojong, and Helen Kuokuo Kimbi. Concurrence of carestart™ malaria hrp2 rdt with microscopy in population screening for plasmodium falciparum infection in the mount cameroon area: predictors for rdt positivity. *Tropical medicine and health*, 47(1):17, 2019.

[185] Jayaraman J Thiagarajan, Prasanna Sattigeri, Deepta Rajan, and Bindya Venkatesh. Calibrating healthcare ai: Towards reliable and interpretable deep predictive models. *arXiv preprint arXiv:2004.14480*, 2020.

[186] Kentaro Toyama. *Geek heresy: Rescuing social change from the cult of technology*. 2015.

[187] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[188] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[189] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[190] Yu-Chih Tung and Kang G Shin. Expansion of human-phone interface by sensing structure-borne sound propagation. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 277–289, 2016.

[191] Gonzalo M Vazquez-Prokopec, Donal Bisanzio, Steven T Stoddard, Valerie Paz-Soldan, Amy C Morrison, John P Elder, Jhon Ramirez-Paredes, Eric S Halsey, Tadeusz J Kochel, Thomas W Scott, et al. Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one*, 8(4):e58802, 2013.

[192] Abinav Ravi Venkatakrishnan, Seong Tae Kim, Rami Eisawy, Franz Pfister, and Nassir Navab. Self-supervised out-of-distribution detection in brain ct scans. *arXiv preprint arXiv:2011.05428*, 2020.

[193] Piia Von Lode. Point-of-care immunotesting: Approaching the analytical performance of central laboratory methods, jul 2005.

[194] Edward Jay Wang, William Li, Doug Hawkins, Terry Gernsheimer, Colette Norby-Slycord, and Shwetak N Patel. Hemaapp: noninvasive blood screening of hemoglobin using smartphone cameras. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 593–604, 2016.

[195] Edward Jay Wang, Junyi Zhu, Mohit Jain, Tien-Jui Lee, Elliot Saba, Lama Nachman, and Shwetak N Patel. Seismo: Blood pressure monitoring using built-in smartphone accelerometer and camera. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2018.

[196] Siriorn Watcharananan, Sasisopin Kiertiburanakul, and Wasun Chantratita. Rapid influenza diagnostic test during the outbreak of the novel influenza A/H1N1 2009 in Thailand: An Experience with Better Test Performance in Resource Limited Setting. *Journal of Infection*, 60(1):86–87, jan 2010.

[197] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[198] Samuel White, Hanjie Ji, and Jeffrey P. Bigham. EasySnap: Real-time audio feedback for blind photography. In *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology, Adjunct Proceedings*, pages 409–410, 2010.

[199] Matt Whitehill, Jake Garrison, and Shwetak Patel. Whosecough: In-the-wild cougher verification using multitask learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 896–900. IEEE, 2020.

[200] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[201] Gaby Anne Wildenbos, Monique WM Jaspers, Marlies P Schijven, and LW Dusseljee-Peute. Mobile health for older adult patients: Using an aging barriers framework to classify usability problems. *International journal of medical informatics*, 124:68–77, 2019.

[202] Rapeeporn Wongkanya, Tippawan Pankam, Shauna Wolf, Supanit Pattanachaiwit, Jureeporn Jantarapakde, Supabhorn Pengnongyang, Prasopsuk Thapwong, Apichat

Udomjirasirichot, Yutthana Churattanakraisri, Nanthika Prawepray, Apiluk Paksornsit, Thidadaow Sitthipau, Sarayut Petchaithong, Raruay Jitsakulchaidejt, Somboon Nookhai, Cheewanan Lertpiriyasuwat, Sumet Ongwandee, Praphan Phanuphak, and Nittaya Phanuphak. HIV rapid diagnostic testing by lay providers in a key population-led health service programme in Thailand. *Journal of virus eradication*, 4(1):12–15, jan 2018.

[203] World Health Organization. World Malaria Report. Technical report, 2019.

[204] Liangwen Yan, Sijung Hu, Abdullah Alzahrani, Samah Alharbi, and Panagiotis Blanos. A multi-wavelength opto-electronic patch sensor to effectively detect physiological changes against human skin types. *Biosensors*, 7(2):22, 2017.

[205] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4477–4488, 2016.

[206] Hassan Zaraket and Reiko Saito. Japanese Surveillance Systems and Treatment for Influenza. *Current Treatment Options in Infectious Diseases*, 8(4):311–328, dec 2016.

[207] Hanrui Zhang, Kaiwen Deng, Hongyang Li, Roger L Albin, and Yuanfang Guan. Deep learning identifies digital biomarkers for self-reported parkinson's disease. *Patterns*, 1(3):100042, 2020.

[208] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020.

[209] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.