# Developing Novel Disease Screening Tools by Measuring Critical Biochemical and Physiological Signals

Manuja Sharma

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

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Shwetak N Patel, Chair

Eric J. Seibel, Chair

Payman Arabshahi

Program Authorized to Offer Degree:

Department of Electrical and Computer Engineering

University of Washington

## Abstract

Developing Novel Disease Screening Tools by Measuring Critical Biochemical and Physiological Signals

Manuja Sharma

Co-Chairs of the Supervisory Committee:
Shwetak N Patel
Department of Electrical and Computer Engineering

Eric J. Seibel
Department of Electrical and Computer Engineering

Health screening tools that are reliable and accurate have successfully reduced the burden on expensive diagnostic testing along with providing the opportunity of quick, lower cost treatments to a larger community. Designing these tools require interdisciplinary efforts, and in this thesis, author explores development of screening tools for two different bacterial diseases in collaboration with clinicians, researchers, and engineers. Oral Caries and Pulmonary Tuberculosis (TB), bacterial diseases affecting millions across the world, face unique challenges in early disease screening. In the case of oral caries, quantitative screening tools are very limited while in TB screening, the lack of low-cost screening device leaves millions of TB cases undetected. This thesis presents novel sensing techniques based on measuring critical biomarkers, identified with the help of clinicians, that can provide rapid, accurate, quantitative, and non-invasive feedback. Particularly, to reduce gaps in dentistry, thesis focuses on developing tools to measure acidity of oral biofilm. This acidity plays a vital role in the formation of oral caries, yet tools to measure it clinically are absent. O-pH, an optical pH monitoring device is presented that measures the acidity of dental plaque using non-invasive, opto-electronic sensors and provides quantitative feedback of oral health to dentists. The results from *in vitro* validation and *in vivo* study with 30 subjects indicate that the device

is reliable and accurate. The spot based device is extended to mapping pH using images and a proof-of-concept device built with multimodal-scanning fiber (mm-SFE) is presented. This technology can enable trend based oral health monitoring using pH. In the next part of the thesis, the similar design process is applied to develop a screening tool for TB that focuses on cough as a biomarker. The production of cough in TB patients are primarily in response to changes in lung tissues, which the thesis hypothesizes can result in differences in cough sounds compared to other respiratory health issues. To validate, the author presents, TBscreen, a ResNet-18 model, based on scalogram images of passive coughs. TBscreen was trained and validated using a controlled dataset of passive cough sounds of subjects with TB and other respiratory health ailments. The analysis indicates the sounds from coughing contain disease biomarkers that increases with bacterial load and has the potential to enable large scale screening. Further, results indicate that passive or natural coughs differ from forced coughs, coughs produced on an external prompt, making it difficult to use forced coughs as a proxy for passive coughs. Through development of these two new sensing tools into human subject studies, the author argues that significant contributions have been made to the engineering and medical communities at large.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# GLOSSARY

CARIES: Dental cavity / Tooth decay

FDA: Food and Drug Administration

PLAQUE: Oral biofilm

PH: Potential of Hydrogen to measure a medium's acidic or basic content

FLUORESCENCE: Fluorescence is the emission of light by a substance that has absorbed light or other electromagnetic radiation

RDT: Rapid Diagnostic Test

PLHIV: People living with HIV

TUBERCULOSIS: Disease caused by inhalation of *Mycobacterium tuberculosis*

PULMONARY TUBERCULOSIS: Tuberculosis that affects lungs

ACCURACY: $\frac{Numberofcorrectpredictions}{Totalnumberofpredictions}$

SENSITIVITY: $\frac{TruePositive}{TruePositive+FalseNegative}$

SPECIFICITY: $\frac{TrueNegative}{TrueNegative+FalsePositive}$

# ACKNOWLEDGMENTS

"*Our greatest glory is not in never falling, but in rising every time we fall.*" - Confucius

Thank you everyone who have helped me rise multiple times in my PhD journey. It was indeed an eventful ride with great collaborators, advisors, numerous high-risk projects, and not to mention the two years of pandemic in between. I am glad to have made it to the other side of the PhD and have a long list of people to thank.

This wouldn't have been possible without my advisors, Eric and Shwetak, and their constant support in helping me steer through various projects. Eric, thank you for helping me navigate through my first clinical study and always being there for brainstorming sessions. Appreciate you being the enthusiastic study participant and volunteering to skip brushing teeth multiple times for science. Shwetak, I am especially grateful for the freedom, trust, and opportunity that you provided to attempt high risk projects. I would also like to thank my undergraduate advisor, Dr. Hung Cao, on introducing me to research on health sensing and helping me decide to pursue PhD. I would especially like to thank Dr. Leonard Nelson for his advising on several dental projects, and teaching thoroughness in literature review.

I appreciate the support of my lab mates at Ubicomp lab and Human Photonics lab (HPL) in brainstorming innumerable issues that research projects, and graduate school throws at you. Edward, thanks for walking me through a project with you and setting an example of how to efficiently execute research ideas. Chunjong Park, we joined the lab together and have shared similar struggles, thanks for listening to my complaints and helping me network for internships. Yaxuan, thanks for on boarding me to the dental projects in HPL and warning me about the possible pitfalls. Matthew Carson, along with our birthdays we have shared

multiple projects around optical pH screening, and I am grateful for your mentorship and guidance.

I am grateful to have had the opportunity to interact across multiple departments at the university. My projects would have been incomplete without the support of dentists from UW School of Dentistry - Dr. Sadr and Dr. Xu, thank you! Understanding the Tuberculosis landscape wouldn't be possible without the support of Dr. Horne and Dr. Hawn from UW Medicine. Dr. Hawn thanks for putting in time to reshape the TBscreen manuscript. The great team at Centre for Respiratory Diseases Research (CRDR), Kenya made it possible to collect the Nairobi dataset on time, I really do appreciate the help. I would also like to acknowledge the various funding agencies that supported my research - National Science Foundation (NSF PFI 1631146 (PI Seibel)), National Institute of Allergy and Infectious Diseases (NIAID)'s RO1 grant 1R01AI150815 (DH, VN, TRH), D43 TW011817 (TRH, VN, DH), and K24AI137310 (TRH), the National Center For Advancing Translational Sciences of the NIH (UL1 TR002319).

To my physicians at UW, UCLA, and back home in India, I would like to extend my thanks, they have been supportive through my precarious health conditions and have always tried to accommodate my professional ambitions along with any treatment plan. It's through discussions with them and experience of the healthcare system that has kept me motivated to work in the field of health screening and diagnostics.

Saving the best for the last - I would like to thank my family and friends like family who have constantly supported me not just in PhD but in life! Debasis, thanks for being there through it all- the ups, and the downs, and making sure that we enjoy all the way through.

Thank you all!

Manuja Sharma

# DEDICATION

to my parents, Ajaymala Sharma and Ajit Kumar Sharma

# Chapter 1
# INTRODUCTION

Screening tools are characterized as lower cost alternatives to expensive diagnostic devices, for large-scale screening, at a lower implementation overhead. These tools generally provide results with lower accuracy than gold standards (diagnostic tools) but should be sufficiently reliable and accurate, since high proportions of false negatives or false positives might represent worse health outcomes and unnecessary diagnostic costs [162, 116, 154]. Convenient accessibility to screening tools enable early disease detection, in return, promoting quicker and at times cheaper medical interventions. Success of screening tools can be especially seen in early detection of cancer - mammography for breast cancer screening, Human papillomavirus (HPV) tests and Pap tests for cervical cancer screening have successfully reduced burden of late-stage cancer diagnosis in women [14, 13]. Until recently, screening tools were assumed to be useful for low-resource settings, but the COVID-19 pandemic has changed the narrative. Prior development of lateral assay based rapid diagnostic tests for influenza and malaria [114, 113] enabled faster translation of at-home COVID-19 antigen tests from research to commercial products. In turn reducing the burden of testing from hospitals across the world.

Looking from a practical and economical point of view, it is crucial to identify diseases that would greatly benefit from a screening tool, to ensure its successful adaptation and prevent fatigue of health workers/patients from over-testing. Diseases that impact a large population, health conditions where early interventions can be game changers in saving lives or improving overall well-being are generally prioritized [94]. In this thesis, author presents, design and development of screening tools for two bacterial diseases. The first tool, O-pH, focuses on oral health which if left unchecked can lead to dental caries, the most prevalent

health condition affecting billions globally [66, 1]. Early screening tools in dentistry can not only prevent loss of tooth but, at times, effectively reverse the damage while significantly reducing the high cost of dental care. The second tool presented, TBscreen, focuses on screening of Pulmonary Tuberculosis (TB), the second leading infectious disease-related cause of death after COVID-19.[24]. Low-cost, reliable, easy to use screening tools can assist in reaching TB patients in remote locations bringing us closer to World Health Organization's (WHO) goal of eliminating TB.

Development of screening tools is supported by the recent advancement in digital health technologies (DTH), a system that uses computing platforms, connectivity, software, and sensors for healthcare and related uses [157, 50]. DTH has enabled digitization of several 'subjective' symptoms, for example, cough counts, step counts, amongst others, helping build a new set of biomarkers - digital biomarkers. To develop screening tools, it is vital to look for biomarkers that is critical in disease progression or transmission and can be measured conveniently. At its core, this is interdisciplinary research requiring teams of clinicians, researchers, engineers to work in tandem to develop the research question and its solution. Fig.1.1 represents the overall process used in designing the two screening tools presented. Present gaps in disease screening were identified along with the critical biomarker whose measurement can help reduce the gaps. This led to the formation of the research question with contributions towards disease screening and engineering. To validate the research question, sensing system was designed and evaluated. Interdisciplinary research enabled us to receive feedback from clinicians and health researchers, helping us to adjust the tool and understand its clinical limitations for further research. In this thesis, screening tools are designed for two bacterial diseases using the described process (Fig.1.1). The requirements of the screening tools are vastly different, giving the opportunity to explore two different health fields by translating engineering skill set.

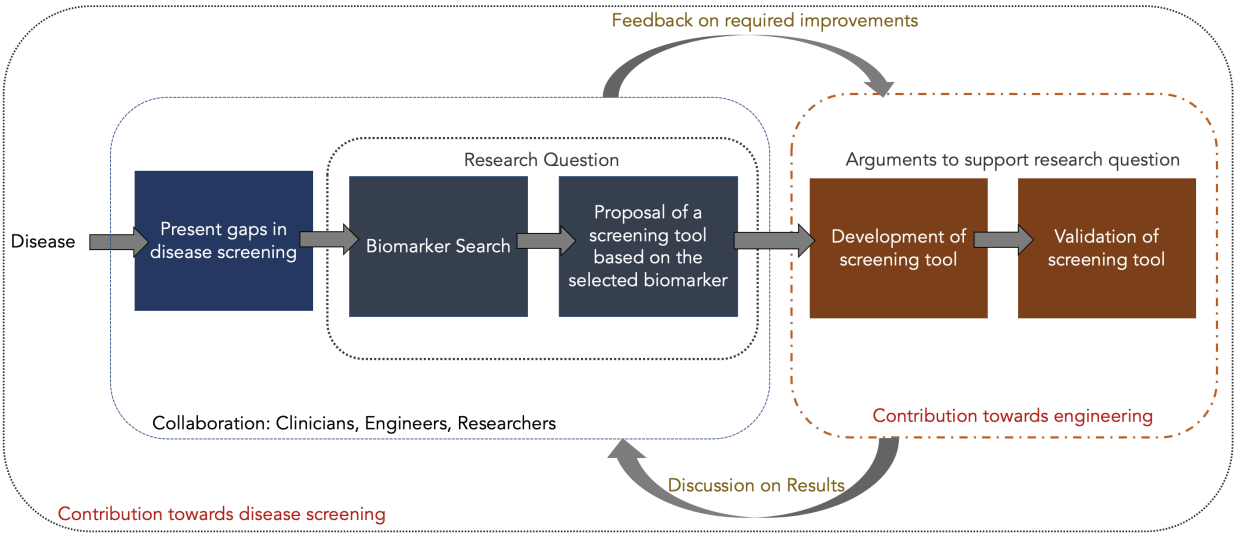


Figure 1.1: Design process for development of screening tools in the thesis

## *1.1 Thesis Statement*

In this thesis, author is looking at reducing gaps in the field of oral health and Tuberculosis screening, by proposing novel sensing techniques to measure critical biomarkers (bio-chemical and physiological disease manifestations). The goal is to provide rapid, quantitative, and accurate feedback to clinicians. Throughout this dissertation, author provides evidence to support the following thesis statement:

**Developing novel disease screening tools by measuring bio-chemical and physiological biomarkers**

The research contributions are divided into two categories based on the two diseases : Oral health screening and Tuberculosis screening, and Fig. 1.1 summarizes the overall development process. These two main categories are then further divided into specific research contributions as summarized below:

1. **Oral Health Screening:** Design of a pH based screening tool to assist in early screening for dental caries by measuring plaque pH, biochemical activity of oral bacteria.

(a) O-pH, an optical pH sensing system, based on fluorescence properties of FDA approved chemical dye, Sodium Fluorescein, is prototyped and validated to measures pH in the range of 4-7.5. *In vitro* analysis shows a mean error of 0.22 pH and standard deviation of 0.16 pH.

(b) In a clinical study with thirty participants, O-pH, indicated that acidity levels significantly differ between adolescent subjects with different cleaning history. In the Pre-Cleaning group, which consisted of a typical patient at a dentist's clinic for a routine re-care visit with no recent professional cleaning, pH measurements for unhealthy surfaces (having lesions, caries) were expectantly lower than the healthy surfaces, though results were not significant.

(c) Based on clinicians' feedback, spot based O-pH system was extended to pH imaging using multimodal-Scanning Fiber endoscope (mm-SFE). The case study showcased its capability in reproducing Stephan curve and a proof of concept design of image based pH sensing of oral plaque.

2. **Pulmonary Tuberculosis Screening:** Design of a passive cough-based tool for TB screening in a population of patients experiencing natural coughs due to TB (physiological biomarker) or non-TB related respiratory ailments.

(a) A unique dataset is collected in Nairobi, Kenya having high TB incidents, consisting of passive (43,200; n= 149) and forced coughs (1,619, n=42) in a controlled setting using three different recording instruments.

(b) In Nairobi dataset, TB subjects did not have significantly higher cough counts than non-TB subjects, passive cough spectral features distinguished these two groups with a five-fold cross validation sensitivity of 0.70 (+0.11 standard deviation across folds) and specificity of 0.71+0.10. In comparison to boundary and condenser microphones, the model trained on smartphones performed best with ROC-AUC (receiver operating characteristic – area under the curve) score of 0.85

(95% C.I. 0.84-0.85, p<0.001) and had better performance in subjects with higher bacterial load 0.87 (95% C.I. 0.87-0.88, p<0.001) or lung cavities 0.89 (95% C.I. 0.88-0.89, p<0.001). Overall, our data suggests that passive cough features distinguish TB from non-TB subjects and are associated with bacterial burden and disease severity.

(c) Findings indicate that passive cough based tools are not translatable to forced coughs as the model trained on passive coughs performed poorly on evaluating with forced coughs (sensitivity: 0.34+0.13). Distribution of scalogram features of passive coughs were also found to be different from forced coughs.

### *1.2 Thesis Document Outline*

This document is divided into eleven chapters with chapter 1 introducing the problem statement and thesis contribution. Chapter 2 details prior research in screening of both diseases as well background work of various technical tools used in the thesis. Chapter 3 to 6 focuses research contributions related to oral screening (1(a), (b), (c)) where the process of Fig. 1.1 is utilized. Chapter 3 explores oral pH as a critical biomarker and critically analyzes prior attempts to measure it. Chapter 4 and 5 represent design and validation of tool and chapter 6 discusses the improvements based on clinicians' feedback. Next, chapters 7-10 focuses on research contributions, 2(a), (b), (c), related to TB screening. Chapter 7 explains how cough can be a critical biomarker and discusses prior research and its limitations. Chapter 8-9 details how the dataset is collected and analyzed to evaluate utility of passive cough. Chapter 9 incorporates clinician feedback in evaluating the utility of the model and presents results that are important in development of such a tool in the field. The final chapter 11 summarizes the results obtained using the screening tool as well as implications of using them in the field.

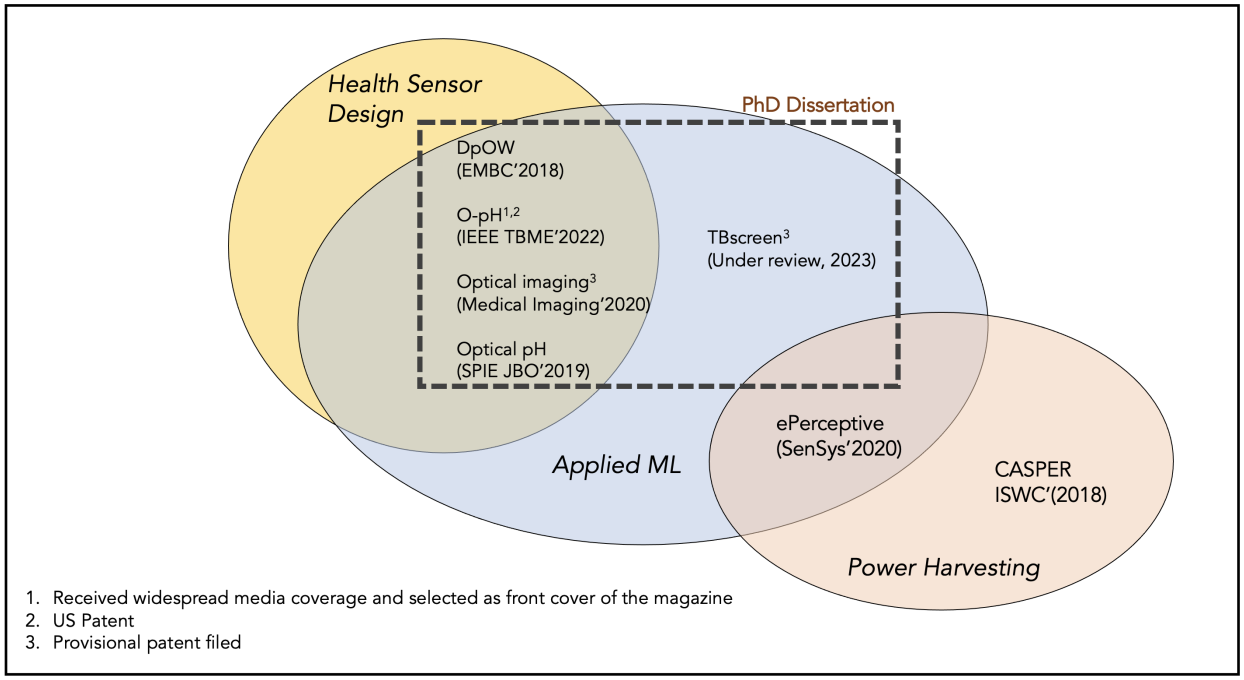


Figure 1.2: Scope of thesis

### *1.3 Dissertation Scope*

Research during author's PhD has focused on designing health sensors, applied ML, and power harvesting applications and their intersections. The dissertation represents subset of the work - designing disease screening tools as seen in Fig. 1.2.

# Chapter 2

# RELATED WORK

## *2.1 Dental Caries and Present Screening Tools*

Chronic caries in teeth, commonly known as tooth decay, is the most prevalent health condition affecting 2.3-3.5 billion people globally [66], [1]. Untreated caries can cause excruciating pain and lead to permanent tooth loss along with adding substantially to a family's medical expenditure [1]. Presently, visualization and tactile inspection is a standard procedure to evaluate dental surfaces. These techniques are the only gold standard for detecting early caries at occlusal (biting) and smooth surfaces (Fig.2.1(a)), while bitewing X-rays (Fig.2.1(b)) are the diagnostic tools used for caries at interproximal (in between teeth) regions. Lesion activity is determined by surface roughness and appearance whereas lesion depth is confirmed using X-rays.

Presently, dentistry is heavily focused on treatment based solutions rather than disease prevention. Probing tools and bite-wing x-rays are the most common procedures used in clinic to evaluate enamel health and look for carious lesion. Though, these techniques can detect or rank early-stage lesions at interproximal regions, they are often too late in detecting diseases in occlusal surfaces, and completely fail to provide any quantitative feedback for enamel hygiene. Recently, fluorescence based devices have been developed to aid in detection of early stage caries [121, 35]. DIAGNOdent, is a laser fluorescence (LF) device, which emits infrared laser light (655 nm), and scales the emitted fluorescence from enamel between 0 and 99 with values higher than 25 indicates caries lesion [102]. It's based on two principles, first, when the infrared light encounters a change in tooth tissue, such as porosity, it results in fluorescent light of a different wavelength. Second, bacterial metabolites such as porphyrines (proto-porphyrine, meso-porphyrine, or proporphyrin), result in the red fluorescence from

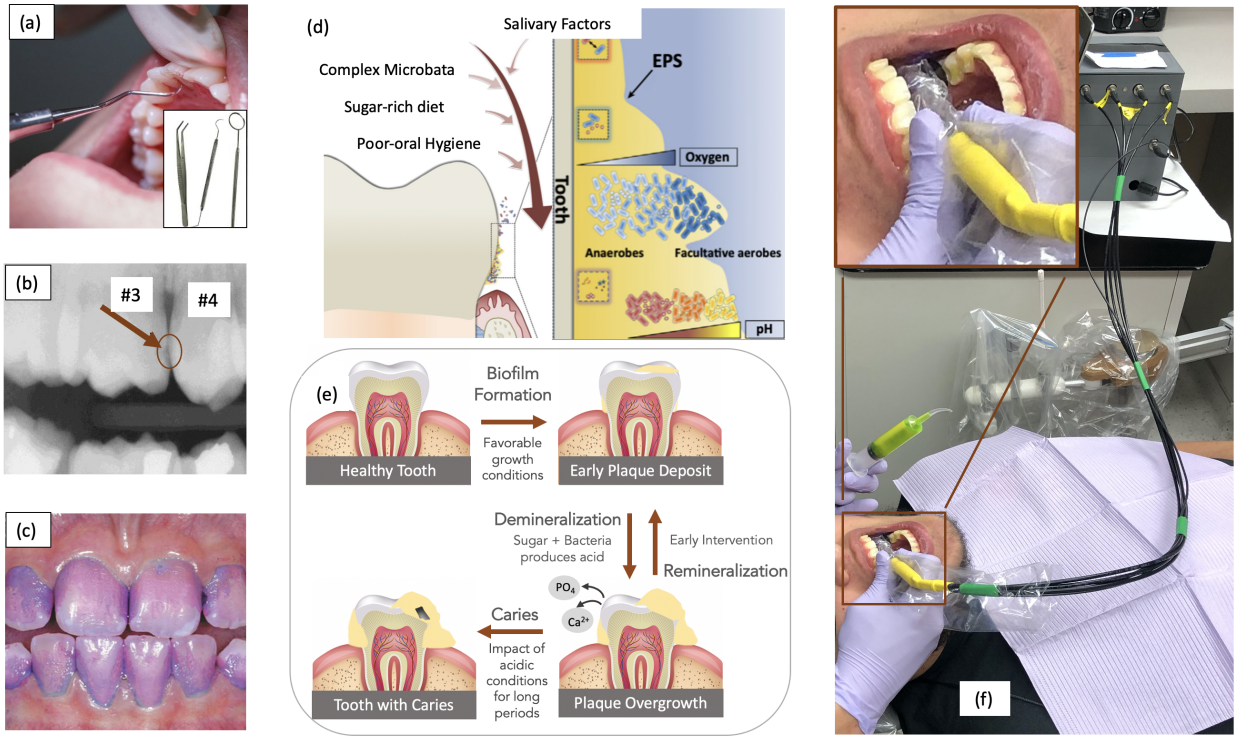


Figure 2.1: (a) Visual assessment using dental tools - gold standard for early occlusal caries. Inset figure shows different kinds of probing instruments used by dentists (b) Bitewing X-ray with an interproximal lesion between teeth 3 and 4 - gold standard for early interproximal caries [151] (c) Patient's mouth after using a biofilm disclosing agent to see dental biofilm coverage (d) Biofilm micro-environment: pH level is lower moving from surface to enamel [16]. Extracellular Polymeric Substance (EPS) composition and characteristic is shown in the inset figure. (e) Caries formation (f) O-pH in operation at a dental clinic with an inset figure showing a closer look of the device inside the mouth. The tip of the probe used to transmit and collect light is hovering over the occlusal surface of the subject. Detailed description of the device is provided in Fig. 4.1 and methods and materials section.

caries lesion [89]. Fluorescence cameras like VistaProof Durr Dental and Spectra Caries Detection Aid are based on similar principle but uses UV instead of IR [35, 12]. Another LED device, MID (Midwest Caries I.D., Dentsply Professional, York, PA, USAOCT), has been developed using red and IR light to analyze the reflectance and refraction of the light from the tooth surface [5]. All the above techniques have been tested in several clinical studies and have been found mostly insufficient in stand-alone detection of early primary or secondary caries [102, 12, 35]. Most of these devices depend on fluorescence properties of specific family of bacteria found around active lesions and is not specific to bacterial activity.

### *2.2 Oral Plaque Based Disease Screening*

Porphyrin based fluorescence used in devices like SOPROcare and has been developed further into low-cost mobile based devices. Angelino et al. [8] designed Plaquefinder, a low-cost, open source 405 nm device, and the associated computer vision algorithm that captured red fluorescence signatures associated with dental plaque and demonstrated comparable performance to commercially available devices. A miniaturized mobile adaptable version of the device was also provided to use the phone's camera to image plaque with additional blue LEDs. In LumiO, Yoshitani et. al [168] added red fluorescence technique to an electric toothbrush custom fitted with a camera to assist in brushing by increased visibility of plaque. They found qualitative evidence that study participants were able to improve awareness of plaque and build confidence on their tooth brushing. These devices though can enable home based plaque index monitoring and aid in practicing oral hygiene but are unable to track acidification of plaque making it less effective in preventing caries formation.

### *2.3 Sodium Fluorescein*

Fluorescein (FL) is an attractive fluorescent dye that offers high quantum efficiency, and it is often selected for intensity-based pH sensing. In addition, FL resides predominantly in the extracellular space (exclusion dye) since it does not penetrate the negatively charged bacterial cell wall. An excitation wavelength near the peak absorption band ($\sim$ 490 nm) is usually

employed and the emission intensity near 520 nm is recorded. Fluorescein emission intensity decreases in a near linear fashion as the pH is reduced. Unfortunately, fluorescence intensity depends upon several factors including the stability of the excitation light source(s), light scattering, dye photobleaching, dye quenching, etc. In a ratiometric pH sensor Rhodamine B, which is relatively insensitive to pH, was used in conjunction with fluorescein to partially compensate for these confounding factors [46, 21]. Encapsulation of the fluorescein and rhodamine dyes [42, 81] in a polymer or glass-like matrix helps prevent leaching of the dye materials into the environment. However, rhodamine is regarded as a health risk [65] and is not approved for *in vivo* human applications where direct contact with human tissue and tooth surfaces is required. A variant of the pH dependent FL intensity methodology appeared in a study of yeast cells [146]. Two spectrofluorometer wavelengths, 490 and 435 nm, were used for fluorescein excitation and the emission intensity at 520 nm was recorded for each excitation wavelength. A ratio of the 520 nm emission intensity at the two excitation wavelengths exhibited a linear log relation to pH over the range of 2.5 to 7. This intensity-based method relies upon the stability of the two excitation wavelengths which may be achieved in a laboratory spectrofluorometer but is difficult to replicate in clinical devices. There are FL lifetime dependent pH measurement techniques, but the change is in few nanoseconds in the desired pH range making it expensive and not designed for routine clinical use [132].

Overlooked in adapting FL fluorescence to pH sensing is the change in the relative proportion of its dianion and anion pH sensitive molecular variants [145, 72]. As the acidity level increases the predominant fluorescent species shifts from dianion to anion. These two species have overlapping absorption and emission spectral characteristics. Unmixing [75, 103] of the overlapping spectral emission data using least-square fitting of the endmember dianion and anion fluorescent species is performed to determine the pH. By a judicious choice of excitation wavelength, it is possible to balance the stronger dianion absorption/emission with the weaker anion absorption/emission to optimize the performance of the unmixing algorithm [139]. Other unwanted noise contributions, like background light and autofluorescence in the

range of 450-650 nm are removed before calculating the biofilm pH.

### *2.4 Tuberculosis and Present Screening Tools*

Tuberculosis (TB), caused by inhalation of *Mycobacterium tuberculosis* (Mtb), is the second leading infectious disease-related cause of death after COVID-19 [24]. After years of decline, the estimated incidence of TB and TB-related deaths increased in 2021, numbering 10.6 million and 1.6 million people, respectively. The current gold standards for TB diagnosis include sputum culture or GeneXpert molecular tests [70, 49]. The availability of these tests is limited in low resource settings, particularly at peripheral health centers. Given its ease of implementation, the WHO recommends symptom screening (assessing for the presence of fever, cough, night sweats, or weight loss), to identify people suspected of having TB. Unfortunately, the accuracy of symptom screening is suboptimal in both people with and without HIV [108, 33]. The WHO target product profile (TPP) for TB triage tests includes a test which is non-sputum based, rapid, low-cost, easy to use with minimal infrastructure requirements, and accurate (>90% sensitive, >70% specific) [106, 69]. Currently available TB screening tests do not meet these criteria [98, 147, 161].

TB disproportionately impacts low resource countries with 87% of new TB cases occurring in the 30 high TB burden countries. Two thirds of these cases were identified in Bangladesh, China, the Democratic Republic of the Congo, India, Indonesia, Nigeria, Pakistan, and the Philippines [24]. TB diagnosis using sputum culture has 100% sensitivity and specificity but takes over three weeks for processing in specialized laboratories, limiting its usage to national facilities [4]. Rapid Polymerase Chain Reaction (PCR) techniques have revolutionized TB detection with results within two hours but the high setup cost and recurrent need for cartridges limits its penetration to peripheral clinics [83]. Various TB diagnostic tools and its challenges in wide spread implementation is summarized in Table 2.1.

Most TB suspects in endemic countries present to peripheral healthcare facilities that may have no electricity, no running water, and limited or no laboratory facilities. This has led to a large number of cases being unidentified in rural areas fueling ongoing transmission in

Table 2.1: Various diagnostic tests for Tuberculosis

| Test | Sensitivity | Specificity | Setup cost | Time to result | Detection limit | Availability | Expert Training | References |
|---|---|---|---|---|---|---|---|---|
| Sputum Culture | 100% | 100% | High | 3 weeks | Low sputum producing patients | Limited to national labs in low resource setting | Yes | [4] |
| GeneXpert MTB/RIF | 82-88% | 96-98% | High | <2 hrs. | Variable sensitivity in HIV/Immunocompromised patients (73–87%) Low sensitivity in smear-negative patients (64–75%) | Limited to secondary clinics | Yes | [83] |
| Smear microscopy | 60-69% | 97-98% | Low | 1 day | Low sensitivity in low bacterial load Poor performance in pediatric and PLHIV | Limited to secondary clinicst | Yes | [31] |
| Chest X-ray | 73–79% | 60–63% | High | 1-2 days | Poor performance in PLHIV | Limited to secondary clinics | Yes | [118] |

the community. WHO estimates that nearly three million people develop active tuberculosis (TB), but are not notified to health authorities [136, 107, 104]. Reaching this 'missing three million' remains one of the top priorities to control TB. WHO end TB epidemic methodology lists the need for a low-cost point of care tool that require minimum infrastructure and expertise to expedite screening beyond the urban centers. Sputum smear microscopy, the global cornerstone of TB diagnosis, can miss half of all people with infectious TB [31, 29], whereas more sensitive tests cannot routinely be implemented at the point of treatment [111]. Present screening tools are primarily based on subjective evaluation of symptoms and has low sensitivity [104]. New techniques, for example, AI based Chest X-ray, C-reactive protein (Table 2.2) are under development but none of these techniques have met the target portfolio set by WHO to enable widespread convenient screening.

## *2.5 Audio Based TB Screening*

TB primarily affects lungs, termed as Pulmonary Tuberculosis, and manifests predominantly as cough. Since a long time, initial symptom screening of TB includes subjective evaluation of prolonged coughs but have low sensitivity. Recently, researchers have focused on digital evaluation of cough frequency and features for TB screening [15, 155], discussed further in Chapter 7. Apart from cough analysis, manually listening of lung sounds for crackles or

Table 2.2: Tuberculosis screening tools

| Test | Above Target cost | Availability at lower clinic | Limitations | References |
|---|---|---|---|---|
| Symptom screening | No | Yes | Subjective | [108] |
| Automated Cough screening | No | Yes | Early stage | [171] |
| Automated Chest X-ray | No | No | Early stage | [123, 167] |
| C-reactive protein | Yes | Yes | Require additional training | [167] |
| Seriological Tests | No | Yes | Inconsistent results | [148, 105] |
| Urine RDTs | No | Yes | Low sensitivity for non-HIV infected population | [96, 47] |

wheezes (lung auscultation) is also used by clinicians to screen for lung damage due to TB. But lung auscultation is often ruled out as a reliable diagnostic technique for TB due to the random distribution of the infection and the varying severity of lung damage [87]. A 1981 study digitally compared respiratory sounds from normal subjects and TB patients and found a distinct difference in wave shapes between sound signals recorded from normal and diseased subjects [92]. Another recent study captured respiratory inhaling and exhaling movements using seven electronic stethoscopes attached using a chest strap. Features were generated in time, frequency domain for adventitious wheezes and crackle analysis and indicated a degree of separation between recordings from healthy lungs and recordings from TB-infected lungs[11].

## *2.6 Audio Features*

Audio signals are captured using microphones and digitized using analog to digital converters. The sampling rate of the ADC conversion determines the maximum frequency content of the audio captured. Using Nyquist equation, an audio sampled at 16KHz contains at maximum 8KHz of frequency content [77]. Along with sampling rate, frequency range of the microphone determines the frequency that can be digitally recorded. Most commercial microphones found in smartphones, recording studio, conference microphone can capture up to 20 KHz of sound frequency[19].

### *2.6.1 Time Domain*

Audio signals are 1-dimensional signals and combinations of various sinusoids as represented in Fig. 2.2(a). Mathematically, sound wave of frequency f can be represented as 2.1. Sampling the signal in digital domain gives us a digital signal x (Eq. 2.2) of length N (Eq. 2.3). Several waveform features like, zero crossing rate, amplitude envelope, root mean square energy of audio are extracted for sound analysis.

$$x(t) = Asin(2\pi ft + \phi) \tag{2.1}$$

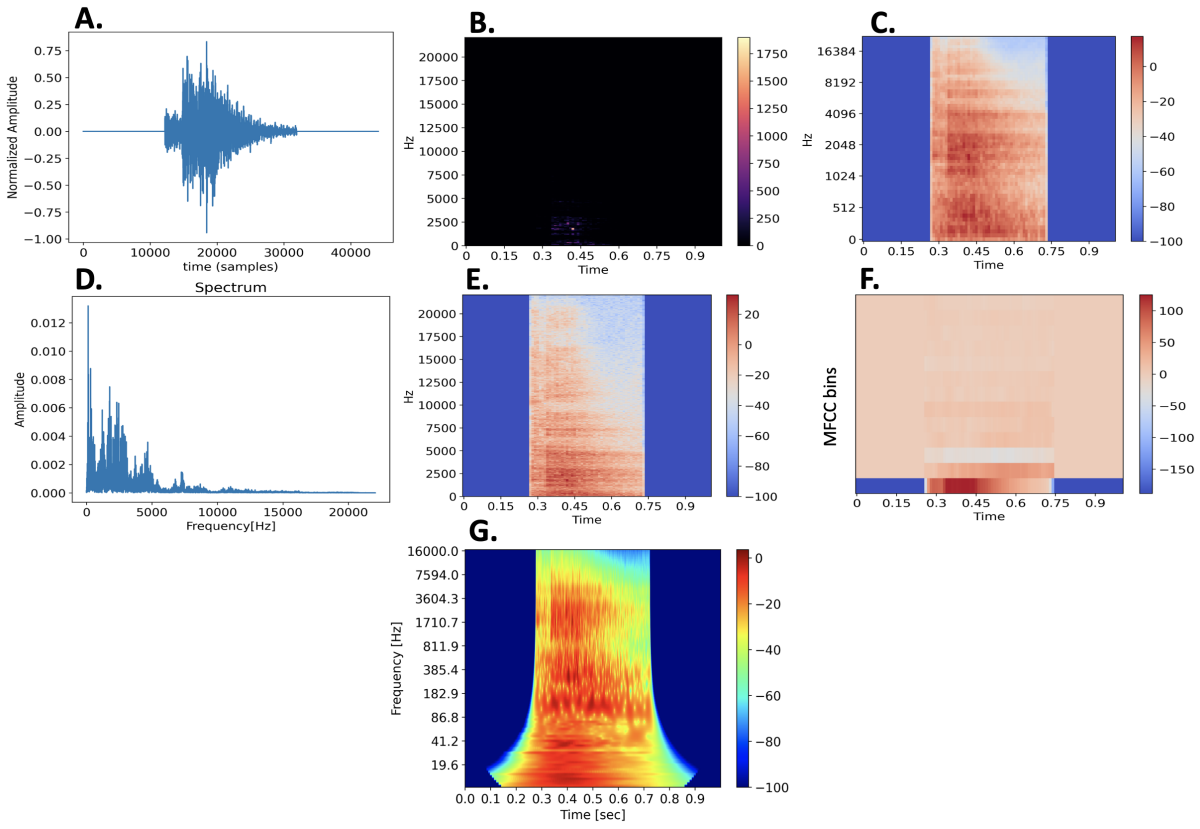


Figure 2.2: (a) Audio signal in time domain (b) Spectrogram (c) Log Melspectrogram (d) Frequency domain (e) Log Spectrogram (f) MFCC (g) Scalogram

$$x = \{x(n)\}, 0 < n < N \tag{2.2}$$

$$N = sampling\ time * sampling\ rate \tag{2.3}$$

### *2.6.2 Frequency domain*

Time domain audio signal can be represented in frequency domain using Fourier transform [17]. Fig 2.2(d) represents the frequency spectrum of Fig.2.2(a), x-axis consists of various frequency present in the signal and y-axis represents the amplitude associated with each frequency. Frequency spectrum, a 1-D representation, is used commonly to separate sound sources of different frequencies in an audio. Mathematically Fourier transform of time varying signal x(t) is represented by equation 2.4, whereas in digital domain, Fourier transform of digital signal x, is computed using Digital Fourier Transform (DFT, Equation 2.5) [28]. This representation provides frequency content of the overall signal but not how the frequency changes over time.

$$X(f) = F(x(t)) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt \tag{2.4}$$

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2\pi i k \frac{n}{N}},\ k = 0..N-1,\ N = length(x) \tag{2.5}$$

### *2.6.3 Spectrogram*

Spectrograms are 2-D image representations of audio signal and captures frequency-time relationship of the audio [172]. In a spectrogram representation plot, one axis represents the time, the second axis represents frequencies, and the colors represent magnitude (amplitude) of the observed frequency at a particular time. Fig. 2.3(b) represents the spectrogram of the audio signal.

It's calculated using Short Time Fourier Transform (STFT), where DFT is applied to short fragments of time, that is, frames taken from a longer signal. For the image representation, individual frames are stacked horizontally, so that time can be read left-to-right, and frequency can be read bottom-to-top. Typically, spectrograms, are magnitude spectrograms,

where the phase component has been discarded and only the DFT magnitudes are retained. There are three main parameters that are tuned for producing [6]:

- Frame Length ($N_F$): The length of each frame that the longer signal is divided into. $N_F$ introduces a time-frequency trade-off as longer frame length provides high frequency resolution at the cost of lower time resolution and vice versa for shorter frame length. For efficient computation, N is generally selected as an integral power of 2.

- Hop Length ($H_F$): The length of audio signal that is skipped between each frame to compute DFT. Small values of the $H_F$ produces redundant outputs (high dimensional image) whereas larger values provide coarser time resolution (lower dimension image). Frequency resolution is independent of $H_F$.

- Windowing Function: For the looped frame to appear continuous, each frame is sample-wise multiplied by a windowing function that tapers to 0 at the beginning and end. Commonly used windowing functions for are Hann, rectangular among others.

Spectrograms are generally further transformed for audio analysis as list below [101, 38]:

- Log spectrogram: Amplitude spectrum is converted to log scale using Eq. 2.6. As we see in Fig. X, log spectrogram provides more information than spectrogram in Fig 2.2(e).

$$S_{dB} = 20log_{10}S \tag{2.6}$$

- Melspectrogram: The mel scale is a non-linear transformation of frequency scale based on the perception of pitches. The mel scale is calculated so that two pairs of frequencies separated by a delta in the mel scale are perceived by humans as being equidistant (Eq.2.7 ). Mel spectrogram when converted to dB scale is known as log melspectrogram and is represented in Fig. 2.2(c).

$$M(f) = 1125ln(1 + \frac{1}{700}) \tag{2.7}$$

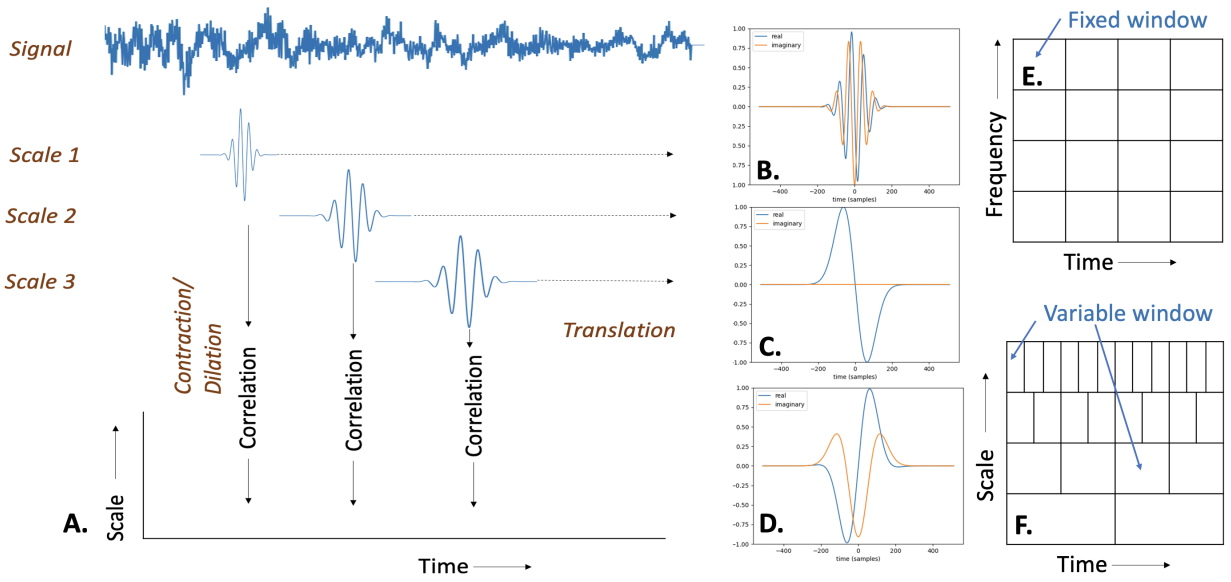


Figure 2.3: (a) Scalogram formation using daughter wavelets (b) Complex Morlet Mother wavelet (c) Mexican hat mother wavelet (d) Complex Gaussian wavelet (e) Time-frequency resolution in a spectrogram (f) Time-frequency resolution in scalogram

- Mel-Frequency Cepstral Coefficients (MFCCs): The MFCCs are calculated by applying the discrete cosine transform (DCT) to a log melspectrogram (Fig. 2.2(f)).

### *2.6.4 Scalogram*

Scalograms are alternate approach to spectrograms for generating frequency-time relationship of a signal using continuous wavelet transform (CWT) [141, 84]. Similar to the STFT, the CWT uses an analysis window called mother wavelet to extract signal segments. Mathematically, CWT is expressed by Eq. 2.8. Scalogram is generated (Fig.2.3(a) by a time plot of the correlation between the signal and the scaled wavelets. Scale and frequency are inversely related. Unlike the STFT, the analysis window or wavelet is not only translated (moved across x-axis) but dilated and contracted (window size varied over y-axis) depending on the scale of activity selected. Wavelet dilation increases the CWT's sensitivity to long time-scale events, and wavelet contraction increases its sensitivity to short time-scale events providing

a better time-frequency resolution than STFTs. Fig. 2.3(e-f) represents the difference in time and frequency resolution of the two representations.

$$C(a,\tau) = \int \frac{1}{\sqrt{a}} \psi \frac{t-\tau}{a} x(t) dt \tag{2.8}$$

$$\psi(t) = wavelet,\ \tau = time\ shift,\ a = dialation$$

Wavelets are signals that oscillate around zero and behave like bandpass filters, Fig. 2.3(b, c, d) represents common mother wavelets which on dilation and contraction produce various daughter wavelets for computing CWTs. Mother wavelets are selected based on the resemblance to the shape of the signal to be analyzed.

CWT requires significant computation time but also produces highly redundant data. Reducing number of scales makes the processing faster while not significantly affecting the information, and often scales are selected using, $2^{N*Scale}$, to put the frequency band for higher frequencies further away than for smaller frequencies [45, 153]. All selected scales are within the Nyquist limit of the sampled signal.

To generate scalograms, we translate the wavelet in time resulting in wavelets to see data outside the observation interval near the boundary. This results in boundary effects termed as cone of influences giving erroneous data at the beginning and end. This error depends on the scale of the wavelet and size of the wavelet is connected to its scale, hence for different scales the cone of influence has different sizes. In case of complex Morlet wavelet, edge effects are seen at $\sqrt{2} * scales$ and are discarded as shown in Fig Fig. 2.2(g) [153].

# Chapter 3

# ORAL pH: CRITICAL BIOMARKER FOR ORAL HEALTH

Present dental tools and procedures provide patients with lagging, non-quantitative feedback assisting inadequately in prevention of new caries or in evaluating site-specific risk of caries development. Despite of oral care playing a significant part of a healthy daily routine, from brushing twice a day, frequent flossing, avoiding foods with excessive sugar, and minimizing snacks in-between meals, in addition to bi-annual dental visits, patients are still unable to evaluate effectiveness of their daily oral-care. Dentists, on the other hand, can't objectively confirm if the patients, especially adolescents, are effectively performing their daily care routine unless a suspicious spot is clinically evident. There is a need to interject this present cycle of waiting-and-watching for a lesion to appear, to evaluate oral well-being using tools that can provide leading indicators for oral health. A leading indicator, a terminology commonly used in occupational health systems [7], provides pro-active, predictive risk assessment unlike lagging tools that assess information after an event has already occurred, particularly in our case, after a carious lesion has formed. Similar to a visit to a general physician where measurements like heart rate, blood pressure, and blood work provide a baseline quantitative information, dentistry could benefit with quantitative measurements of the risk factors that are directly correlated with caries formation and can be safely monitored over time to understand the status of oral health. The current adjunct diagnostic tools are focused on measuring the presence of the disease, rather than assessing the risk of developing active caries.

One of the techniques to obtain quantitative measurement of caries risk is by developing tools to monitor oral enamel biofilm - the sticky, yellowish coating found on teeth surfaces which plays a crucial role in early caries. Presently, dental biofilm (also referred

to as plaque) is evaluated using visual quantitative measurement techniques like Quigley Hein plaque index [156] that measures and ranks dental biofilm coverage with help of probing tools but is unable to objectively evaluate cariogenesis of biofilm. Similarly, disclosing dyes (as shown in Fig.2.1(c)) assist in visual inspection of dental biofilm, though staining of teeth makes the use uncommon. There are fluorescent based devices like SOPROcare and Q-Ray that capture fluorescence by exciting porphyrin found in oral biofilm [127, 120] with blue light. These devices increase dental biofilm visibility and also indicate dental biofilm maturity which is proportional to the intensity of porphyrin's red fluorescence. Though these fluorescent devices provide leading indicators, they focus on very specific porphyrin producing bacterial groups (*Streptococcus mutans*, etc.) [57, 121], ignoring the impact of vast number of (over 700) microbes found across different oral cavities [71, 164] and are confounded by food stains, lowering specificity as a stand-alone leading indicator of caries. Several low-cost, at-home, dental biofilm monitoring devices have been proposed, for example, Angelino et al. [8] designed Plaquefinder, a low-cost, open source 405 nm LED (Light Emitting Diode) based device, and the associated computer vision algorithm that captured red fluorescence signatures associated with dental biofilm and demonstrated comparable performance to commercially available devices. Similarly, with LumiO, Yoshitani et. al [168] added red fluorescence technique to an electric toothbrush custom fitted with a camera to assist in brushing by increasing visibility of dental biofilm. They found qualitative evidence that study participants were able to improve awareness of dental biofilm and build confidence on their toothbrushing. These devices can enable home based dental biofilm index monitoring and aid in practicing oral hygiene but are unable to track acidification of dental biofilm making it less effective in preventing caries formation.

### *3.1 Formation of Dental Caries*

Our mouth with its optimum temperature (35-37°C), neutral pH, and frequent access to nutrients is a breeding ground for several hundred species of micro-organisms, found around tooth surfaces and gum lines [93]. On consumption of carbohydrates, bacteria in the dental

biofilm produce acid which is slowly neutralized by the action of saliva. This compensating mechanism can be disturbed with frequent consumption of sugar rich food, lack of proper dental hygiene, disruption in flow of saliva, and other lifestyle habits, increasing the acid production, its frequency, and duration of acid exposure to enamel. This leads to a change in micro-environment favoring growth of harmful bacteria that can survive in low-pH and anaerobic conditions as shown in Fig.2.1(d). If left unmonitored without intervention, extended exposure to acid can degrade the tooth enamel of minerals to become a demineralized lesion and ultimately cause carious cavitation as depicted in Fig.2.1(e).

Thus, routine monitoring of the acid producing function of the biofilm which plays an early critical role in the degradation of enamel can help us understand pH changes as a leading site-specific risk indicator to caries formation.

### *3.2 Oral Biofilm Acidity*

In 1938, Dr. R.M. Stephan, pioneered the research of measuring oral plaque acidity by examining sampled plaque mixed with pH indicator dyes *in vitro*; and later he measured enamel plaque pH *in situ* using a custom antimony based pH micro-electrode [149, 90]. In the 1944 study [91], Stephan found that the plaque pH decreased from a baseline measurement (resting pH) after a glucose rinse, and takes up to 40 min to return to its resting value from the buffering action of saliva. This change in pH over time after a sugar rinse (from resting to rapid drop with sugar metabolism, to a slow rise from consumption of sugar and dilution from saliva), is now called the Stephan Curve' as shown in Fig.3.1(a). Since then, several studies have examined Stephan curve and found different sections of the curve, resting pH [41, 76], minimum pH after the sugar rinse [130, 86], time taken to return to resting pH [37], related to caries activity. Though studies have noted that Stephan curve measured on individual tooth surfaces is noisier than the averaged curve reported by Stephan [41, 86]. Difference in study variables like concentration and type of carbohydrate, duration of carbohydrate rinse, site of measurement, pH of rinse, initial pH of enamel plaque, number of days from the cessation of tooth-brushing, have made it difficult to compare pH quantitatively across different studies.

Nonetheless, the overall average Stephan Curve trend is exhibited consistently across studies.

In a 2000 study, Lingstorm et al. [86] studied pH in low and high caries risk group, with first group subjects who are caries free and latter having subjects with both sound teeth and white spot lesions. They found statistical difference for minimum pH (from sugar rinse drop) for sound teeth in low caries risk subjects vs. white spot sites in higher-caries risk subjects as well as for sound and white spot lesions in the latter. Though, the resting pH was similar regardless of subject or tooth caries status. This appears to be the only work that studies variability of pH from specific tooth surfaces from within a mouth. Like previous studies [91, 37] the measurements were averaged for teeth surfaces, though sound and non-sound data was analyzed separately.

In a follow-up study [122], on a hundred adolescence subjects in Sweden, Quiroz et. al validated prior results and showed that oral plaque pH can be used as a method for discriminating between individuals with varying caries risk. Similar study by the team [125] was repeated on the Karen tribe in Thailand known to have low caries risk, showing the dental plaque in both Karen children and adults responded to a sucrose rinse but to a smaller extent than seen in Caucasians. On average, the Karen tribe reported a higher resting pH than the parallel study in Sweden indicating that resting pH could be associated with caries risk.

Along with acidity of oral biofilm, pH of saliva (the buffering agent responsible to maintain neutral pH) plays an important role in evaluating overall oral health. Prior studies have shown that alkalotic salivary pH is associated with generalized chronic gingivitis whereas acidic pH is associated with generalized chronic periodontitis [112, 73, 9]. Monitoring of salivary pH has potential to indicate generalized oral health though enamel localized information is limited.

### *3.3 Plaque pH Measurement Techniques*

*In vivo* oral plaque pH measurements are mainly performed with pH micro-electrodes. Over the years, different kinds of pH electrodes were used to measure plaque pH *in vivo* with Beetrode iridium–iridium oxide micro-electrode being the most common. To measure pH,

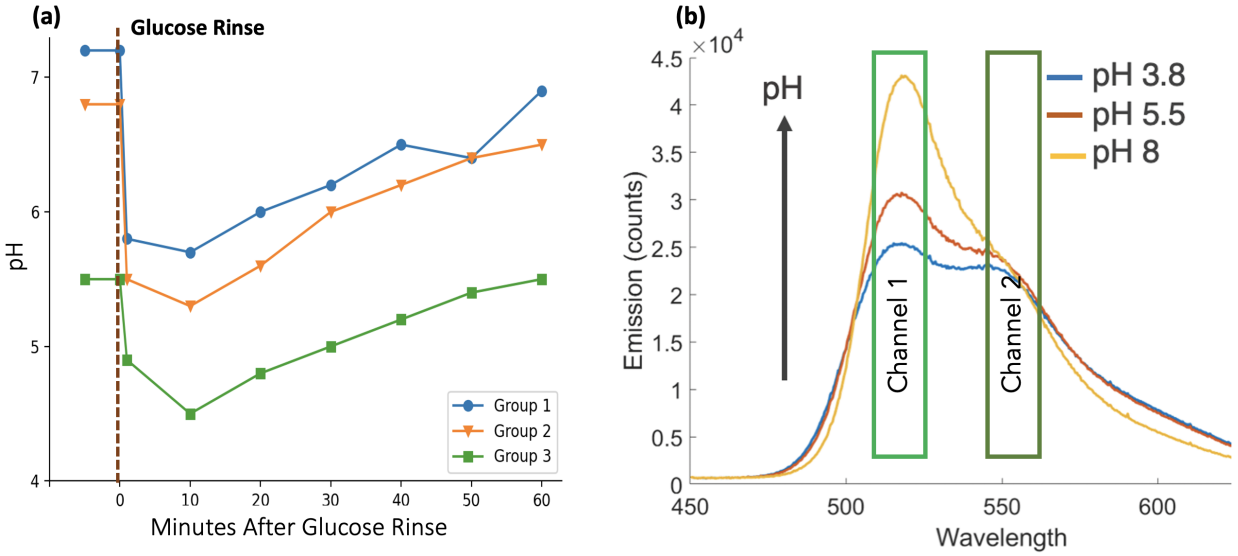


Figure 3.1: (a) The Stephan curve, pH response of oral film immediately after a sugar rinse and monitored upto 1 hr. in three subject groups with different caries risk (Group 1: caries free, 2: slight caries activity, 3: extreme caries activity.) Several studies have shown that drop in pH after sucrose rinse is dependent on caries activity in the region [91]. The graph includes three of the 5 categories of subjects represented in 1944's Stephan Curve. (b) Fluorescence spectrum of aqueous solution of sodium fluorescein in different pH solutions obtained using 420 nm LED excitation and captured with a spectrometer. O-pH uses peak at 520 and 550 nm to measure pH.

a salt bridge is created with a reference electrode and subject's fingers placed in 3 M KCL solution [119]. Modern electrodes are only 0.1 mm in diameter making them suitable for many interproximal plaque pH measurements, though their fine structure and need for a reference makes them prone to fragility, breakage, and inconvenience. Another critical issue encountered is the measurement of pH at the plaque-saliva interface [119]. Lingstrom et al., 1993 measured pH oral biofilm grown on indwelling electrodes as surrogate enamel substrates which demonstrate values >1 pH unit lower compared to more common micro-electrodes that measure at the saliva interface. With our device, the pH-sensitive fluorescein dye rapidly diffuses extracellularly throughout the biofilm which enables a pH measurement that is more representative of bacterial activity at the plaque-enamel interface. Indwelling electrodes

are mounted on removable partial dentures and used for transmitting pH of interdental plaque but suffers from measuring pH of biofilm that has grown on glass electrode instead of enamel [119]. Techniques like plaque sampling, where plaque is removed from different teeth surfaces (∼20) has been used in several studies to measure pH *in vitro*. This method though capable of providing an average pH of biofilm loses teeth location accuracy while also causing biofilm disturbance. Recent work has also used pH strips to measure pH at interproximal sites and found high correlation with reference electrode pH measurement [23, 27]. But, pH strips are difficult to insert at interproximal spots without wedging and unable to measure in deeply pitted occlusal surfaces. Confocal laser scanning microscopy (CLSM) has been used to measure the pH depth profiles of dental bacteria biofilms [165, 133]. Fluorescent dyes were embedded into the biofilm matrix and image processing algorithms are applied to separate the bacterial biomass and extracellular regions. Separating the two regions is important since only the organic acids residing in the extracellular regions lead to enamel demineralization. The confocal studies reveal the spatially heterogeneous nature of the biofilm matrix where the lowest pH values are found in the deep anaerobic regions attached to the host surface. Biofilm heterogeneity is often attributed to the diffusion limited transport of nutrients through the extracellular exopolysaccharide (EPS) matrix surrounding the mixed species bacterial microcolonies. CLSM investigations provide a valuable insight into the 3-D structure of plaque-like biofilms but adapting this technology to the dental clinic is currently not feasible. Endogenous oral bacteria porphyrin fluorophores have been used to measure pH levels in suspensions of oral bacteria [60, 59]. Porphyrins have an intense absorption peak near 405 nm and emit a red fluorescence that is associated with plaque deposits. As the pH environment of the porphyrin was decreased the peak fluorescence wavelength (635 nm) broadened and a new feature appears as a blue-shifted shoulder at 622 nm [60]. It is likely that the new feature is associated with porphyrin aggregation [134]. Relative intensity of the two spectral features was shown to be a linear function of pH [60]. However, production of porphyrin fluorophores depends on the thickness and age [159] and bacterial species in the plaque [160, 82] and the porphyrin species are not photostable

[61]. We also found that porphyrin emission was either absent or weak in human plaque scrapings. Fluorescent nanoparticles and quantum dots have been extensively explored for pH sensing [58, 20, 39]. They offer a wide range of fluorescent wavelengths which depend upon the materials and geometry of the core-shell particles. Low temperature sol-gel fabrication permits the inclusion of organic fluorophores which can afford chemical sensing of the local environment for nanobiotechnology (lab-on-a particle) applications. Diffusion into dense biofilms will be slow unless the particle size is less than 10 nm [117]. However, these materials are costly to produce and are not approved for human studies. A water-soluble polymer material with a ratiometric fluorescent pH sensing dye tethered to the polymer backbone was demonstrated using bacterial cells [170]. The intended application of the polymer bound fluorescent dye material is in a cell plate reader rather than an in situ oral biofilm pH sensor.

All of these pH measurement techniques are challenging to be used in a clinical environment for a routine checkup on all enamel surfaces.

# Chapter 4

# O-pH: OPTICAL pH SENSING FOR ENAMEL HEALTH MONITORING

This chapter focuses on designing a spot based oral pH monitoring tool, O-pH which can be conveniently used in clinical setting with relatively low cost and high efficiency after a one-time calibration [140, 135].

## *4.1 System Design*

### *4.1.1 Sodium Fluorescein Properties*

Sodium Fluorescein (Fl), is a dye commonly used as diagnostic tool in ophthalmology and approved by FDA for human use. In the aqueous solution it has a peak absorption band at ∼490 nm and fluoresces with a wide spectra from 500 to 650 nm with a distinct peak at 520 nm. This emission intensity is directly proportional to the extracellular biofilm pH. Additionally, Fl has been shown to rapidly penetrate dental biofilm extracellular matrix making it an ideal candidate for pH measurement of dental biofilm [138], [126], [22].

Sjoback et al. [145] have shown that in aqueous solution, Fl exhibits an equilibrium mixture of four different species: cation, neutral, anion, and dianion. Out of the four, only the dianion and anion species are fluorescent, having different absorption and emission peak, and pH dependent concentration in the solution. For example, at pH 4 and lower, a Fl solution consists of predominantly anions, and at a pH 9, the solution mainly has dianions resulting in different spectral properties in the 450-650 nm range [139]. Solutions between pH 4-7.5 contain both dianion and anion species resulting in a fluorescent spectral profile that is a mixture of individual emission profiles [Fig.3.1(b)] distinctly observed by selecting an excitation wavelength that can excite both species (∼420 nm). As previously demonstrated,

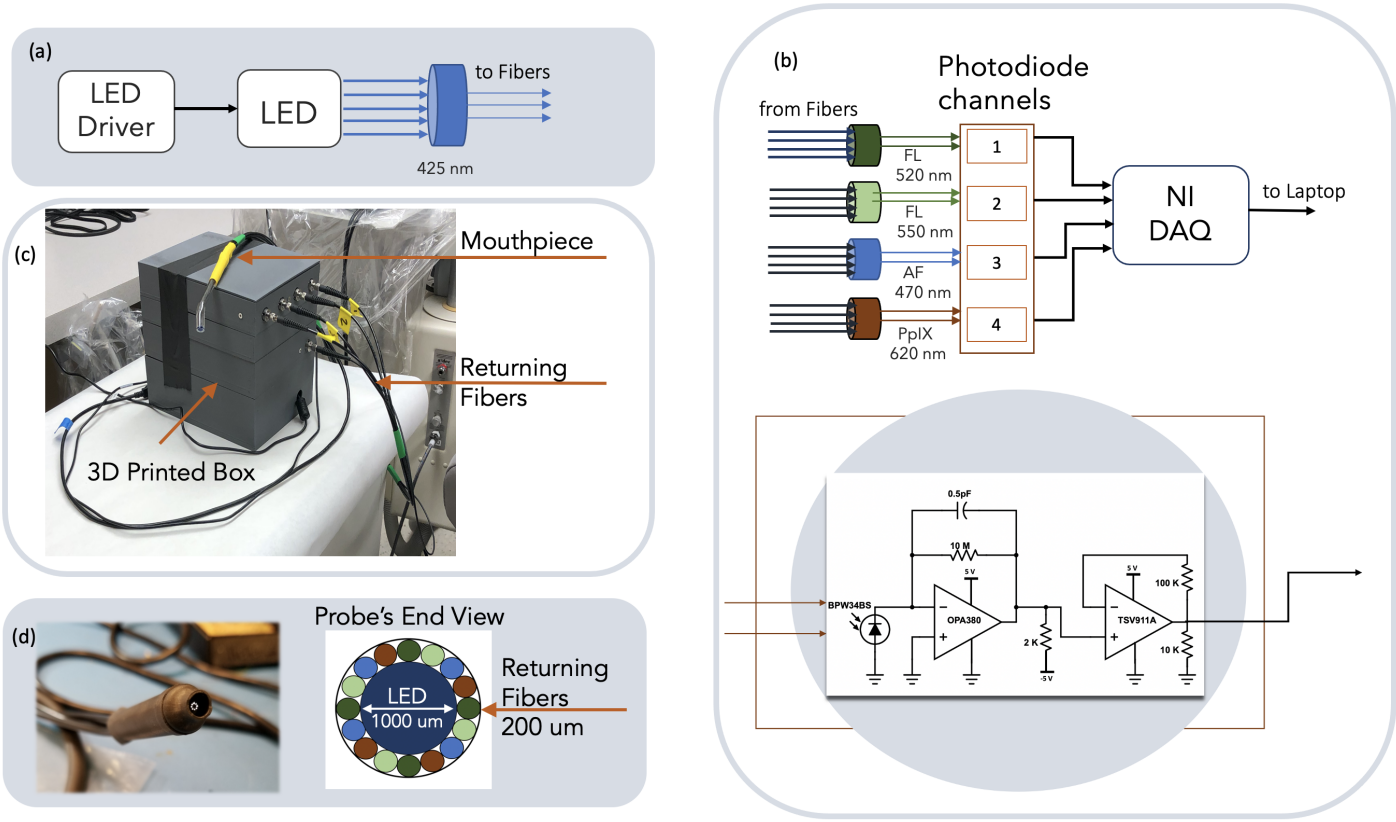


Figure 4.1: Device Architecture: (a) Excitation Unit (b) Photodetector Unit [FL: Fluorescein, AF: Auto Fluorescence, PpIX: Porphyrin] with schematic for photodiode channel (c) 3-D printed box with optical fibers attached (d) Fiber optics probe and its end view.

Fl emission spectra captured using spectrometer can be unmixed with least mean square to predict pH [139]. Our prototype, O-pH, uses distinct fluorescence properties of Fl dianions and anions species, but instead of using the entire spectra, it utilizes only the two peaks at 520 and 550 nm to calculate pH in the range of 4-7.5 [Fig.3.1(b)].

### *4.1.2 Device Architecture*

The device architecture consists of three components: (a) excitation unit (b) detection unit (c) mouth probe.

The excitation unit is used to excite the Fl solution and comprises a LED driver (Thorlabs,

LEDD1B) pulsing a blue LED (ThorLabs, M420F1) at 500 Hz with 5W. The pulsing LED light is filtered using a fluorescence, band pass filter (Semrock, FF01-425/26-25) centered at 425 nm to limit the bandwidth of the excitation wavelength (Fig. 4.1 (a)) and block out-of-band emissions [67].

The emitted fluorescence on absorption of LED light is measured using the detector unit which consists of four independent, optically filtered, photodiode channels Fig. 4.1(b). Different channels of the detector unit are used to capture Fl fluorescence and low signal emissions. Channels 1 and 2 of the photodiode board is used to detect Fl anion and dianion fluorescence intensity. Channel 1 uses a band-pass filter (BP) centered at ∼520 nm(Semrock, FF01-524/24-25) to measure emitted photons from dianions and Channel 2 uses a BP filter centered at ∼550 nm(Semrock,FF01-549/12-25) to measure emission from anions. Channels 3 and 4 are used to detect low level fluorescence in the mouth that can be excited by the 420 nm LED light, namely auto-fluorescence (AF) and porphyrin's (PpiX) fluorescence. These channels use a filter centered at 475 nm (Semrock,FF02-475/20-25) and another centered at 632 nm (Semrock,FF02-632/22-25) for AF and PpiX respectively. Each photodiode circuit, shown in Fig. 4.1(b), consists of a Silicon photodiode (BPW34BS) where the incoming photon is collected, generating current which is then converted to voltage using a transimpedance amplifier (TI, OPA380) with a gain of 10M V/A. The output voltage of the transimpedance amplifier is amplified using a non-inverting amplifier (TSV911A) with a gain of 11 V/V. The final output voltage is sampled using National Instrument's data acquisition unit (NI,DAQ600) at 10KHz frequency.

The above two units are housed inside a 3D printed box, shown in the Fig. 4.1(c) with jacketed optical fibers coming out of the box. The fiber optics bundle consists of central 1000$\mu$m fiber (ESKA, Mitsubishi) that carries the excitation light from the LED, and surrounded by sixteen returning 200$\mu$m fibers carrying the emitted fluorescent light to photodiodes. Each photodiode channel inside the box is coupled to four optical fibers to receive emitted photons. The length of all fibers is one meter to provide flexibility for the operator to probe far back in the mouth with the device. These fibers terminate in a

hand-held dental probe; such that the tip of the probe has the excitation fiber in the center surrounded by returning sixteen fibers in a circular ring. Fig.4.1(d) shows the image of the probe's tip and it's end view. A rubber barrier is used at the tip of the probe to avoid physically touching the fibers tip to subject's teeth and is changed for every subject.

### *4.1.3 Algorithm*

The sampled voltages from the DAQ is transformed to frequency domain using Fast Fourier Transform. The amplitude of signal corresponding to 500 Hz is recorded for each photodiode channel. This is the frequency of the pulsing blue LED and selecting the voltage amplitude corresponding to this frequency helps in discriminating against background light. Extracted fluorescence reading from channel 1 and channel 2 is then used to calculate pH. Channel 3 recording is utilized to measure the AF noise which acts as a threshold to accept or reject estimated pH. This threshold is estimated during the calibration process. Channel 4 data is used to measure PpiX fluorescence as another indicator of dental health.

## *4.2 Device Calibration*

O-pH requires a one-time calibration for pH measurement. We describe the calibration process and device accuracy in subsequent sections.

*Chemical Preparation:*

1 Molar stock solution of sodium fluorescein (Sigma Aldrich and ScienceLab) was prepared in deionized water. The fluorescein solution was diluted in phosphate citrate buffer (0.2M dibasic sodium phosphate, 0.1M citric acid, pH indicated for each experiment), 0.1 M sodium bicarbonate buffer, or chemically defined medium (CDM) buffer to form solutions in the range of 4 to 7.5 pH [139]. These solutions of $200\mu M$ concentration were used for calibration of pH device with a conventional pH meter (ThermoFisher Scientific).

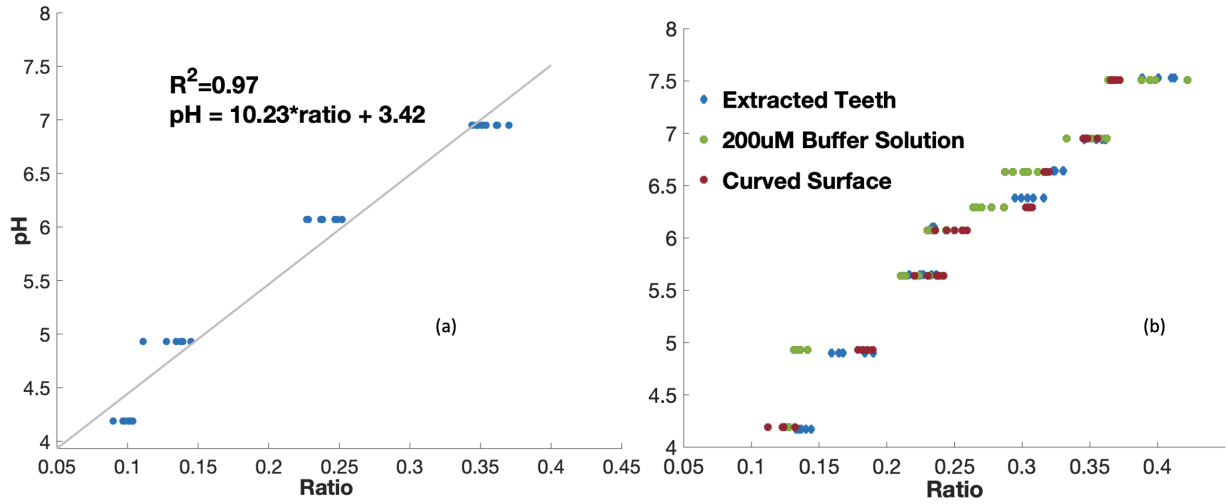


Figure 4.2: (a) Calibration curve using buffer solution in a 1mm cuvette. Ratio is given by equation 1. (b) Verification of calibration curve using 200uM buffered fluorescein in 1mm cuvette, on extracted human teeth, and on artificial curved teeth surfaces (occlusal, interproximal, and buccal surfaces of artificial teeth). A drop of fluorescein is added on different teeth surfaces and pH is measured using O-pH.

*Fluorescence Measurement:*

Using a 1 mm glass cuvette, we measured fluorescence of $10\mu L$ of four different $200\mu M$ Fl buffers ranging from pH 4 to pH 7.5. Each measurement was repeated ten times to obtain the calibration curve as shown in Fig. 4.2(a). A linear relationship was obtained between pH and ratio defined in Equation 1 with a correlation coefficient of 0.97.

$$ratio = \frac{Ch1 - Ch2}{Ch1 + Ch2} \tag{4.1}$$

$$pH = 10.34 * \frac{Ch1 - Ch2}{Ch1 + Ch2} + 3.42 \tag{4.2}$$

Table 4.1: O-pH Accuracy

| pH Range | Mean Error | Std Deviation |
|---|---|---|
| 4-4.5 | 0.57 | 0.09 |
| 4.5-5.5 | 0.27 | 0.15 |
| 5.5-6.5 | 0.18 | 0.09 |
| 6.5-7.5 | 0.13 | 0.08 |
| Overall(4-7.5) | 0.22 | 0.16 |

### *4.3 In Vitro Verification*

We verified the calibration curve by measuring different Fl buffers in the same pH range using the 1mm cuvette used in calibration. Since the calibration curve was obtained using a flat surface but *in vivo* testing would be performed on irregular surfaces, so the device was verified on artificial teeth surfaces (Perio 525 Typodont, frasaco GmbH). We dispensed Fl on occlusal, interproximal, and buccal surfaces to measure pH values. Next, we tested Fl on extracted human teeth to see the effect of low signal levels of AF.

We found the pH measurement was robust to AF if the AF signal is below a threshold. This threshold was noted and used in clinical testing to discard measurements. All the predicted pH values are plotted in Fig.4.2(b), obtaining an overall correlation coefficient of 0.92. The device had an overall error of 0.22 pH with 0.16 standard deviation. O-pH device accuracy in various pH ranges are listed in Table 4.1. We found that fluorescence readings of channel 1 and channel 2 made inaccurate predictions if the fluorescence was too low, but this signal could be amplified by increasing the excitation power. Distance of the probe from measuring surface doesn't affect the accuracy if the fluorescence signal strength is above this threshold. With our maximum current and voltage setting, we found that at a separation distance up to 3mm, the device probe provided accurate results. To note, we bounded our measurements between 4 and 7.5 pH, discarding any values outside this range as inaccurate.

### *4.4 Limitation*

The accuracy of O-pH was verified with *in vitro* studies using buffered fluorescein solutions and pH meter. *In vitro* study to understand sugar response using lab grown biofilm was not performed. Alternatively, resting pH could be verified *in vitro* by collecting biofilm from the subject's mouth and measuring pH after dilution with water. But this method was not adopted as it would have caused disruption of dental-biofilm and also reduced the number of spots to measure drop pH in the mouth.

# Chapter 5
# O-pH: CLINICAL STUDY

O-pH was tested in a clinical setting to evaluate its reliability and utility as a chair side screening device [140].

## *5.1 Study Design*

The clinical study, the first optical based pH measurement of dental biofilm, was designed with pediatric patients to monitor dental biofilm pH before and after a sugar rinse for both healthy and unhealthy teeth surfaces.

### *5.1.1 Recruitment*

Pediatric patients categorized as high caries risk after clinical exam at University of Washington's Center of Pediatric Dentistry (CPD) were recruited along with a control group comprised of low caries risk patients. The inclusion criteria for the high caries risk group include at least one active lesion (cavitated or non-cavitated) either at interproximal region between maxillary posterior teeth, or at occlusal surface of mandibular posterior teeth. The inclusion criteria for the low risk control group included absence of active caries lesion or any existing restorations.

We excluded subjects undergoing active orthodontic treatment at study selected sites, having asthma, eczema, or any known allergy to yellow dyes. The high risk group is further divided into "Post-Cleaning group" and "Pre-Cleaning group" based on their recent history of professional dental cleaning. A total of 30 subjects were recruited, the "Post-Cleaning group" (n = 18) has subjects with professional dental cleaning within last three months, the "Pre-Cleaning group" (n = 7) has subjects without professional dental cleaning for

over 3 months, and lastly, a control group with subjects in low-caries risk category and a professional dental cleaning within three weeks (n = 5), see Table 5.1. Subjects were given a remuneration gift card for participating in the study and the study was approved under our institution's IRB (IRB ID: STUDY00007002).

### *5.1.2 Protocol*

The study protocol used ICDAS II ranking scheme to rank maxillary interproximal and mandibular occlusal surfaces [34], performed by a dentist at CPD using bitewing radiographs and clinical exam charting at a routine patient visit. Ranking was performed three weeks before the O-pH appointment for the Post-Clean group and within a week after the O-pH appointment for Pre-Clean group. Additionally, all teeth surfaces with no caries activity were ranked as 0 and with any carious lesion as 1, giving us a binary distinction between teeth surfaces. For every subject, we had a high number of 0 ranked tooth surfaces and only a few ranked 1. There was a minimum interval of three weeks between cleaning and pH measurements using O-pH for the Post-Clean group to allow the dental biofilm to mature.

At O-pH testing in the University's dental clinic, third- and second-year dental students (n=5) performed the pH measurements under the supervision of a dental faculty. The dental students were aware of the inclusion/exclusion criteria but blinded to group designation and surface rankings. Before the measurement, subjects were asked to rinse their oral cavity with water. Subjects were asked to produce 10 mL of saliva in a measuring cup and it's pH was measured using a conventional pH meter, followed by a baseline measurement of test surfaces (maxillary interproximal and mandibular teeth occlusal surfaces) to detect teeth AF. Next, we measured, the "rest pH" after applying Fl on the same set of teeth surfaces using a blunt hyperdermic needle one tooth at a time. Subjects then retained 10 ml of 0.3 M sucrose solution in their oral cavity for fifteen seconds. They were instructed to either swallow or spit out the sucrose solution. One minute after the sugar rinse, we measured the "drop pH" by re-applying Fl. Difference between rest pH and drop pH was calculated and called "diff pH". Application of fluorescein and pH measurement at each spot took a few

Table 5.1: Subject Statistics

| Subjects | Post-Cleaning | Pre-Cleaning | Control |
|---|---|---|---|
| Total | 18 | 7 | 5 |
| Age | 16.5 | 15 | 15 |
| Mean Cleaning Interval | 31 days | 114 days | 14 days |

seconds. At maximum, it took an additional two minutes between the measurement of first and last tooth. Each set of pH measurements (rest, drop pH) were taken with mouth open, but patients were allowed to close their mouth or speak in between measurements if it was too uncomfortable. Each measurement with O-pH at a tooth surface was repeated thrice and average of the three was used for analysis. Subjects were not provided with any prior instructions on skipping meals or to avoid brushing. Since, saliva pH is generally neutral across subjects, we used it as a stable baseline to normalize pH values across subjects. For analysis, we normalized rest and drop pH w.r.t to saliva pH and compared across different surfaces. This is an additional metric that we looked at as it takes in account impact of saliva on caries formation.

### *5.1.3 Statistical Analysis*

To measure variability in device measurement, we collected three readings per spot for rest and drop pH. Each triplet's mean and standard deviation were used to calculate the pool standard deviation of the device. This gives the average spread of all data points about their group (triplet) mean. For clinical data analysis, groups with normal distributions but unequal amount of data (pH measurements of Post vs Pre-Cleaning group) were compared using Welch's t-test [169] and permutation test [99] at 0.05 significance level. In case of groups without a normal distribution (pH of Pre/Post Cleaning group having surfaces with rank 1), only permutation test was used for significance analysis. Shapiro-Wilk's normality test

was used to test normal distribution of data distribution of data [137]. Different Groups and the statistical tests used are elaborated in the Results section. All analyses were performed using SciPy 1.7 package in Python 3.

## *5.2 Results*

### *5.2.1 Device Verification*

In the clinic, we relied on the device accuracy from *in vitro* testing and verified whether the device can take repeatable measurements. In total we measured rest pH at 85 surfaces and drop pH values at 95 surfaces, giving us a total of 180 readings. Since, each reading was measured thrice, we had a total of 540 readings. For a few measurements (<1%), we had lesser than three readings, as data points had to be discarded because of low quality or out of range pH prediction. To verify repeatability, we calculated mean and standard deviation of each rest/drop measurement triplet and then calculated pooled standard deviation. We obtained 0.23 pH of pooled standard deviation with our data, i.e., the actual readings were within 0.23 pH from the measured mean value of a triplet. Lack of clinically approved oral pH measurement devices hindered us from verifying the accuracy of the device *in vivo*.

### *5.2.2 Clinical Findings*

Assuming Pre-Clean group has higher dental biofilm level, we analyzed Pre-Clean and Post-Clean group to understand differences in pH measurements. The control group comprising caries free subjects was tested within three weeks of professional dental cleaning and lacked significant biofilm growth resulting in reduced Fl absorbance and low fluorescence emission for pH detection. The result helped us modify the clinical protocol to maintain at least a three week interval between professional cleaning and testing.

We hypothesize that lower rest and drop pH, and higher diff pH, are associated with higher level of "unhealthy" dental biofilm contributing to elevated caries risk in a certain subject. To test this hypothesis, we compared the resting, drop and difference of pH obtained

between the two recruited groups. We found Pre-Cleaning group had a lower resting and drop pH than the Post-Cleaning group. Similarly, the difference in pH was higher in Pre-cleaning group than the Post-Clean indicating higher bacterial acidification. Fig. 5.1(a), (c), (e) shows the distribution of rest pH, drop pH, and diff pH obtained in the two groups. Since, we had unequal number of data in each group, we used Welch's t-test and permutation test to measure if the pH differences between the two groups were significant. We found that drop pH was significantly lower (alpha $< 0.05$) in Pre-Cleaning compared to Post-Cleaning with $p = 0.0008$ using both tests, diff pH was significant only using permutation test ($p = 0.014$). The rest pH was lower for Pre-Clean group but we didn't find significant difference. We also compared pH between groups with the same ranking, i.e., surfaces with rank 0 in Pre-/Post-Cleaning were compared and did not find any significant difference. For subjects with rank 1, rest pH and drop pH had a significant difference with $p = 0.004$ and 0.003 respectively using permutation test (data did not have a normal distribution). Fig. 5.2(a), (b), (c), shows distribution for both ranks along with number of teeth surfaces measured.

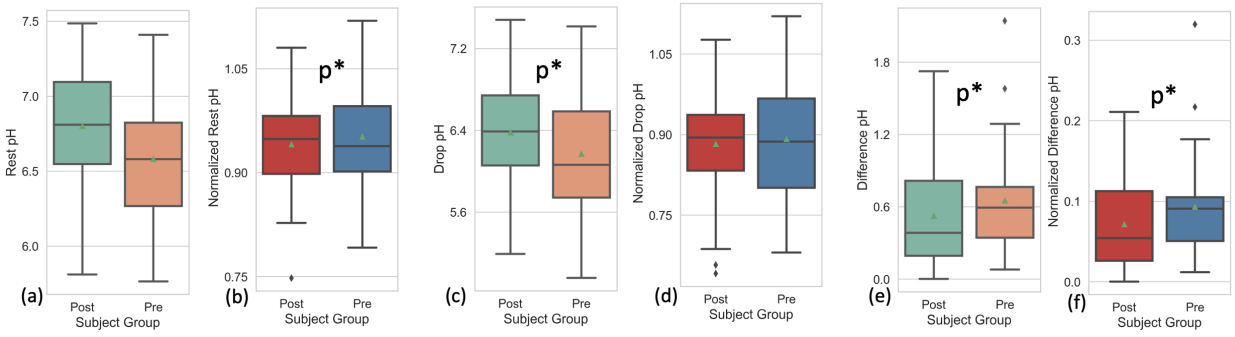


Figure 5.1: Box plots of Post and Pre Cleaning group for (a) Rest pH (b) Saliva normalized Rest pH (c) Drop pH (d) Saliva normalized Drop pH (e) Difference pH (f) Saliva normalized Difference pH with p* indicating significance with $p<0.05$

Next, comparing saliva pH between the two groups, it was observed that Pre-Clean had a lower pH than the Post-Clean group though average difference was not significant. On normalizing pH measurements with subject's saliva pH (measured before the sugar rinse),

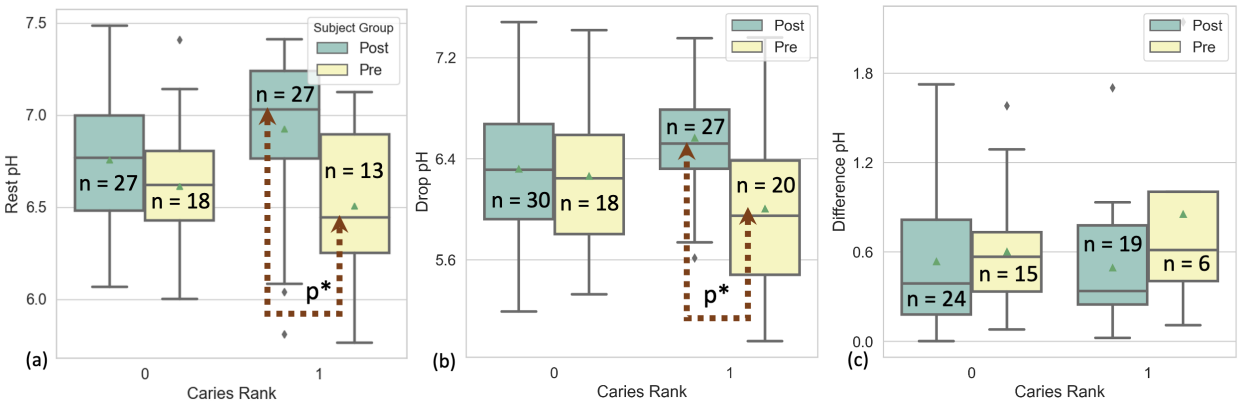


Figure 5.2: Box plot of pH measurements for different ranks per group using (a) Rest pH (b) Drop pH (c) Difference pH, with p* indicating significance with p<0.05 and n = number of teeth surfaces measured

significant difference was obtained for rest pH (Welch's t-test, p = 0.003) and diff pH (Permutation t-test, p = 0.014), see Fig.5.1(b), (d), (f). Since, the data is normalized using saliva pH, it is difficult to predict the direction of the difference unlike pH measurements in Fig.5.1(a), (c), (e) where a low "rest" or "drop" pH means higher acidity. For rank based normalized pH analysis for each group, we did not find any significant difference.

We also examined all the subjects irrespective of the cleaning group to see difference between caries and non-caries surfaces. We found average rest, drop, and diff pH for non-caries surfaces are: 6.73, 6.3, and 0.55 whereas for caries surfaces are : 6.81, 6.36, 0.56 respectively.

### *5.3 Summary of Learning & Limitations*

In terms of measuring capability, the device performed best in the Pre-Cleaning group in comparison to other groups as we measured 40 surfaces amongst 8 subjects whereas only 45 surfaces across 18 subjects in the Post-Cleaning group. Higher Fl fluorescence signal in Pre-Cleaning group along with lower AF signal assisted in obtaining repeatable measurements. We measured at least 4-5 surfaces per subject but many readings in Post-Cleaning group

were discarded because of high AF, indicating fluorescence by enamel or underlying tissues. Presence of higher AF in Post-Cleaning group vs Pre-Cleaning group could be indicative of thinner dental biofilm coverage resulting in capture of higher fluorescence from enamel. Across both groups, we noticed that surfaces to which fluorescein application was convenient, for example, upper-distal-interproximal, and lower-occlusal surfaces, had a higher signal to noise ratio. Drop pH values were more repeatable than rest pH value and perhaps the combination of sugar and fluorescein made the dye adhere to the biofilm more. Biofilm index (Quigley Hein plaque index) of teeth surfaces weren't measured but we observed that areas with low growth of biofilm had higher auto fluorescence signal. The device algorithm was found to be robust to clinical light settings. The linear fit for calibration does cause lower accuracy in lower pH range (pH 4-4.5, Table 4.1) but avoids overfitting of curve. To make device robust to noisy fluorescence, we decided to use AF as a threshold to discard pH measurements, but future versions can be built to adjust the calibration curve based on captured AF signal.

Mean rest/drop pH values of healthy/unhealthy surfaces were comparable on combining both the Post and Pre-cleaning group data. In the Pre-Cleaning group, which consists of a typical patient at a dentist's clinic for a routine recare visit, resting pH and drop pH (pH after the sugar rinse) for unhealthy surfaces (rank 1) are lower than the healthy surfaces (rank 0), though larger studies are needed to show significance. Population based standard levels of rest and drop pH could be established using clinical studies to help dentists/patients evaluate oral health quantitatively. The pH trend was opposite in Post-Cleaning group. Though this seems contrary to popular cariology concepts, prior studies have shown a wide range of variation in pH profile for unhealthy and sound enamel. P. Lingström et.al [86] measured similar rest pH and drop pH at sound and white spot regions. In another study of sound and carious (past the early caries stage) root surfaces in the same subjects yielded indistinguishable biofilm pH profiles [2]. A number of reasons could have caused the confounding results in our case, for example, it's possible that the Post-Cleaning group perhaps isn't representative of 'true enamel environment' as it consists of young dental biofilm, resulting in a pH pro-

file different from Pre-Cleaning group. Additionally, subjects in Post-Cleaning group were informed three weeks prior to the O-pH appointment about presence of unhealthy/carious surfaces. This could have prompted some of the subjects to improve their oral hygiene preventing build-up of harmful biofilm. The amount of dental biofilm in the Pre-Cleaning group is generally higher than the Post-cleaning group but it is not the amount but the composition of biofilm that plays critical role in caries formation. Unfortunately, the study didn't include microbial analysis of biofilm and we need further studies to confirm whether both young and mature biofilm at unhealthy surface has different bacterial profile or not. If the profile is indeed different, it will further strengthen the need of a pH monitoring device in clinic as it can measure 'present' biofilm activity and aid as a tool to assess oral hygiene.

The significant difference of drop and diff pH in Pre- vs Post- Cleaning group (Fig. 5.1 (c), (e)) indicates that O-pH could be used in the dental clinic as a hygiene tool to measure the growth of acid producing dental biofilm. It can also be useful as an educative tool to help patients, younger patients in particular, understand the immediate harmful impact of sugar rich diets on mouth's micro-environments and assert importance of professional dental cleaning. In comparison to Stephan's 1944 study [91], we obtained a smaller average diff pH (0.84 and 0.48 for Pre- and Post- respectively, Fig. 5.2(c)), lower than 1 pH unit for caries surfaces. The diff pH was similar to difference reported in Lingström's 2000 study [86] between sound and white spot lesions. One of the reasons could be the averaging technique, Stephan's study had categories with different caries activity and reading was averaged across all surfaces (sound and unhealthy surfaces) per category but the Lingström study looked at difference between sound and white spot surfaces and averaged only for similar surfaces, similar to analysis represented in Fig.5.2. We haven't used any subject based averaging as that reduces teeth/surface specificity. Though our study analyzed both carious and caries-free surfaces from same subjects, it lacks evaluation using contralateral surfaces in the oral cavity. Additionally, to have sufficient enrollment we did not advise subjects to skip oral routines (brushing, flossing, etc.) or increase intake of sugar. Recruiting subjects who have abstained from brushing for couple of days and sub-dividing them into groups of low and

high sugar consumption would have helped in better understanding impact of sugar as well as oral hygiene on pH.

O-pH requires moderate biofilm build up to measure pH with high signal to noise ratio as indicated from the lack of sensitive measurement in the control group. This is a device limitation that it needs medium/high biofilm deposit to measure pH and can be improved using higher excitation power and Fl concentration. Interestingly, prior studies [55, 23, 27] have had subjects skip brushing for 1-3 days to obtain Stephan curve with biofilm mass above 0.5-0.75 mg per site to have reproducible results [86]. This indicates that higher level of biofilm build up is needed to differentiate between healthy/unhealthy surfaces using acidity monitoring.

Further, as saliva pH also plays a role in caries formation [36], another metric, normalized pH measurements (biofilm pH/saliva pH), was used to understand if the trend is different for (healthy/unhealthy) surfaces and found results similar to non-normalized data. Though, saliva pH is an important factor to consider, normalized pH takes away the intuitiveness of biofilm pH as an acidity indicator.

Lack of micro-electrodes approved for intra-oral use in United States limited our study from verifying O-pH's accuracy *in vivo*. Micro-electrodes measure pH at the saliva/biofilm interface and isn't an ideal ground truth for O-pH that measures pH of extracellular oral biofilm. Microelectrodes, as previously mentioned is a contact based approach and could have caused disturbance in the biofilm impacting readings with O-pH. Therefore, our approach for verifying the O-pH performance *in vivo* was based on comparison to prior studies that used pH measurement systems in research settings, as well as demonstrating acceptable repeatability of multiple measurements from the device.

Although O-pH has the potential to be non-contact and thus nondestructive to the dental biofilm, this current spot-based pH sensing has clear drawbacks, especially in reliably testing the same spot before and after a sugar rinse. The lack of replicability in probe placement directly impacts the accuracy of pH drop measurements and has been identified as a source of variability in previous microelectrode measurements [41]. Imaging plays an important role

in mapping as dental biofilm pH is highly variable spatially and is a critical enhancement for measuring pH difference.

# Chapter 6

# O-pH IMAGING: EXTENDING BEYOND SPOT pH MEASUREMENTS

With the present spot based system it is difficult to perform trend analysis over short times for the Stephan Curve within a single visit, let alone months-long gaps in time across multiple visits. These challenges can be overcome by using an imaging system, image co-registration, and an improved clinical protocol. A non-contact optical imaging method would be ideal, but pH indicator dyes typically require full spectral analysis or multiple bio-compatible dyes which limit clinical translation. Currently, pH imaging of solid tumors are performed using pH sensitive positron emission tomography (PET) radiotracers [158], Magnetic Resonance (MR) spectroscopy [44], Magnetic Resonance Imaging (MRI) [43], and optical imaging using fluorescence [52] . These techniques are very expensive to be used for oral health screening of all patients at dental clinics.

## *6.1 Device Design and Calibration*

As a proof of concept, we modified the multi-modal Scanning Fiber Endoscope (mm-SFE) to use the two wavelength technique employed by O-pH for optical pH image-based mapping. The mmSFE scans the distal end of a single 80 micron diameter optical fiber in a spiral pattern at 10-12 KHz using a custom tubular piezoelectric actuator and a custom lens assembly [80]. The vibrating singlemode fiber emits 424 nm light (Nichia laser diode with Thor Labs Fiberport and clean up Semrock Brightline bandpass filter at 420+/-5nm) that is nearly collimated for a forward view from the mmSFE tip. By collecting backscattered reflectance (B-channel) and emitted fluorescence channels (G channel centered at 520 nm and R channel centered at 549 nm) in a ring of multimode plastic optical fibers, three spectral

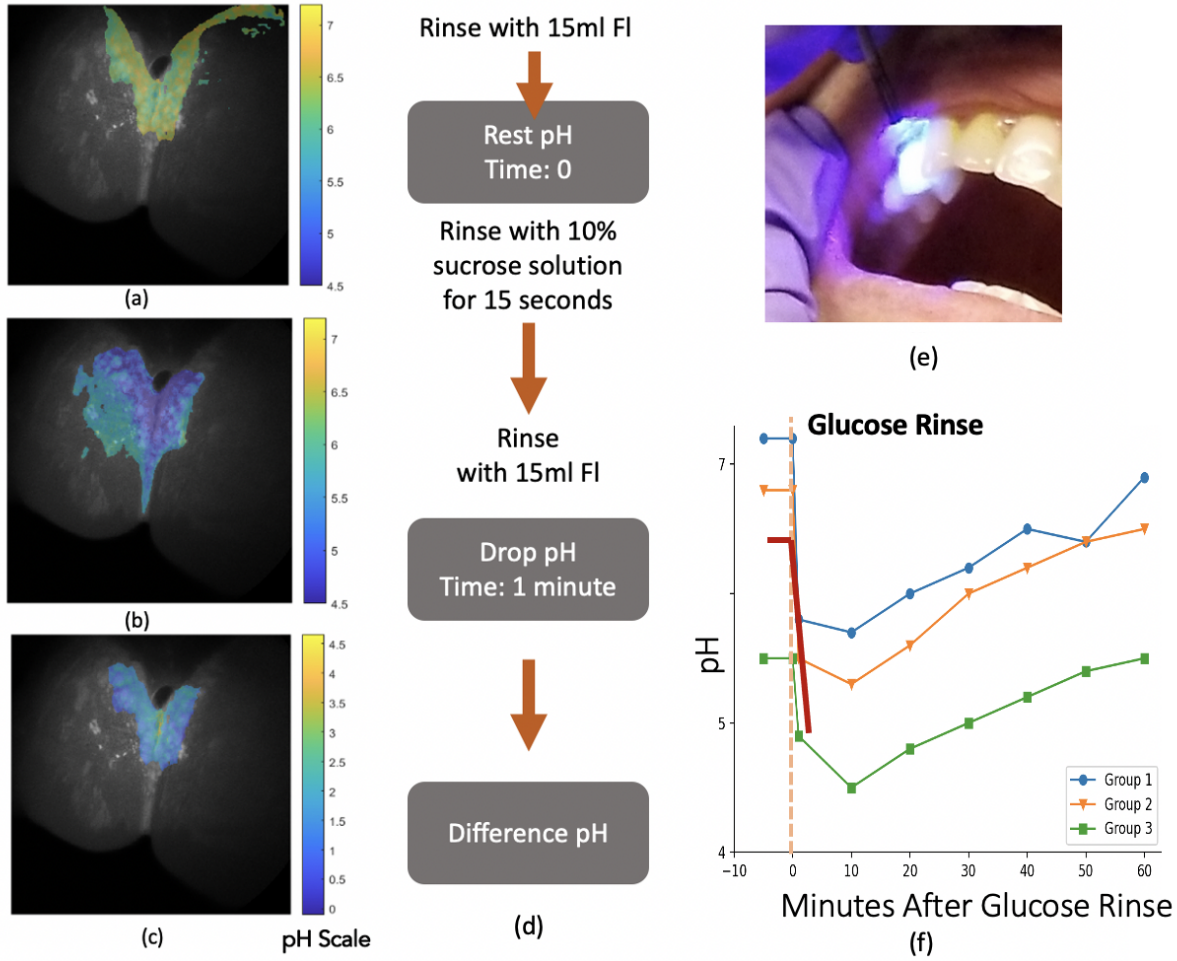


Figure 6.1: Case study with mm-SFE based pH sensing. The subject had not received professional cleaning for over seven months and had skipped brushing for 5 days prior to the examination. (a) Interproximal dental biofilm image with pH heatmap (b) pH heatmap after a sugar rinse (c) Difference between resting and drop pH (d) Protocol used for testing with mm-SFE. Fluorescein is rinsed instead of applied on each tooth surface using a blunt hyperdermic needle unlike the previous clinical study (e) mm-SFE pH probe (f) Stephan curve with red line indicating the average pH obtained using images at each stage. Group 1 to 3 are same as Fig.3.1(a).

bands of RGB are created after filtering and photomultiplier detection [63]. Similar to O-pH, we verified the imaging based device *in vitro* and built a calibration curve using the ratio, (G-R)/(G+R) w.r.t to pH. The relationship for each pH value was obtained by averaging 10 video frames acquired over 10 seconds [63].

### *6.2 Case Study*

A low-caries risk subject without a professional cleaning in last seven months was examined using the O-pH-scope after skipping brushing for 5 days. We used the modified protocol from the clinical study to enable faster measurement. Instead of applying Fl with syringe one tooth at a time, subject rinsed mouth with Fl before resting and drop pH measurement (Fig. 6.1). This study was approved under our institution's IRB (IRB ID: STUDY00002579).

The reflectance image of teeth overlaid with pH information enables tracking of regions before and after the sugar region. As shown in the images, rest pH around 6.4-7 was obtained, with 5-5.5 drop pH, and diff pH around 1.5 pH, similar to group 2 of Stephan's study (Fig. 6.1 (f)).

### *6.3 Discussion*

The mmSFE system uses highly sensitive photomultiplier optical detection which may provide sensitive pH sensing with thinner and less mature biofilms. But, the imaging system poses its own image processing challenges because enamel surfaces lack features, making it difficult to align and stitch images. Additionally, air bubbles in mmSFE images hindered accurate pH measurements in the pilot study. However, this challenge may be overcomed in the dental clinic by using compressed air to remove air bubbles. In addition, optical imaging system equipped in some dental offices can create full 3-D images of teeth thus reducing challenges in registering images taken over time. Upcoming hyperspectral cameras can be utilized instead of mm-SFE to map and measure oral pH [163].

The clinical protocol suggested can be further improved and validated in larger studies. For example, the level of 10% sucrose solution used for the O-pH and mmSFE case study

could be raised to 20% sucrose concentration which shown by Lingström's et al.[86] results in higher diff pH. In another example, several studies [85, 86, 41] have shown that at times it may take up to 5 mins to reach the lowest pH after a sugar rinse. So, monitoring the drop pH every minute for 5 minutes can perhaps give a better pH differentiation between caries and sound enamel surfaces. We avoided measuring the entire Stephan curve because it would be difficult to implement a testing protocol that lasts 60-90 minutes in routine clinical practice.

# Chapter 7

# COUGH: CRITICAL BIOMARKER FOR PULMONARY TUBERCULOSIS

Inhalation of Mtb along with incapability of immune system in preventing the growth of bacteria leads to development of active TB. Over 85%, Mtb infection originate in lungs and are known as pulmonary tuberculosis. Mtb attacks lung tissues prompting immune response and results in formation of granuloma (Fig. 7.1). If left unchecked, tissue damage continuous, resulting in bigger granulomas, causing discomfort and inflammation. These structure changes causes subjects to produce cough and can progress to release sputum and blood along with cough. Presently, wide number of screening/diagnostic tools are based on analysis of cough sputum [111].

## *7.1 Tuberculosis and Cough*

TB coughs are produced by physical damage to the lung tissue and is an early stage symptom. Historically, presence/absence of cough is used for initial TB screening but has low sensitivity [108]. Several studies have shown the importance of automatic cough counts to improve the present accuracy of self-reported coughs. Since the physical structure of the tissues changes due to infection, we hypothesize that it is not just the cough count but also the sound itself that differs in comparison to other respiratory health issues [171]. For example, coughing in asthma is caused due to allergen irritation in the air passage whereas in pneumonia air sacs are inflamed with fluid aggravating cough reflex. Since the sources of cough production changes based on disease, the frequency content of source could also vary for these coughs. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as “a characteristic that is objectively measured and evaluated as an indicator

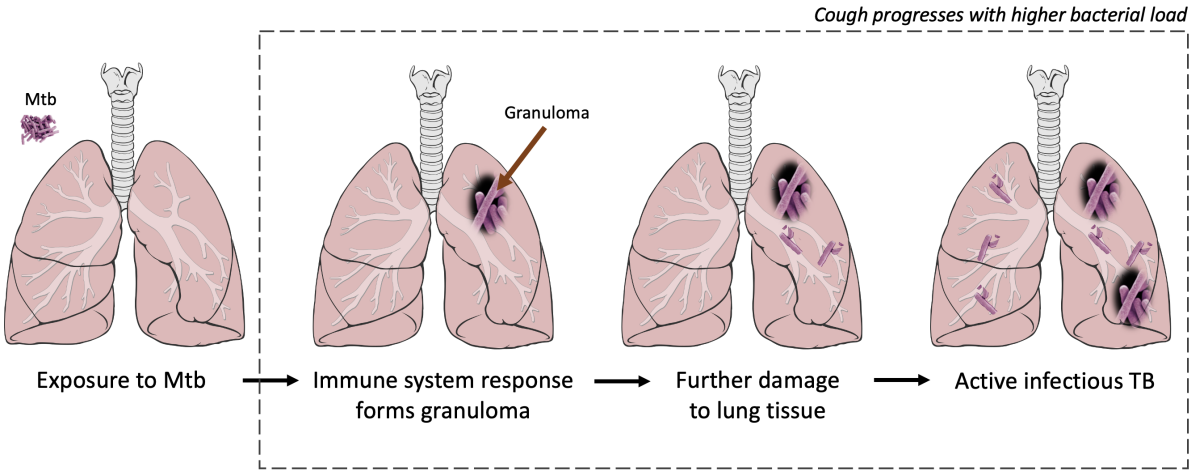


Figure 7.1: Progression of unchecked Pulmonary Tuberculosis

of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"[48]. With advances in audio processing, digital cough features has the potential to be objectively measured and is indicative of pathogenic processes showing the potential to be digital biomarker for pulmonary TB screening.

## *7.2 Digital Cough based Disease Screening*

The mechanism of cough production varies according to mucus properties, respiratory muscle strength, mechanosensitivity, and chemosensitivity of airways and other factors resulting in diverse cough sounds [26]. Some of the coughs sound "dry" or "wet" and their difference is discernable by human ears while other sound frequencies present in the cough are outside the frequency range of human ear. Cough's frequency and time domain features [88] have been studied as a bio-marker for several pulmonary diseases, like asthma [74], pneumonia [3], and TB[155]. The recent COVID-19 pandemic has pivoted the focus of machine learning community on cough analysis and several publications have explored machine learning (ML) algorithms to differentiate between COVID-19 coughs from coughs of healthy subjects.

These studies have used classical ML modeling tools like logistical regression, support vector machine [110] to modern deep learning tools like convolutional neural networks[51] and transformers [166] to achieve sensitivity between 0.65 to 0.98 and specificity between 0.69 to 0.97 [51, 110, 64, 166] in classifying COVID-19 vs non-COVID-19 coughs.

### *7.3 Digital Cough Analysis for TB Screening*

Particularly, in the field of tuberculosis cough sensing, the 2018 study by Botha et. al [15] with a total of 38 subjects (746 coughs) showed a sensitivity of 62% and specificity of 95% in classifying between TB and healthy coughs by training a logistical regression model on the voluntary cough dataset. In their 2021 publication, they extended their previous work by including unhealthy non-TB subjects (35) (coughing due to other underlying health condition) along with TB subjects (n =16) giving them a total of 1358 voluntary coughs. They achieved a sensitivity of 81% and specificity to 90% showing that there is potential in using quality of coughs in TB cough screening [109]. The authors mention that the dataset is collected at a location in a medical facility that has high environmental noise to replicate real-life cough based screening scenario. Cough classifiers need to be robust to environmental noise to be applicable in field, but unfortunately noisy dataset makes it difficult to ascertain whether the classifier model is learning the differences in the background noise or is it actually training on the features of disease of interest for classification.

In spite of a number of publications with good performance metrics for cough based disease classification, cough feature based disease screening models have not been reproducible/translatable and the reason why the model is learning to differentiate between coughs is still evasive [131, 152]. A number of issues as highlighted by researchers [51, 171], for example, (a) dataset imbalance in terms of gender, age, subjects between the coughs from control and disease group (b) difference in demographics/ environment of control and disease group, (c) variation in recording devices, (d) to inconsistencies in training methods which can lead to overestimated metrics, are applicable to TB cough classifiers as well. Additionally, there is a debate whether voluntary coughs which have been used to train the cough classifiers in

most publications are a correct representation of natural (passive) coughs since the mechanism of cough sound generation is different in two scenarios [18, 54]. Therefore, analysis using voluntary coughs adds another layer of confusion of whether or not the AI models are learning true characteristics of a disease cough. Unlike using artificial intelligence (AI) to interpret chest radiographs (x-rays) where the ML models are trying to replicate visual human findings, in a cough based AI models we still haven't located the discriminatory features between diseases (TB vs non-TB in our case). This makes it extremely important to focus not only the performance metric but also critically analyze the features/biases that can impact the model.

To overcome these challenges, it's critical to model data that has minimum background noise and environmental variability between control group (non-TB) and disease group (TB) to be certain that the model is truly learning differences in diseases rather than the ambiance noise. In this work, we have collected such a dataset and trained a binary cough classifier, TBscreen, that provides us the opportunity to answer the question- Are there discriminatory features in frequency content of TB coughs in comparison to other respiratory diseases?

# Chapter 8

# NAIROBI DATASET

## *8.1 Enrollment*

Audio cough recordings of participants with pulmonary TB and control participants having non-TB-related cough (non-TB) were collected at the Centre for Respiratory Diseases Research (CRDR), Kenya Medical Research Institute (KEMRI), Nairobi, Kenya. We recruited adult outpatients with TB from National Treatment Program clinics prior to starting anti-tuberculosis treatment. Pulmonary TB was diagnosed based on a spontaneous sputum sample that was GeneXpert (MTB/RIF or Ultra) positive, which was subsequently confirmed by AFB-culture. At the CRDR, sputum samples were decontaminated using N-acetyl-L-cysteine and sodium hydroxide and examined using fluorescence microscopy. If one or more acid-fast bacilli (AFB) per equivalent of 100 immersion fields was observed, the slide was considered positive and graded on a 0 to 3+ scale. After re-suspension with phosphate buffer, equal sample volumes were used to perform mycobacterial culture and GeneXpert MTB/RIF or Ultra (Cepheid, Sunnyvale, CA). The GeneXpert assay assigns a semiquantitative category to positive tests for M. tuberculosis based on cycle threshold (Ct) values. GeneXpert MTB/RIF categories are high, medium, low, and very low; Ultra has an additional level, trace positive, the lowest level of detection. Mycobacterial culture was performed using MGIT Manual Mycobacterial Growth System (Becton-Dickinson, Franklin Lakes, NJ). Isolates were identified as M. tuberculosis using the Capilia TB Test Kit (TAUNS, Numazu, Japan). Sputum evaluations were performed on fresh samples. Per Kenya policy, patients with TB who were not known to be HIV positive underwent HIV testing. Participants with non-TB-related cough were adult outpatients recruited from the same National Treatment Program clinics or other clinics and were GeneXpert negative, had chest X-rays not typical for TB, and were

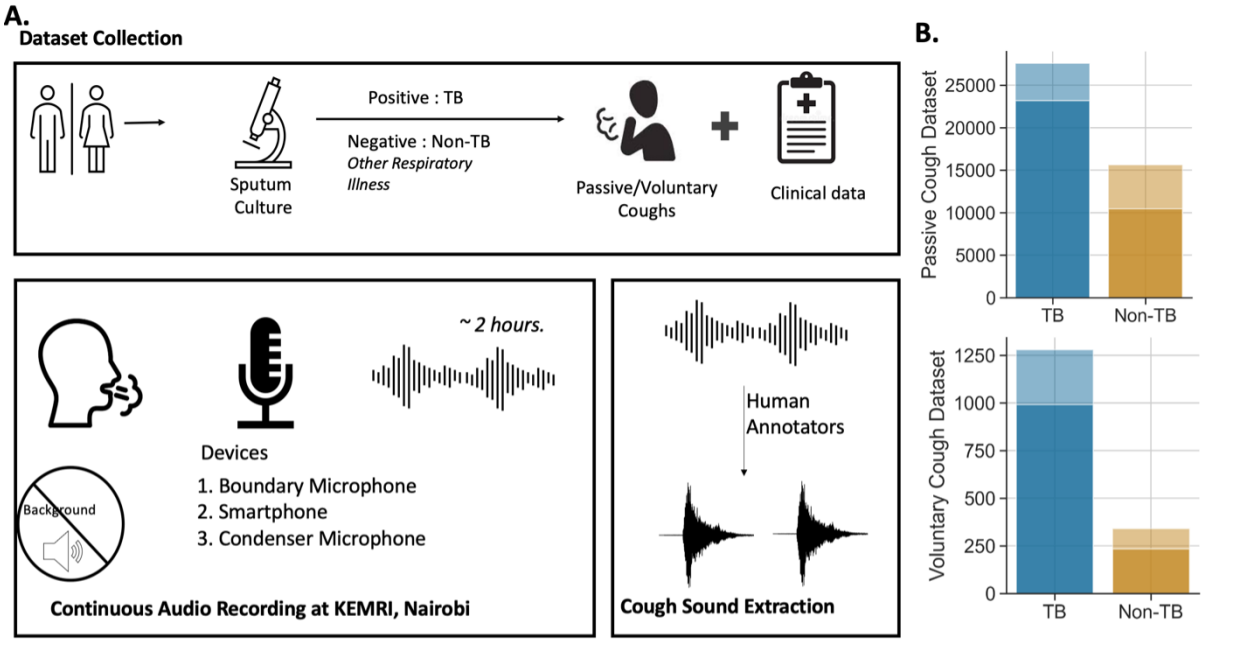


Figure 8.1: Dataset summary (a) Study protocol for the audio data collection at Kenya Medical Research Institute (KEMRI), Nairobi and subsequent cough annotation at University of Washington (UW), Seattle. Subjects with Tuberculosis (TB) and a control group of subjects having pulmonary symptoms other than Tuberculosis (non-TB) had natural cough sounds (passive coughs) recorded using three recording devices in a quiet room for two hours. A subset of the subjects provided forced coughs (voluntary coughs) at the beginning of each audio recording. These recordings were annotated using Audacity software and cough sounds with minimum background noise or distortion were selected. (b) The bar graphs represent the total passive and voluntary coughs (including all recording devices) in the Nairobi cough dataset. The lighter shade in the bar graphs indicates cough discarded due to environmental noise or audio distortion and darker shade represents the selected coughs per group. The adjacent boxplot represents distribution of total selected cough counts per subject including all three recording devices.

determined to have cough due to conditions other than TB by study clinicians. Chest X-rays were evaluated for cavitary disease by a study investigator (DJH). Various sub-categories of semi-quantitative GeneXprt and Sputum smear results were grouped into two categories – high and low. For GeneXprt, low category included trace, very low, and low results, and high category included medium and high semiquantitative readings. Similarly for sputum smear result, negative, scanty, and 1+ were classified as low and 2+, and 3+ results were categorized as high.

### *8.2 Study Protocol*

After obtaining informed consent (in English or Swahili), each participant sat in a quiet room for 2 hours with three recording devices recording continuous audio. The audio recording was annotated by humans to mark coughs in the audio file. Three audio devices; a smartphone (Google Pixel 2), a low-cost boundary microphone (Codec), and a high-end condenser microphone (Yeti) were used to record the audio at a sampling frequency of 44.1 KHz and were placed on a table in front of the subject at a fixed distance. The latter two devices were plugged into a laptop for recording. The audio recording room was selected to be in a quieter location of the hospital grounds to minimize background noise interference and participants were advised to minimize any phone conversation while the microphones were recording. Forced coughs from participants, where individuals were prompted to produce 10 coughs, were also recorded for a subset of participants. Raw audio data was uploaded to the Amazon Cloud Services S3 platform by the data collection team in Kenya. Along with audio recording, we collected demographic (age, gender, smoking history, pulmonary health history, HIV history) and clinical data (sputum analysis, chest-Xray, and blood samples) (Table 1). Study data were collected and managed using REDCap electronic data capture tools hosted at University of Washington. The study was approved by the University of Washington (STUDY00009209) and KEMRI (KEMRI/SERU/CRDR/048/3988) Institutional Review Boards. Audio files were annotated by human annotators at the University of Washington using Audacity software. Coughs with any background noise such as

fan, door, speech, or any other respiratory sounds like a sneeze or clearing of nose/throat were discarded. Additionally, cough audio files with any waveform distortion were removed from the dataset. Each cough sound was processed to have a fixed length of 1 second, and recordings greater than a second were divided into multiple audio files. Files with a length of less than 1 second were centered and padded with zeroes to make them one second long. Audio segments less than 0.1 seconds were discarded.

### *8.3 Dataset*

The passive cough dataset consists of 43,200 coughs, each one second long, from 149 subjects. After rejecting 4,390 TB and 5,169 Non-TB coughs due to background noise and clipping, the total number of passive coughs in the Nairobi Cough dataset was 33,641 (TB: 23,191 and Non-TB: 10,450) from all three recording devices and 149 subjects (TB: 103 and Non-TB: 46) The forced cough set was reduced from 1,619 to 1,225 coughs (TB: 991 (42 subjects), non-TB: 234 are (8 subjects)) after discarding 394 coughs due to clipping or background noise.

*Dataset-T1*

We built a balanced subset having an equal number of TB/non-TB subjects to (Fig. 8.2(a)) to train and evaluate the binary classifier since the number of subjects in the non-TB group is lower than the TB group (Table 8.1). The balanced dataset consists of 45 non-TB subjects (1 subject was removed due to lack of sex information) and 45 TB subjects randomly sampled from the TB dataset. The sex distribution of TB (male: 27, female: 18) and non-TB (male: 27, female: 18) subjects was identical. We limited the maximum number of coughs per subject to 225 (the average number of coughs per subject in the dataset) to avoid signatures from any one subject dominating in training or evaluation. A dataset with 21,133 coughs (10,728 TB and 10,346 non-TB) was trained and tested with 5-fold nested cross-validation.

Table 8.1: Demographic and Clinical Information of cohort

| Demographic and clinical information of cohorts | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Full Cohort (T2) | | Training & Evaluation Cohort (T1) | | Forced Cough Cohort (T3) | |
| | Category | Sub-Category | TB | Non-TB | TB | Non-TB | TB | Non-TB |
| Total Subjects | - | - | 103 | 46 | 45 | 45 | 29 | 8 |
| | Recruitment | Pulmonary TB (Cohort A) | 103 (100%) | - | 45 (100%) | - | 29 (100%) | - |
| | | Non-TB (Failed Cohort A) | - | 36 (78%) | - | 35 (77%) | - | 2 (25%) |
| | | Non-TB (Cohort C) | - | 10 (22%) | | 10 (22%) | | 6 (75%) |
| Demographic | Gender | Male | 75 (73%) | 27 (57%) | 27 (60%) | 27 (60%) | 21 (73%) | 3 (38%) |
| | Age Group | (18-40] | 69 (67%) | 23 (50%) | 33 (73%) | 23 (51%) | 21 (72%) | 5 (63%) |
| | | (40-60] | 31 (30%) | 18 (40%) | 11 (24%) | 18 (40%) | 7 (24%) | 2 (25%) |
| | | Median Age | 36 years | 40 years | 33 years | 40 years | 39 years | 32 years |
| Clinical History | HIV History | Yes | 12 (12%) | 16 (35%) | 5 (11%) | 16 (36%) | 0 | 3 (36%) |
| | Smoking History | Yes | 41 (40%) | 5 (10%) | 12 (27%) | 5 (11%) | 12 (41%) | 1 (13%) |
| | Coughing Status (Any duration) | Coughing | 95 (92%) | 41 (89%) | 41 (91%) | 41 (91%) | 26 (90%) | 6 (75%) |
| | Coughing Duration (>2 weeks) | Yes | 88 (85%) | 35 (76%) | 40 (89%) | 35 (78%) | 25 (86%) | 6 (75%) |
| | Hemoptysis | Yes | 29 (28%) | 10 (22%) | 12 (27%) | 10 (22%) | 7 (24%) | 6 (75%) |
| | Fever | Yes | 76 (74%) | 23 (50%) | 39 (87%) | 23 (51%) | 18 (62%) | 2 (25%) |
| | Weight Loss | Yes | 89 (86%) | 27 (59%) | 38 (84%) | 27 (60%) | 25 (86%) | 4 (50%) |
| | Night sweats | Yes | 77 (75%) | 17 (37%) | 35 (78%) | 17 (38%) | 18 (62%) | 3 (38%) |
| | Comorbidity | Diabetes | 3 (3%) | - | 3 (6%) | - | 0 | - |
| | | Asthma | 1 (1%) | - | 1 (2%) | - | 0 | - |
| | | Prior TB history | 15 (15%) | 11(24%) | 5(11%) | 11(24%) | 3(10%) | 2(25%) |
| | Chest x-ray findings | Cavitary disease | 74(72%) | 0 | 32(71%) | 0 | 7(24%) | 0 |
| | | Lung consolidation | 55(53%) | 2(4%) | 22(49%) | 2(4%) | 6(21%) | 1(13%) |
| | | Abnormal Lung quadrants | 98(95%) | 4(8%) | 39(87%) | 4(8%) | 11(34%) | 1(13%) |
| | | Normal | 0 | 27(57%) | 0 | 27(60%) | 0 | 4(50%) |
| TB Presentations | GeneXpert | Negative | 0 | | 0 | | 0 | |
| | | Trace | 6 (6%) | | 4 (9%) | | 4 (14%) | |
| | | Very Low | 7 (7%) | | 4 (9%) | | 1 (3%) | |
| | | Low | 19 (18%) | | 8 (18%) | | 8 (38%) | |
| | | Medium | 40 (39%) | | 14 (31%) | | 8 (28%) | |
| | | High | 30 (29%) | | 15 (33%) | | 8 (28%) | |
| | Sputum Smear | Negative | 3 (3%) | | 1 (2%) | | 0 | |
| | | Scanty | 4 (4%) | | 2 (4%) | | 1 (3%) | |
| | | 1+ | 24 (23%) | | 10 (22%) | | 4 (14%) | |
| | | 2+ | 26 (25%) | | 7 (16%) | | 8 (28%) | |
| | | 3+ | 39 (38%) | | 21 (47%) | | 11 (38%) | |

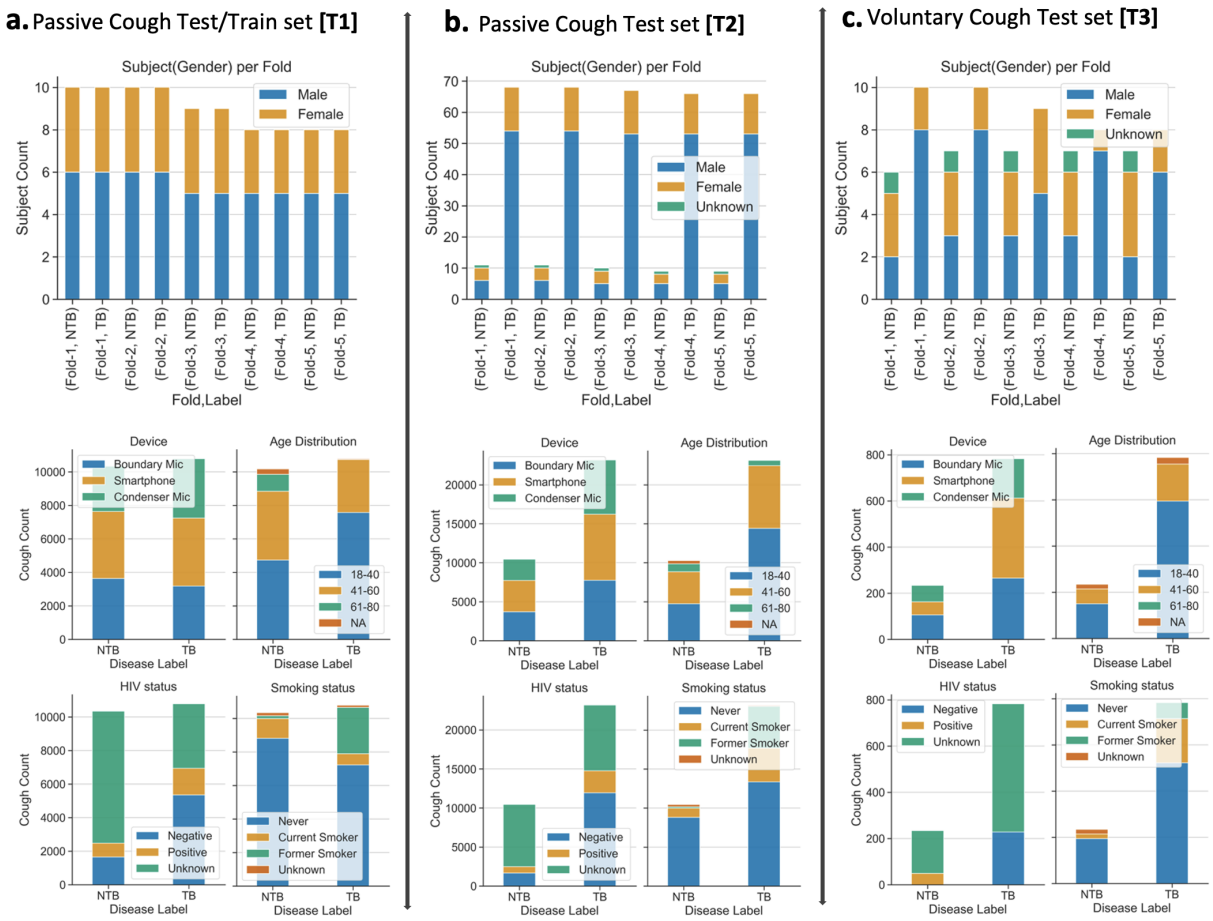


Figure 8.2: Datasets used for training and testing of passive and voluntary binary cough classifier. (a) T1 Dataset: it has 5 subject-independent (unique subjects) folds with equal number of TB (n=45) and non-TB (n=45) subjects and identical gender distribution for both the classes. Distribution of cough in T1 with respect to recording devices, age, HIV status and smoking status is summarized. (b) T2 Dataset: an unbalanced test set consisting of coughs from all TB subjects (n=103) and all non-TB subjects (N=46) coughs in the dataset. T1 folds are extended to include all non-training data in the dataset, each fold is constructed such that there is no overlap of subjects in training and testing data. Distribution of cough in T2 w.r.t to recording devices, age, HIV status and smoking status is depicted. (c) T3 Dataset: this dataset contains voluntary coughs from TB (N=29) and non-TB (N=8) subjects, each fold is constructed such that there is no overlap of subjects in training and testing data. Distribution of cough in T3 with regards to to recording devices, age, HIV status and smoking status is depicted.

*Dataset-T2*

T2 is an extension of the T1 dataset and includes all non-TB (n=1) and TB (n=58) subjects not part of T1. It provides a 5-fold unbalanced dataset with 33,641 passive coughs (TB: 23,191, non-TB:10,450). Distribution of coughs based on the recording device, age, HIV infection, and smoking status is summarized in Fig. 8.2(b). This dataset is used as a test dataset giving us performance metric for 5 folds using models rained on T1. Test folds are independent of training sets, for example, a model trained/validated on Fold 2-5 of the T1 dataset are tested using Fold 1 of T2.

*Dataset-T3*

Forced coughs were used to evaluate the passive cough model trained using T1. A test set T3 was built for each fold such that coughs (passive or forced) from the same subject were not present in the training and the testing set simultaneously.

*Dataset-Multiclass*

Three sub-datasets were built using passive coughs from the Nairobi dataset, each with 3 classes: non-TB (class-0), low-TB presentation (class-1) and high-TB presentation (class-2). Classes 1 and 2 were assigned by estimated Mtb bacillary burden (low or high for class 1 or 2, respectively)based GeneXpert semi-quantitative grade (class 1: trace, very low, and low; class 2: medium, and high), sputum smear result (class 1: negative/scanty/1+; class 2: 2+/3+), or chest X-ray findings of cavitary disease absent (class 1) or present (class 2). Therefore, 3 different models based on GeneXpert, sputum smear, and chest-Xray were trained and evaluated using 5-fold cross-validation. Each dataset is divided into 5 folds such that all the classes have equal number of subjects and similar gender distribution.

# Chapter 9

# TBSCREEN: COUGH CLASSIFIER

Nairobi Dataset was used to train and evaluate a cough classifier to differentiate between TB cough and cough from other respiratory ailment. This chapter details processing of cough features to feed as input to the model, model architecture, and training/evaluation of the cough model.

## *9.1 Cough Features*

The cough audio sampled at 44.1 KHz was trimmed/adjusted to have duration of 1 second by either padding the shorter audio file with zeros or dividing the longer cough data into multiple files. Next, to extract both time and frequency domain information from cough data, 1-D audio was converted to frequency domain (2-D) by generating scalograms with Complex Morlet Transformation. These scalograms are used for training and evaluating TBscreen classifier. TBscreen was also compared to baseline models trained using cough spectrograms. Sub-sections below describe the methodology to pre-process audio data as input to cough models.

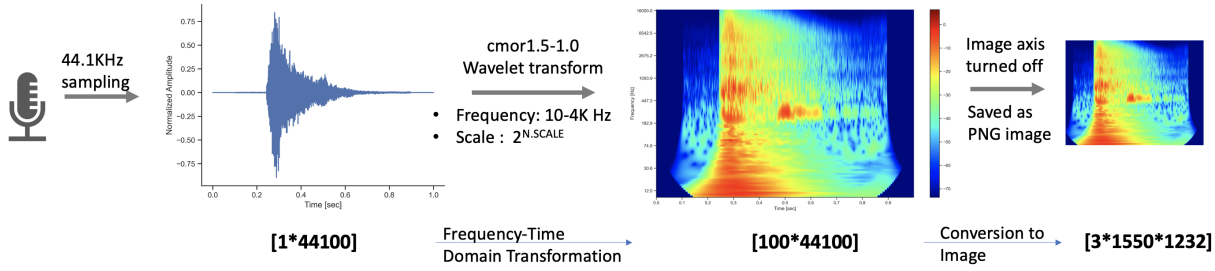


Figure 9.1: Scalogram image generation.

### *9.1.1 Scalogram*

The shape of the complex Morlet mother wavelet, a sine wave tapered by a Gaussian, resembles the shape of an audio waveform which is nearly sinusoidal, and was used for scalogram generation. Complex Morlet waveform transformation is defined by equation 9.1, a bandwidth (B) of 1.5 and center frequency (C) of 1 was selected for our application.

$$\psi(t) = \frac{1}{\sqrt{\pi B}} exp^{-t^2/B} exp^{j2\pi Ct} \tag{9.1}$$

Hundred scales of the scalogram with spacing defined by – was selected between different frequency ranges: 10-4KHz, . To enhance the features, we used log normalization of the scalograms. This generated 2-D log scalogram of size 100*44100. To reduce the dimensionality, instead of feeding the log scalogram directly, color image of the log scalogram was was stored as a PNG image of size 1150*1232*3 and fed to the model. PNG images generated using matlplotlib library with time on x-axis and scales on y-axis in increasing order was used (Fig. 9.1). To study impact of sampling rate, log scalogram images were also generated for audio files resampled to 8KHz.

### *9.1.2 Spectrogram*

For the ResNet18 baseline model, audio sample was converted to log melspectrogram using PyTorch's torchaudio transformation. In case of VGGish, we generated log melspectrogram using the technique proposed in the initial publication [56].

## *9.2 Model Architecture*

### *9.2.1 TBScreen*

*Binary Classification Model:*

Several models were evaluated as summarized with pretrained ResNet18 [53] model performing the best. ResNet18 has four residual convolutional blocks called the feature layers,

followed by an adaptive average pooling layer and a final dense layer which is called the classification layer of the model. The original model architecture was modified by replacing the classification layer (a single dense layer) with two dense layers. Dropout layers were added before each dense layer to prevent the model from overfitting, and ReLu activation was used between the two dense layers (Fig. 9.2). The input image was reduced to resized to 448x224x3 and mean and standard deviation of image dataset was used for training and evaluation.

*Multi-Class Classification Model:*

This model was similar to the binary classifier, apart from the output layer which had three classes instead of one (Non-TB, low TB load, and high TB load) and SoftMax activation.

### *9.2.2 Baseline Model*

For baseline, we used two models: (a) pretrained ResNet18 model (Fig. 9.2 (b)) , (b) pretrained VGGish [56] with log mel spectrograms as input.

## *9.3 Training and Evaluation*

We used transfer learning with ResNet18, pre-trained with millions of images using ImageNet [32] to train our TB vs non-TB binary classifier. The scalogram image was resized to 448x224x3, normalized to have values between 0 to 1, and adjusted for mean and variance using ResNet18 pre-trained model's data to train the model. A similar transformation was applied to the testing and validation sets. The scalogram image is different from the ImageNet dataset which was used to train the ResNet18 requiring to fine-tune both the feature layer along with training the classification layers of the model. We used binary cross entropy loss with Adam optimizer to train the model. The model hyperparameters - learning rate for feature/classification layers, learning rate scheduler, and batch size were tuned to adjust the model performance. The model was trained for at least 20 epochs after which training,

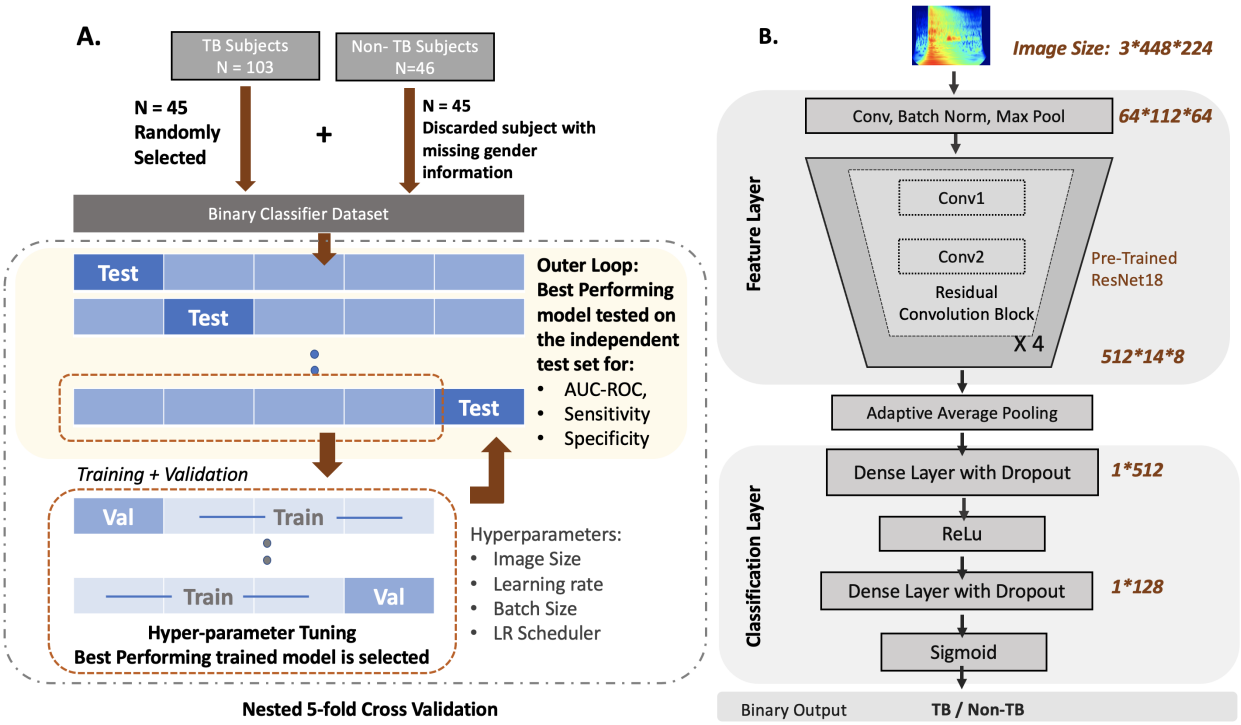


Figure 9.2: (a) 5-fold cross validation (b) ResNet18 model for TBscreen.

and validation loss was monitored to stop the model training early. The model training was stopped when the training loss didn't improve for 10 consecutive epochs, or the validation loss increased for 10 continuous epochs. The knee point of the training curve, the point at which training stabilizes was calculated and the model thereafter with the best validation accuracy was selected. The best performing binary model (highest ROC-AUC score) across different folds was trained with a learning rate of 0.000001 for feature layers, 0.00001 for the classification layer, a batch size of 32, and a scheduler that decreases the learning rate by 0.1 every 20 epochs.

In training multi-class model, we used cross entropy loss with SoftMax activation for the output layer. Performance was measured using overall accuracy and class-specific accuracy calculated from the normalized confusion matrix. Additionally, per class accuracy, sensitivity and specificity was also calculated.

For ResNet18 baseline model, the generated log melspectrogram (sampling rate: 44.1KHz, hop time = 0.01s, window length = 0.025 s) was normalized using min-max normalization to convert the amplitudes in the range of 0 to 1. In case of VGGish, we used log melspectrogram without normalization to fine tune the VGGish model (sampling rate: 16KHz, hop time = 0.01s, window length = 0.025s). These models were trained similarly using Adam optimizer with hyper-parameter tuning, and similar criteria to select the model.

All models were trained and tested using 5-fold, nested, cross-validation with the balanced dataset divided into subject independent 5 folds and having an equal number of TB and non-TB subjects. Gender distribution for TB and non-TB subjects were kept identical in each fold (Fig. 8.2. In each of the five iterations, the model was tested on one of the folds while trained and tuned using the rest of the four folds (Fig. 9.2). This was repeated 5 times so that each sub-fold acted as an independent test set giving us five sets of model metrics. In each training, three of the four training folds were used for training, and one-fold was set aside for validation to adjust the model hyperparameters. Since there were four folds in total for training and validation, we trained and validated four models and the best-performing model was selected to evaluate the independent test set. All models were trained with PyTorch using multiple GPUs part of the Hyak supercomputer system at the University of Washington. Model performance was evaluated using Receiver Operating Curve-Area Under the curve score (ROC-AUC score), sensitivity, and specificity.

### *9.4 Statistical Analysis*

The Mann-Whitney U test was used to assess statistical significance without any assumption of normality in the dataset (implemented using Python's SciPy library). Differences between ROC curves were tested for significance using DeLong's test. We assessed associations between cough frequency and TB characteristics using multivariate generalized linear regression models in which we included predictor variables (age, sex, HIV status GeneXpert grade, AFB-smear grade, chest X-ray cavitation) with p-values $<0.05$ in bivariate analyses. All statistical tests were two-sided with $\alpha = 0.05$. Analyses were performed using Stata

14 (StataCorp, College Station, TX) and R: A Language and Environment for Statistical Computing.

### *9.5 Result*

In order to examine whether cough counts or features are TB-specific, we enrolled participants with cough seeking healthcare who were diagnosed with TB (N=103) and without TB (N=46, Table 8.1) in Nairobi, Kenya. Participants with TB were recruited from National Treatment Program clinics and were GeneXpert (MTB/RIF or MTB/RIF Ultra) positive, subsequently confirmed by AFB-culture. All study interventions were performed prior to the initiation of anti-tuberculosis therapy. Participants with non-TB-related cough were recruited from the same National Treatment Program clinics or other clinics and were all GeneXpert negative, had chest X-rays not compatible with TB, and were determined to have cough due to conditions other than TB (e.g., bacterial pneumonia, viral upper respiratory infection, asthma) by study clinicians. Both groups had similar demographic and clinical characteristics with many subjects self-reporting cough for longer than two weeks, fever, and night sweats (Table 8.1). We extracted 43,200 passive (natural) coughs from these subjects (N = 149) by annotating continuous two-hour audio recordings from three devices used simultaneously (smartphone, low-cost boundary microphone, and high-cost condenser microphone), in a dedicated and relatively quiet room (Fig. 8.1). To examine the relationship of cough features between participants with TB and non-TB-related cough, we selected coughs with minimum background noise and audio distortion for analysis, bringing the total clean passive cough count to 33,641. In addition, we collected forced coughs from a subset of participants (42 TB and 8 non-TB) giving us a total of 1,619 coughs from all three recording devices (Fig. 8.1). Amongst these, 1,225 coughs with minimum background noise or audio clippings were selected for cough feature analysis. This dataset provides a unique opportunity to compare disease (TB) and control (non-TB) cough features recorded with low demographic variability and minimum ambient noise interference.

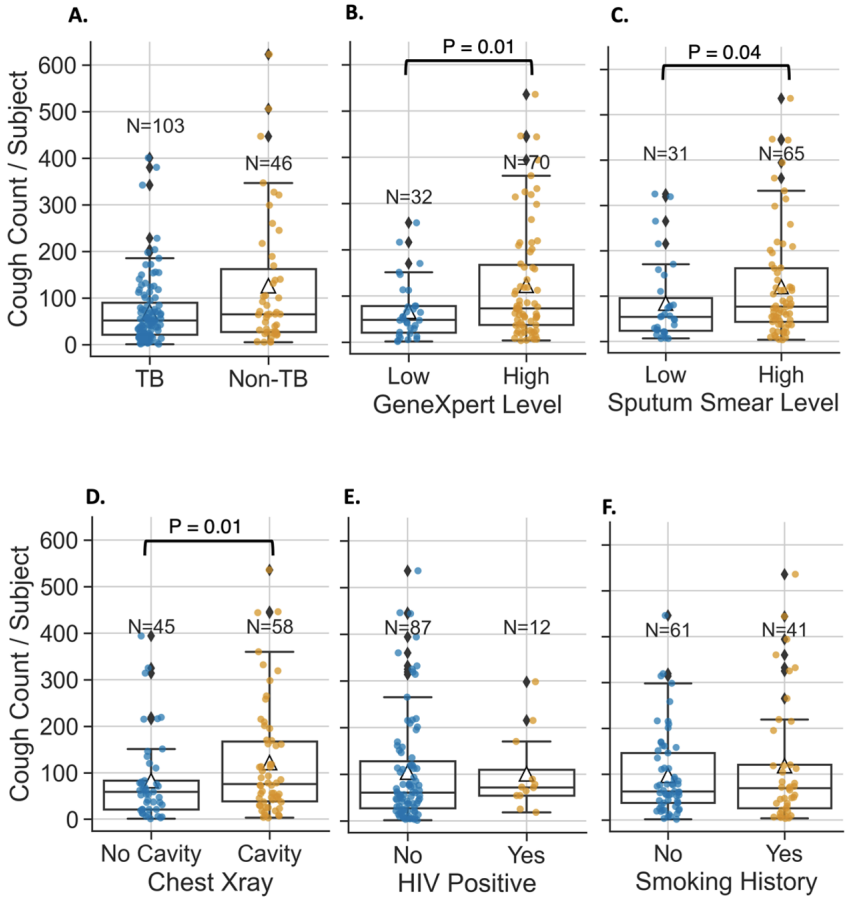


Figure 9.3: Summary of cough counts in a two-hour interval. Box plots in (a) represent cough distribution of TB vs non-TB subjects. Mean number of coughs in each box plot is depicted by a triangle. Similarly, cough distribution for various sub-categories of TB subjects is summarized with (b) low, and high PCR test result using GeneXpert; (c) low or high sputum smear score; (d) with or without cavity on chest X-ray findings; (e) with or without HIV infections; (f) with or without a history of smoking. In each graph, total number of subjects in the category is shown by N. Data for subjects with missing sub-category information is not shown. *=P<0.05 with univariate testing by Mann Whitney U statistical test.

### *9.5.1 Cough counts and TB presentation*

We first examined whether passive cough counts were associated with TB compared to those without TB related cough. Median cough counts in participants with TB were similar to those without TB (64 vs 65, coughs, P=0.64); Fig. 9.3). We also evaluated whether cough count associated with different TB presentations. Increased cough counts were associated with an increase in sputum Mtb bacterial load measured by GeneXpert (low (48 coughs), and high (73 coughs), P=0.01; Fig. 8.1(b)), and sputum AFB-smear (low (54 coughs) or high (76 coughs) sputum smear, P=0.04; Fig. 8.1c)). Participants with lung cavities on chest radiographs (76 coughs) had higher median cough counts compared to those without lung cavities (59 coughs; P=0.04, Fig. 8.1(d)). The median number of coughs were similar in TB subjects with and without HIV infection (72 vs 60 coughs, P=0.50), and with and without history of smoking (69 vs 62 coughs, P=0.93) (Fig. 8.1(e)-(f)). Using a generalized linear regression model to test associations between clinical variables and cough counts, we found that only semi-quantitative grading of GeneXpert remained significant with an increase of 22.8 counts with each additional grade (P=0.02, additional data in Appendix Table A.1). Overall, the data suggest that participants with TB and without TB have similar cough counts and that the cough counts among participants with TB increases with mycobacterial load.

### *9.5.2 Binary TBscreen performance*

We next examined whether features of cough audio, including variations in signal energy over time, and frequency, distinguished TB from non-TB-related coughs. We developed TBscreen, a ResNet18 based TB vs. non-TB classifier using RGB images of scalogram (time-frequency feature map generated using complex Morlet wavelet transform) of passive cough sounds from all three recording devices. The model was trained and tested using 5-fold cross-validation on dataset T1 (N=90: TB=45, non-TB=45; Table 8.1 and Fig. 8.2(a)). Ninety subjects were divided into 5 groups called folds, with each fold having a balanced number of unique subjects

(subject-independent folds) and identical gender distribution between the two classes (TB and non-TB, Fig. 8.2(a)). For 5-fold cross-validation, the binary classifier was first trained and validated on four of the five folds and then evaluated on the independent reserved fold, with the entire process repeated four more times (Fig. 9.2(a)). The ResNet18 classifier (Fig. 9.2(b)) generated an average receiver operating characteristic – area under the curve (ROC-AUC) of 0.79 and a standard deviation of 0.06 (sensitivity: 0.70±0.11, specificity: 0.71±0.10) across five folds on the subject balanced dataset T1 (Table 2 9.1, ROC curve for T1 with standard deviation across different folds in Fig. 9.4(a)). The five test folds of T1 dataset was expanded to form T2 (N=149: TB=103, non-TB=46; Fig. 8.2(b)) by including all non-TB (n=1) and TB (n=58) data in the Nairobi dataset that wasn't used for balanced classifier training/evaluation. The model results on dataset T2 across 5 folds had a ROC-AUC score of 0.82±0.03 (sensitivity: 0.74±0.02, specificity: 0.72±0.10, Table 9.1, ROC curve in Fig. 9.4(a)).

We next performed several secondary analyses. To understand if the model trained on passive coughs is translatable to analyze forced coughs, we evaluated the classifier's performance on a test set consisting of only forced coughs (Dataset T3, TB=29, non-TB=8, Fig. 8.2(c)). The classifier's sensitivity dropped to 0.34±0.13 while specificity increased to 0.81±0.12 with a ROC-AUC score of 0.64±0.05. This indicated that the forced cough model performed poorly and classified the majority of the cough data as non-TB (Table 9.1).

We also examined whether there were device-related performance differences. We used a subset of the cough dataset by including coughs from only one recording device to train and evaluate the model performance on the T1 and T2 subsets. The smartphone-based model performed best with 0.83±0.11 ROC-AUC for T1 subset, and 0.86±0.03 for T2 subset (Table 9.1, Table 9.3, Appendix Table A.2, Appendix Table A.3). Since our dataset contains multiple coughs from the same participant, along with evaluating performance per cough, we evaluated the model for accuracy per participant. The model on an average correctly classified TB coughs per TB participant with an accuracy of 0.68 (95% CI: (0.61, 0.75), N=103, Table **??**), and 0.78 (95% CI: (0.70, 0.86), n=45) for non-TB coughs per non-TB

subject. Together, these data demonstrate that TBscreen can discriminate between TB and non-TB cough features, but a model trained on passive cough is not translatable to predict forced cough features. Interestingly, the model based on smartphone coughs had the best performance on a reduced test set.

Table 9.1: Performance across datasets using all recording devices

**Performance across datasets:** Average ROC-AUC score, sensitivity, and specificity with standard deviation across 5-folds using different training data (coughs from all devices vs coughs from smartphone) and test sets (T1, T2, T3). Two variations of the classifier are tested on three different test sets, T1: Subject balanced passive cough dataset (for gender and number of subjects) and used for 5-fold training and testing of the classifier; T2: Expanded T1 consisting of all non-TB subjects and TB cough data not included for training the 5-fold classifier; T3: a voluntary cough dataset consisting of coughs from TB and non-TB subjects.

| | **Model Training Parameters** | **Test Set** | **ROC-AUC score** (Average of 5-folds ± S.D. across folds) | **Sensitivity** (Average of 5-folds ± S.D. across folds, threshold = 0.5) | **Specificity** (Average of 5-folds ± S.D. across folds, threshold =0.5) | **Sensitivity @70% specificity** (Average of 5-folds ± S.D. across folds) | **Combined ROC-AUC score of 5 folds** (Average after combining results from all 5-folds (De-Long's Confidence Interval)) |
|---|---|---|---|---|---|---|---|
| TBscreen | Device: All,<br>Scalogram: 10 Hz – 4 KHz,<br>Sampling rate: 44.1 KHz | T1: Subject Balanced CV | 0.79±0.06 | 0.70±0.11 | 0.71±0.10 | 0.72±0.10 | 0.80 (0.79-0.80) |
| | | T2: Expanded T1, Unbalanced set | 0.82±0.03 | 0.74±0.02 | 0.72±0.10 | 0.76±0.04 | 0.80 (0.79-0.80) |
| | | T3: Voluntary cough, Unbalanced set | 0.64±0.05 | 0.34±0.13 | 0.81±0.12 | 0.47±0.06 | 0.64 (0.62-0.66) |
| TBscreen<br>Trained/ Evaluated on coughs from Smartphone | Device: Smartphone,<br>Scalogram: 10 Hz – 4 KHz,<br>Sampling rate: 44.1 KHz | T1 subset: Subject Balanced CV | 0.83±0.11 | 0.76±0.12 | 0.74±0.10 | 0.76±0.20 | 0.85 (0.84-0.85) |
| | | T2 subset: Expanded T1, Unbalanced set | 0.86±0.03 | 0.80±0.03 | 0.74±0.10 | 0.83±0.05 | 0.86 (0.85-0.87) |
| | | T3 subset: Voluntary cough, Unbalanced set | 0.61±0.14 | 0.16±0.11 | 0.95±0.05 | 0.51±0.18 | 0.66 (0.62-0.70) |

*Impact of various parameters*

We next analyzed the impact of transient cough features on model performance by training and testing cough classifiers (dataset T1) using scalogram features generated from different frequency ranges (10 Hz-4 KHz, 4 KHz-8 KHz, 10 Hz-16 KHz) and sampling rates. The model performed best in the frequency range of 10 Hz-4 KHz (sensitivity: 0.70±0.11, specificity: 0.71±0.10) and had lower sensitivity and specificity in frequency ranges above 4KHz

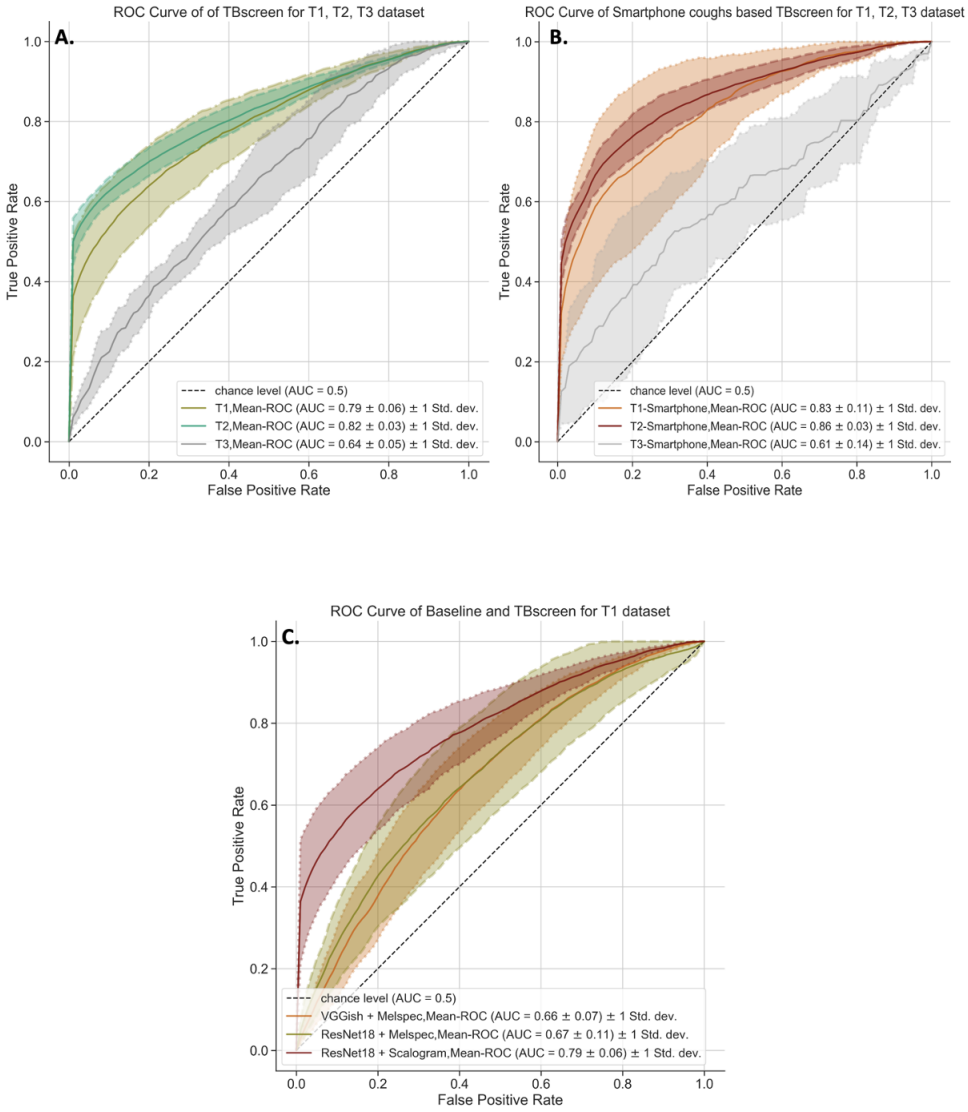


Figure 9.4: Passive Binary Cough ROC plot. (a) ROC-curve with standard deviation across 5-folds for model trained using coughs from all devices and evaluated on T1: subject balanced passive cough dataset (w.r.t. to gender and number of subjects) and used for 5-fold training and testing of the classifier; T2: expanded T1 consisting of all non-TB subjects and TB cough data not included for training the 5-fold classifier; T3: a voluntary cough dataset consisting of coughs from TB and non-TB subjects. ROC curve with standard deviation for s second model trained on and validated on coughs from smartphone is also represented (b) ROC-curve with standard deviation across 5-folds for model trained using coughs from smartphone and evaluated on T1, T2, T3 (c) Comparison of ROC curve of the binary cough classifier trained using scalogram images of cough and baseline cough models trained on mel-spectrogram features.

Table 9.2: Performance across datasets using particular recording device

| | **Model Training Parameters** | **Test Set** | **ROC-AUC score** (Average of 5-folds ± S.D. across folds) | **Sensitivity** (Average of 5-folds ± S.D. across folds) | **Specificity** (Average of 5-folds ± S.D. across folds) | **Sensitivity @70% specificity** (Average of 5-folds ± S.D. across folds) | **Combined ROC-AUC score of 5 folds** (Average of 5-folds with Confidence Interval) |
|---|---|---|---|---|---|---|---|
| TBscreen<br><br>Trained/Evaluated on coughs from Smartphone | Device: Smartphone,<br><br>Scalogram: 10 Hz – 4 KHz, Sampling rate: 44.1 KHz | T1 subset: Subject Balanced CV | 0.83±0.11 | 0.76±0.12 | 0.74±0.10 | 0.76±0.20 | 0.85 (0.84-0.85) |
| | | T2 subset: Expanded T1, Unbalanced set | 0.86±0.03 | 0.80±0.03 | 0.74±0.10 | 0.83±0.05 | 0.86 (0.85-0.87) |
| | | T3 subset: Voluntary cough, Unbalanced set | 0.61±0.14 | 0.16±0.11 | 0.95±0.05 | 0.51±0.18 | 0.66 (0.62-0.70) |
| TBscreen<br><br>Trained/Evaluated on coughs from Boundary Microphone | Device: Boundary Microphone,<br><br>Scalogram: 10 Hz – 4 KHz, Sampling rate: 44.1 KHz | T1 subset: Subject Balanced CV | 0.77±0.10 | 0.69±0.09 | 0.67±0.20 | 0.69±0.13 | 0.78 (0.77-0.78) |
| | | T2 subset: Expanded T1, Unbalanced set | 0.81±0.06 | 0.73±0.04 | 0.69±0.18 | 0.73±0.07 | 0.79 (0.78-0.80) |
| | | T3 subset: Voluntary cough, Unbalanced set | 0.61±0.08 | 0.44±0.13 | 0.68±0.11 | 0.47±0.13 | 0.62 (0.59-0.66) |
| TBscreen<br><br>Trained/Evaluated on coughs from Condenser Microphone | Device: Condenser Microphone,<br><br>Scalogram: 10 Hz – 4 KHz, Sampling rate: 44.1 KHz | T1 subset: Subject Balanced CV | 0.73±0.14 | 0.65±0.19 | 0.65±0.13 | 0.62±0.22 | 0.73 (0.72-0.74) |
| | | T2 subset: Expanded T1, Unbalanced set | 0.80±0.04 | 0.75±0.06 | 0.65±0.12 | 0.73±0.06 | 0.74 (0.73-0.75) |
| | | T3 subset: Voluntary cough, Unbalanced set | 0.69±0.07 | 0.46±0.19 | 0.74±0.14 | 0.60±0.10 | 0.66 (0.62-0.70) |

(0.64±0.19, 0.68±0.08, $p < 0.001$, Table 9.3,Appendix Table A.3,Appendix Table A.4). Next, we analyzed features using different sampling rates (rate at which audio data is sampled by the device: 8 kHz, 44.1 KHz) to verify its impact on the model performance. We observe that the model performance degrades with a lower sampling rate ($p < 0.001$, 9.3,Appendix Table A.2,Appendix Table A.3). The binary cough model performs best in the frequency range of 10 Hz – 4 KHz with a sampling rate of 44.1 KHz.

Further, we assessed the influence of recording devices by training and evaluating device specific cough classification models. The cough model was trained and tested on each fold in T1 using data only from a specific recording device (T1 subset), scalogram features in frequency range of 10 Hz - 4 KHz, and audio sampling rate of 44.1KHz. Overall, the model trained with smartphone data performed best with an average AUC-ROC score of 0.83±0.1, sensitivity of 0.76±0.12 and specificity of 0.74±0.1, followed by boundary microphone (sensitivity: 0.69±0.09, specificity: 0.67±0.20, $p < 0.001$), and then condenser microphone (sensitivity: 0.65±0.19, specificity: 0.65±0.13, $p < 0.001$) (Fig. 9.4(b), Table 9.2, 9.3, Appendix Table A.2, Appendix Table A.3 ). On evaluating device specific models on complete T1 test set (with all the recording device) we found a drop in specificity across all device models. Results indicate that smartphone-based model performs best in comparison to the other two recording devices and a model trained on one device has lower performance while evaluating cough data from multiple recording devices.

*Performance bias of smartphone based cough model*

We further examined the best performing model (Smartphone cough-based model, frequency range: 10 HZ-4 KHz, and audio sampling rate: 44.1 KHz) within different demographic/clinical sub-categories like gender, age, smoking status, HIV infection, and different presentations of TB subjects (Table 9.4 and Appendix Table A.4)) using T1 subset (only coughs recorded with smartphone). The model demonstrated better performance with male coughs (ROC-AUC: 0.87+0.15) over female coughs (ROC-AUC:0.78±0.12, $p < 0.001$). The model performed better for older age group (ROC-AUC in 18-40: 0.80±0.15, 40-60:

Table 9.3: Various cough features and model performance

**Various cough features and model performance:** Average ROC-AUC score, sensitivity, and specificity with standard deviation across 5-folds using different training inputs and performance evaluated on T1 dataset.

| | **Model Training Parameters** | **Test Set** | **ROC-AUC score** (Average of 5-folds ± S.D. across folds) | **Sensitivity** (Average of 5-folds ± S.D. across folds, threshold = 0.5) | **Specificity** (Average of 5-folds ± S.D. across folds, threshold =0.5) | **Sensitivity @70% specificity** (Average of 5-folds ± S.D. across folds) | **Combined ROC-AUC score of 5 folds** (Average after combining results from all 5-folds (De-Long's Confidence Interval)) |
|---|---|---|---|---|---|---|---|
| Frequency range of scalogram<br><br>Device: All, Sampling rate: 44.1 KHz | 10 Hz – 4 KHz | T1: Subject Balanced CV | 0.79±0.06 | 0.70±0.11 | 0.71±0.10 | 0.72±0.10 | 0.80 (0.79-0.80) |
| | 4 KHz – 8 KHz | T1: Subject Balanced CV | 0.75±0.09 | 0.64±0.19 | 0.68±0.08 | 0.64±0.14 | 0.75 (0.75-0.76) |
| | 10 KHz – 16 KHz | T1: Subject Balanced CV | 0.79±0.07 | 0.72±0.07 | 0.69±0.10 | 0.71±0.12 | 0.80 (0.79-0.80) |
| Sampling rate<br><br>Device: All | 8 KHz Scalogram: 10 Hz – 4KHz | T1: Subject Balanced CV | 0.65±0.09 | 0.64±0.18 | 0.57±0.1 | 0.51±0.14 | 0.63 (0.62-0.63) |
| | 44.1 KHz Scalogram: 10 Hz – 4KHz | T1: Subject Balanced CV | 0.79±0.06 | 0.70±0.11 | 0.71±0.10 | 0.72±0.10 | 0.80 (0.79-0.80) |
| Recording device<br><br>Scalogram: 10 Hz - 4KHz, Sampling rate: 44.1 KHz | Smartphone | T1 subset: Subject Balanced CV (Smartphone coughs) | 0.83±0.11 | 0.76±0.12 | 0.74±0.10 | 0.76±0.20 | 0.85 (0.84-0.85) |
| | | T1: Subject Balanced CV | 0.79±0.03 | 0.76±0.05 | 0.62±0.04 | 0.72±0.06 | 0.80 (0.79-0.80) |
| | Boundary Mic. | T1 subset: Subject Balanced CV (Boundary Microphone coughs) | 0.77±0.10 | 0.69±0.09 | 0.67±0.20 | 0.69±0.13 | 0.78 (0.77-0.78) |
| | | T1: Subject Balanced CV | 0.77±0.09 | 0.72±0.12 | 0.66±0.14 | 0.68±0.14 | 0.79 (0.79-0.80) |
| | Condenser Mic. | T1 subset: Subject Balanced CV (Condenser Microphone coughs) | 0.73±0.14 | 0.65±0.19 | 0.65±0.13 | 0.62±0.22 | 0.73 (0.72-0.74) |
| | | T1: Subject Balanced CV | 0.69±0.17 | 0.67±0.18 | 0.53±0.18 | 0.56±0.25 | 0.69 (0.68-0.69) |
| Alternative Frequency Representation<br><br>Mel-spectrogram Device: All Baseline Model | Model: VGGish<br><br>Sampling rate: 16 KHz | T1: Subject Balanced CV | 0.66±0.07 | 0.62±0.19 | 0.61±0.09 | 0.53±0.12 | 0.63 (0.63-0.64) |
| | Model: ResNet18<br><br>Sampling rate: 44.1 KHz | T1: Subject Balanced CV | 0.67±0.11 | 0.66±0.16 | 0.58±0.10 | 0.55±0.16 | 0.65 (0.65-0.66) |

0.87±0.11, $p = 0.01$). The model performance was also better for PLHIV subjects versus subjects with no HIV infection ($p < 0.001$). Subjects with a smoking history had higher ROC-AUC score (0.87±0.17, $p < 0.001$) over subjects with no smoking history (0.80±0.10). The ROC-AUC score of the model in differentiating between TB and non-TB coughs was higher for TB subjects with high GeneXpert semi-quantitative grade (0.86±0.12, $p < 0.001$) versus TB subjects with low grade (0.69±0.12). The model performed better in TB subjects with a lung cavity (0.86±0.12, $p < 0.001$) versus no lung cavity (0.71±0.05). Overall, the smartphone-based cough model had better performance with male subjects and TB subjects having a higher GeneXpert semi-quantitative grade or a cavitary chest x-ray and was unaffected by age.

### *9.5.3 Baseline Model performance*

We compared VGGish and ResNet18 models trained on mel-spectrogram (alternative audio frequency representations) as a baseline against the scalogram model. The mel-spectrogram feature set was able to classify TB/non-TB coughs (AUC-ROC score: 0.61±0.06/0.62±0.08, VGGish/ResNet18), but had lower sensitivity (0.62±0.19/0.66±0.16, VGGish/ResNet18), and specificity (0.61±0.09/0.58±0.10, VGGish/ResNet18) in comparison to TBscreen (AUC-ROC score: 0.79+0.06, $p < 0.001$) built using scalogram features (Table 9.3, Fig. 9.4(c), and Appendix Table A.3). Together, these subgroup analyses demonstrated that scalogram cough features generated in a frequency range of 10 Hz-4 KHz from cough audio recorded using smartphone at a sampling rate of 44.1 KHz had the best performance in differentiating between TB and non-TB coughs.

### *9.5.4 Multi-class TBscreen performance*

We extended the binary classification model (TB vs non-TB) to a multiclass classifier to examine whether cough features can differentiate between non-TB and various clinical presentations of TB. The model included three distinct classes- non-TB (class-0), low-TB burden presentation (class-1) and high-TB burden presentation (class-2). The level of TB burden

Table 9.4: Smartphone based model performance

| **Smartphone classification and performance bias analysis of binary cough model:** The classification results are represented for different demographic, clinical and TB presentations in T1 dataset. | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Category** | **Sub-Category** | **ROC-AUC score** (Average of 5-folds ± S.D. across folds) | **Sensitivity** (Average of 5-folds ± S.D. across folds, threshold = 0.5) | **Specificity** (Average of 5-folds ± S.D. across folds, threshold =0.5) | **Sensitivity @70% specificity** (Average of 5-folds ± S.D. across folds) | **Combined ROC-AUC score of 5 folds** (Average after combining results from all 5-folds (De-Long's Confidence Interval)) |
| Overall Model<br><br>Device: Smartphone, Scalogram: 10 Hz – 4KHz, Sampling rate: 44.1 KHz | All inclusive | - | 0.83±0.11 | 0.76±0.12 | 0.74±0.10 | 0.76±0.20 | 0.85 (0.84-0.85) |
| Demographic Bias | Gender | Male | 0.87±0.15 | 0.85±0.14 | 0.72±0.13 | 0.84±0.25 | 0.89 (0.89-0.90) |
| | | Female | 0.78±0.12 | 0.56±0.13 | 0.81±0.14 | 0.70±0.23 | 0.76 (0.74-0.78) |
| | Age Group | [18,40] | 0.80±0.15 | 0.74±0.16 | 0.71±0.14 | 0.70±0.25 | 0.85 (0.84-0.86) |
| | | (40,60] | 0.87±0.11 | 0.81±0.12 | 0.78±0.16 | 0.86±0.16 | 0.83 (0.81-0.84) |
| Clinical Bias | HIV History | No HIV history | 0.82±0.10 | 0.74±0.13 | 0.72±0.13 | 0.77±0.17 | 0.84 (0.83-0.85) |
| | | With HIV history | 0.92±0.08 | 0.89±0.07 | 0.81±0.17 | 0.90±0.10 | 0.91 (0.90-0.93) |
| | Smoking History | No Smoking History | 0.80±0.10 | 0.71±0.11 | 0.73±0.10 | 0.72±0.18 | 0.81 (0.80-0.83) |
| | | With Smoking History | 0.87±0.17 | 0.83±0.15 | 0.81±0.18 | 0.83±0.27 | 0.90 (0.89-0.91) |
| TB Presentations | GeneXpert | Low Bacterial Load | 0.69±0.12 | 0.52±0.22 | 0.74±0.1 | 0.58±0.24 | 0.74 (0.73-0.76) |
| | | High Bacterial Load | 0.86±0.12 | 0.82±0.13 | 0.74±0.1 | 0.82±0.21 | 0.87 (0.87-0.88) |
| | Sputum Smear | Low Bacterial Load | 0.84±0.15 | 0.76±0.21 | 0.74±0.1 | 0.78±0.23 | 0.85 (0.83-0.87) |
| | | High Bacterial Load | 0.85±0.11 | 0.79±0.12 | 0.74±0.1 | 0.80±0.21 | 0.86 (0.85-0.87) |
| | Lung Cavity | No Cavity | 0.71±0.05 | 0.52±0.1 | 0.74±0.1 | 0.56±0.13 | 0.71 (0.69-0.72) |
| | | With Cavity | 0.86±0.12 | 0.83±0.12 | 0.74±0.1 | 0.83±0.21 | 0.89 (0.88-0.89) |

Table 9.5: Subject wise smartphone based model performance

| **Subject Category**<br><br>Model: Smartphone coughs based, 10-4KHz, 44.1KHz | **No. of Subject** | **Mean accuracy (per subject)** | **95 % Confidence Interval** | **Mean accuracy (per cough)** |
|---|---|---|---|---|
| TB, Dataset T1 | 45 | 0.66 | (0.56, 0.76) | 0.76±0.12 |
| TB, Dataset T2 | 103 | 0.68 | (0.61, 0.75) | 0.76±0.12 |
| Non-TB, Dataset T1 | 45 | 0.78 | (0.70, 0.86) | 0.74±0.10 |

was either based on Mtb bacillary load estimated using GeneXpert or sputum smear result (low vs high bacterial load); or on the presence (high)/absence (low) of lung cavities. Therefore, 3 different models based on GeneXpert, sputum smear, and chest-Xray were trained and evaluated using 5-fold cross-validation. Each fold had coughs from unique participants (subject-independent), balanced number of participants, and equal gender distribution across three classes (Fig. 9.5). The accuracy score for a model based on GeneXpert was 0.40±0.09 (n = 90; Class 0/1/2: 30), sputum smear 0.45±0.05 (n = 105; Class 0/1/2: 35), and chest x-ray 0.44±0.03 (n = 117; Class 0/1/2: 39). All three models had the highest sensitivity for class 0 (GeneXpert: 0.58±0.23, sputum smear: 0.58±0.05, chest x-ray: 0.62±0.07) and lowest sensitivity for class 1 (GeneXpert: 0.21±0.08, sputum smear: 0.33±0.15, chest x-ray: 0.23±0.08). In summary, the chest x-ray based multi-class model had better sensitivity for all three classes in comparison to the other two models but overall, the accuracy and sensitivity of all three models were sub-optimal due to lower performance in distinguishing between

class 1 and class 2 (Fig. 9.5, Table. 9.6).

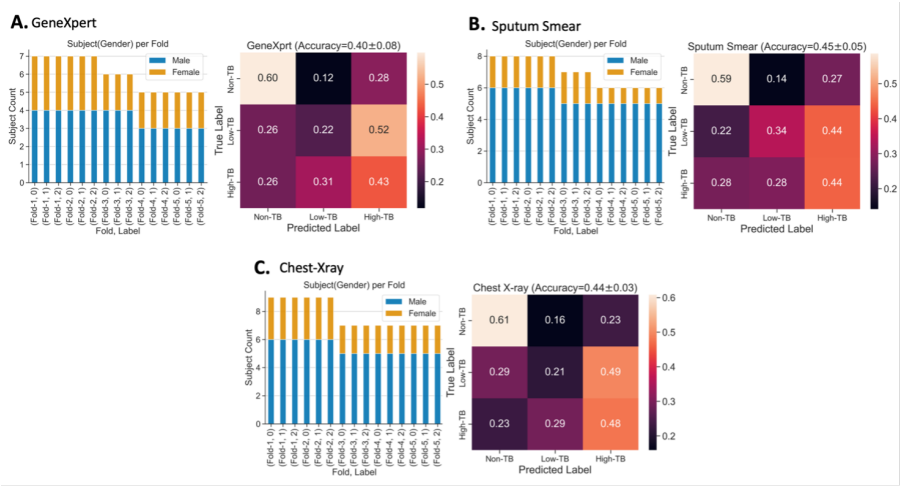


Figure 9.5: Dataset and performance for multi-class classifier. Model using (a) GeneXpert levels, (b) sputum smear result, or (c) chest X-ray. Multi-Class normalized confusion matrix for four different types of classification is presented along with the subject/gender distribution in the 5-fold cross validation dataset. The confusion matrix summarizes classification results from all five folds.

## *9.6 Summary of Learning & Limitations*

We investigated whether cough characteristics discriminate between TB and non-TB-related coughs. Although cough counts did not discriminate between cough related to TB versus other conditions, cough scalogram characteristics were associated with identification of coughs due to pulmonary TB. Our initial ResNet18 classifier model distinguished TB vs non-TB coughs with ROC curve value 0.79±0.06 using scalogram features (signal energy vs time and frequency) generated in the frequency range of 10-4 KHz and sampling rate of 44.1 KHz using a balanced dataset. We found that the best performing model was based on recordings from a Pixel smartphone (ROC curve 0.83±0.10). Further improvements in accuracy of the smartphone-based model were noted in detecting participants with a high GeneXpert semi-quantitative grade, who are likely most infectious, compared to those with

Table 9.6: Multi-class model of TB presentation. Performance metrics of models based on GeneXpert, Sputum smear and Chest X-ray.

| TB Presentation | Sub-Category | No. of subjects | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|
| GeneXpert | Overall | 90 | 0.40±0.09 | - | - |
| | Class 0: Non-TB | 30 | 0.69±0.05 | 0.58±0.23 | 0.74±0.11 |
| | Class 1: GeneXpert 1-3 | 30 | 0.60±0.06 | 0.21±0.08 | 0.77±0.08 |
| | Class 2: GeneXpert 4-5 | 30 | 0.51±0.11 | 0.42±0.17 | 0.57±0.15 |
| Sputum Smear | Overall | 117 | 0.45±0.05 | - | - |
| | Class 0: Non-TB | 39 | 0.70+0.10 | 0.58+0.05 | 0.76+0.13 |
| | Class 1: Sputum smear 0-3 | 39 | 0.64+0.44 | 0.33+0.15 | 0.78+0.07 |
| | Class 2: Sputum smear 4-5 | 39 | 0.57+0.04 | 0.45+0.13 | 0.64+0.03 |
| Chest X-ray | Overall | 105 | 0.44±0.03 | - | - |
| | Class 0: Non-TB | 35 | 0.70±0.06 | 0.62±0.07 | 0.74±0.05 |
| | Class 1: TB, without cavity | 35 | 0.60±0.05 | 0.23±0.08 | 0.78±0.06 |
| | Class 2: TB, with cavity | 35 | 0.58±0.04 | 0.49±0.11 | 0.65±0.11 |

non-TB-related cough (ROC-AUC 0.86, 95% CI 0.87-0.88). This model achieves a sensitivity of 82% (±21%) at 70% specificity, a level that approaches the WHO TPP for a TB triage test of 90% sensitivity and 70% specificity.

While several recent studies have shown similar or greater accuracy in TB cough classifiers [15, 109, 115], we believe that several strengths of the present study elevate the validity of our findings and support the potential of cough classifiers as TB triage tests. All three prior studies of machine language algorithms for TB cough-based detection recorded forced

(voluntary) coughs. In our study we found that forced coughs performed poorly when applied to a model trained on passive coughs, highlighting differences in these cough types. Two of the prior studies enrolled healthy volunteers as the control group [15, 115], which would not be reflective of the expected clinical role (persons suspected of having pulmonary TB) for cough screening algorithms. Other important limitations to the prior studies include gender imbalance between classes and the presence of significant ambient noise in the dataset [109]. Additional strengths of our study include rigorous evaluations to exclude TB in non-TB-related controls, cough recordings obtained prior to TB treatment initiation, strict recording conditions and standardized lengths of recordings (2 hours), and training/evaluation using a gender balanced dataset (T1).

We assessed both scalogram and mel-spectrogram cough representation methods due to a lack of consensus in the field on the optimal methodology. Scalograms generated using continuous wavelet transform provide better frequency vs time resolution compared to other frequency domain transformations [30, 40, 153]. This approach has been used in the analysis of other 1D data like electroencephalogram [10], DNA analysis [100], lung auscultation [143]. The resolution of scalograms come at the cost of increased feature size, especially for audio data, limiting its applications. To reduce the size of generated scalograms, we used colored (RGB) images as an encoding for scalograms to train TBscreen [129, 95] making it easier to use medium size deep learning models for training. We could then leverage pretrained image classification models reducing the need for a very large training dataset. Traditional cough-based disease classifiers have been developed using mel-spectrogram [110, 51, 15, 115] by transforming Short Time Fourier Transform (STFT)-spectrograms with mel-filter banks. These filter banks mimic frequency sensitivity of a human ear as well reduces dimension of an STFT based spectrogram [150]. It is difficult for clinicians to screen patients by listening to their cough sounds which makes mel-spectrograms (based on response to human ears) as suboptimal approach. Our results confirm that a model using scalogram features performs better than mel-spectrogram features in TB disease classification. Additionally, impact of various factors (demographic, recording devices, cough features, and TB presentation) on

model performance was assessed to provide better transparency of the TB cough model.

Models trained using different recording devices show a varied response with the smartphone-based model having the best performance. The condenser microphone was overly sensitive resulting in the capture of more ambient sound and a smaller training dataset due to audio distortion. The boundary microphone is a low-cost microphone which might not have captured the cough sound adequately, though it's built-in algorithm to reduce ambient sound could have positively impacted the performance. Audio recording devices that take in feedback to adjust its gain based on an incoming cough's volume are ideal for recording coughs for analysis. Narrowing the frequency of interest would help mitigate issues of designing a universal cough model as it can enable development of tools that are sensitive to audio frequencies only in the range of interest.

Our study has several limitations. We found the model performed better with males and those with a higher bacillary load and the presence of cavities. Overall, number of female subjects and coughs from female participants (n = 36, coughs = 6,559 (31%)) were lower than male subjects in the dataset (n = 54, coughs = 14,574 (69%)) which could cause the difference in performance for the two genders. In addition, the improved performance in subjects with a higher bacterial load or chest cavity could be attributed to more disease signal in coughs of subjects with advanced TB. The overall study results are limited in terms of number of subjects and needs evaluation on a larger independent dataset. Increasing the training data can possibly improve the present sensitivity which is currently lower than WHO requirement's (Sensitivity: 90%, Specificity: 70%) for a TB screening tool [108]. Additionally, TBscreen was trained on a controlled dataset and is not optimized for time efficiency or robustness to any ambient noise essential for real-world deployment.

# Chapter 10

# PASSIVE COUGHS AND FORCED COUGHS

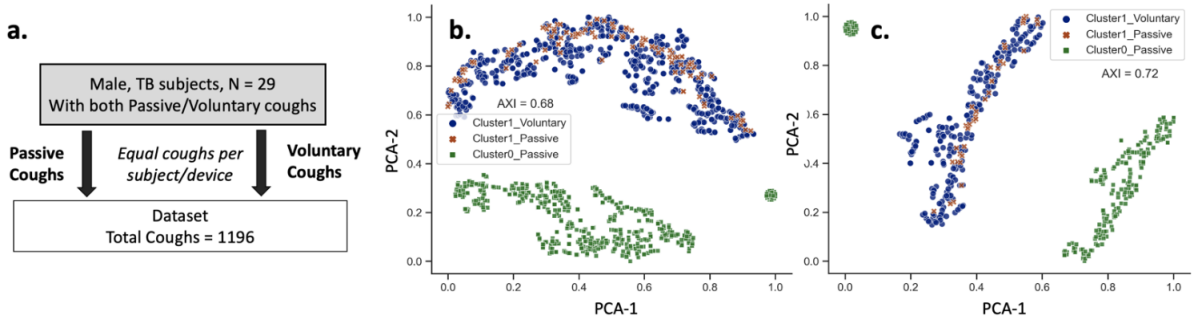


Figure 10.1: Voluntary Cough Analysis: (a) Dataset to analyze voluntary vs passive cough clus- ters (b) Cluster using voluntary and passive cough scalograms from Male (TB) subjects plotted with t-sne. (f) A similar t-sne plot for clustering result using coughs of TB (Male) subjects recorded using smartphone only.

Forced coughs have generated a lot of interest for cough screening and one of the reason for it's popularity is the ease of implementation [51, 171]. The Nairobi dataset provided an unique opportunity to compare passive and voluntary coughs of same subjects and to answer the research question - Can minimal prompt based forced coughs be a proxy for passive coughs in TB screening?

## *10.1 Prior Work*

Passive and voluntary cough are comprised of three phases - inspiration, compression, and expulsion and sound 'similar' to human ear. Despite the similarities in the sound, researchers have identified some physiological differences between the cough types [78, 79]. Researchers

have shown differences in the functional organization and coordination of muscular activity between reflex and voluntary cough [78]. For example, it was found that abdominal expiratory muscleelectromyography (EMG) activity for reflex cough was greater compared to forced coughs with equivalent airflow rates. In a 2015 study [18], with 25 subjects measured physical properties of cough using respiratory inductance plethysmography, it was found, Lung volume initiation (LVI; p = and lung volume excursion were significantly greater for forced cough compared to passive coughs. Differences between these two cough types likely reflect differences in the coordination of the respiratory and laryngeal subsystems [54]. Hegland et. al in a study with subjects having Parkinson's disease (PD) state that clinicians should be aware that evaluation of cough function using voluntary cough tasks overestimates the peak expiratory flow rate (PEFR) and cough expired volume (CEV) that would be achieved during passive cough in PD patients [54]. Previous works have focused on mechanical evaluation of coughs but are limited in terms of time and frequency domain analysis of cough sounds.

### *10.2 Model*

We verify if the voluntary and passive coughs from the same individuals are identical or not. Equal number of passive and voluntary coughs per device and subject is selected from the Nairobi dataset, giving us a total of 874 passive coughs and 874 voluntary coughs from 41 individuals with TB (all male). Coughs are converted to scalogram images, normalized, flattened (3, 448*224) and then reduced using PCA (variance of 0.98). The data is fitted using K- Means clustering with n=2 and clustering accuracy is measured using adjusted random index (AXI). The PCA features are plotted using 2-D tsne to visualize the k-mean clusters. Clustering process is summarized in Fig. 10.1(a)

### *10.3 Result*

We have used K-Means clustering to compare clusters of voluntary vs passive coughs to understand if they have similar features. To reduce variability in data, we have included only TB, male subjects, and equal number of coughs from each recording devices giving us

a dataset of 1196 coughs from 29 male, TB subjects (Fig. 10.1(a)). The clustering score, AXI is 68% indicating that the characteristics of voluntary coughs do differ from passive cough. The clusters are visualized using t-sne and plotted in Fig. 10.1(b) showing two clear clusters, cluster 0 only consists of passive coughs though cluster 1 is a mixture of few passive coughs and all of the voluntary coughs. On performing clustering on a subset of this dataset by including coughs only from one recording device (smartphone), we find the clustering accuracy improves further to 72% as shown in Fig. 10.1(c)

# Chapter 11

# CONCLUSION

In this thesis, I presented design and validation of screening tools designed for two different bacterial diseases. The overall design process included selection of a critical biomarker (pH for oral health and cough for Pulmonary TB) based on bacterial activity, determining the knowledge gap in sensing the biomarker, and development/validation of new sensing system to bridge the gap. This chapter summarizes the results for the individual tools and discusses implications of its introduction in disease screening.

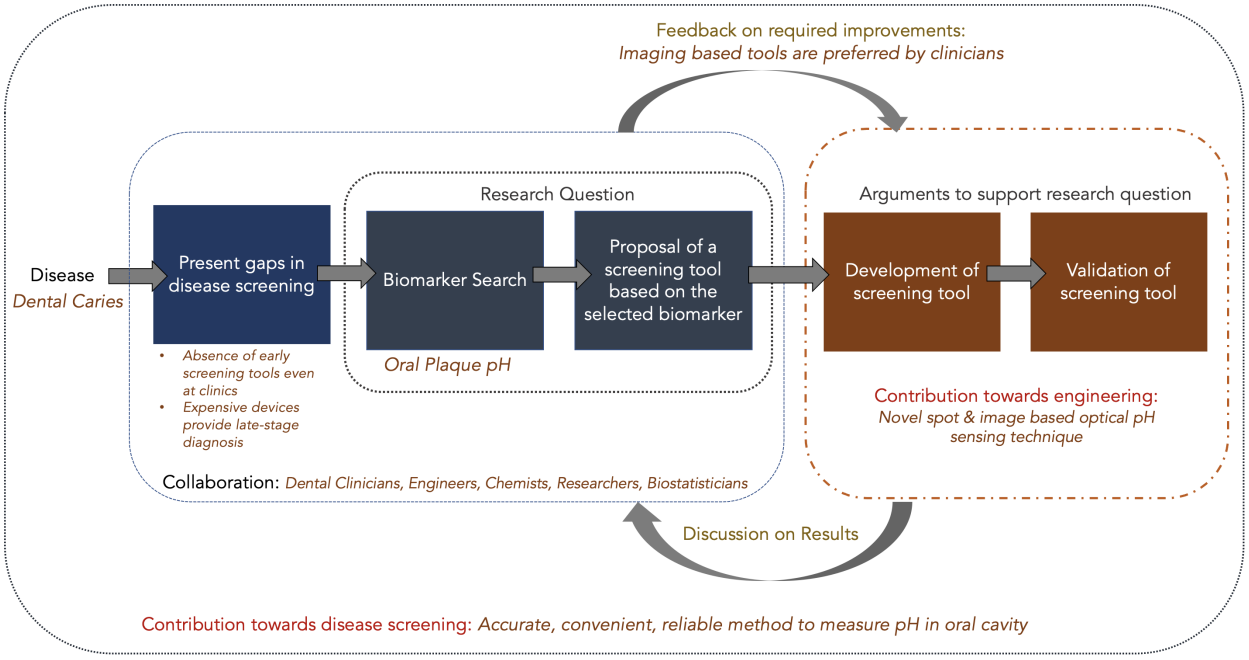


Figure 11.1: Design process for O-pH.

### *11.1 O-pH: Summary and Implications*

O-pH measures acidification of the oral biofilm which is a critical step in the caries process, unlike indirect optical methods that rely on the presence of specific bacterial species in the biofilm. The device is capable of measuring pH of plaque at occlusal pits and fissures and interproximal surfaces with repeatable measurements (Fig. 11.1). Fast diffusion of sodium fluorescein dye into the biofilm enables measurement of pH inside the biofilm's micro-environment rather than pH on the saliva surface. Additionally, the dye-based methodology allows measurement of extracellular pH without disturbing the biofilm. The initial clinical study with 30 subjects has shown O-pH's capability to differentiate between low and high plaque load subjects using pH measurements. Future studies are needed to confirm its utility as a hygiene monitoring device and to measure pH trends within groups with low plaque load. We noticed, one of the drawbacks of a point-based device was uncertainty of probing the same region before and after a sugar rinse. This limitation was addressed by proposing an imaging-based pH monitoring device developed on the same principle as O-pH and tested on one subject. mm-SFE scope results indicated its ability to track rest and drop pH with images. Further clinical studies are needed to evaluate its usability and accuracy.

In the clinic, O-pH could be part of each dental visit to capture the pH heatmap of patient's oral cavity. Trends across visits can provide feedback to the dentist and patient about changes in lifestyle that is positively or adversely affect the oral health, ultimately providing a method to check growth of harmful oral bacteria. Thus, O-pH and mm-SFE scope are a step towards development of tools that can break the cycle of lagging dental indicators by providing site-specific trends that monitors direct bio-chemical properties affecting enamel health. Along with plaque pH, O-pH can be used to monitor salivary pH that provides a generalized oral health and is under investigation as an important biomarker for oral cancer and other health diseases [142, 128, 68, 124].

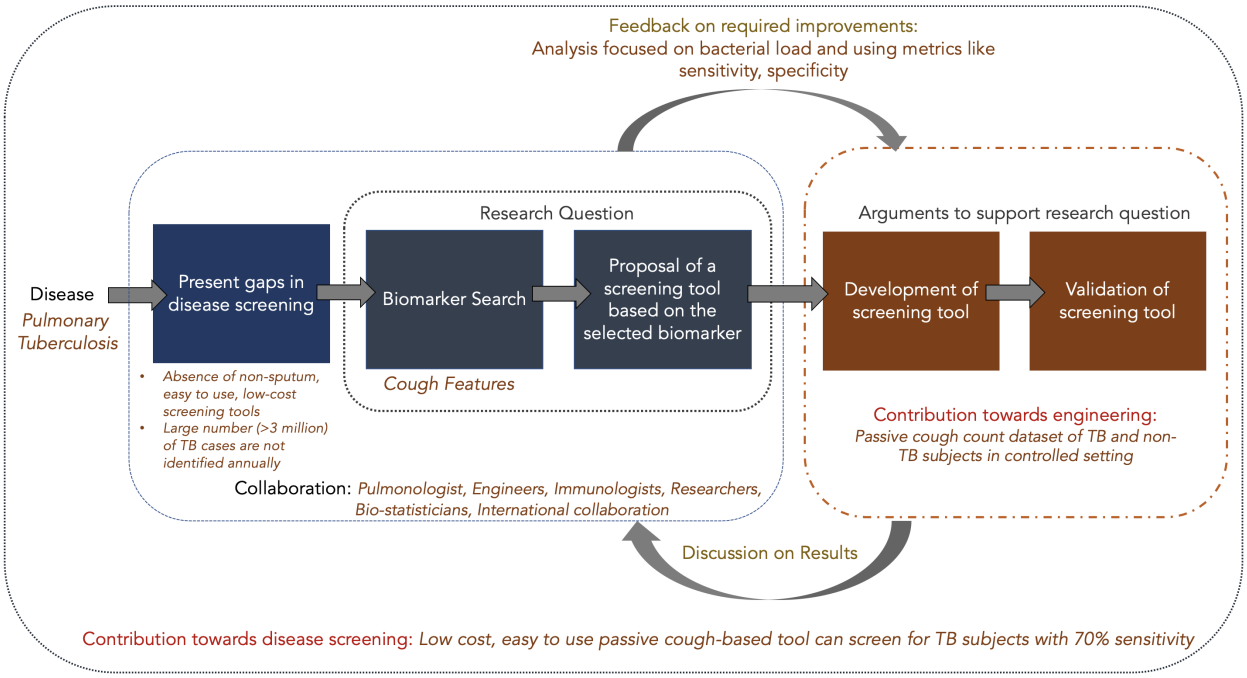


Figure 11.2: Design process for TBscreen.

## *11.2 TBscreen: Summary and Implications*

The Nairobi cough dataset provides a unique access to passive coughs of both disease (TB) and control group (non-TB) with minimal ambient interference and recorded in an identical environmental setup. Our findings support the feasibility of using a widely available recording device (smartphones) for a point-of-care cough-based TB screen (Fig. 11.2). The smartphone-based model performed best in identifying participants with high GeneXpert grades or cavitary findings on lung disease supporting a role for cough detection in identifying persons with pulmonary TB who are most infectious. Our preliminary results indicate that spectral signature of passive and forced coughs differ and need to be cautious in using forced coughs as proxy for passive coughs in TB screening. Looking at the present gaps in TB diagnosis/screening (Fig. 11.3), this tool can be vital in reducing the load of missing three million cases as it is easy to implement and requires minimal expertise to operate.

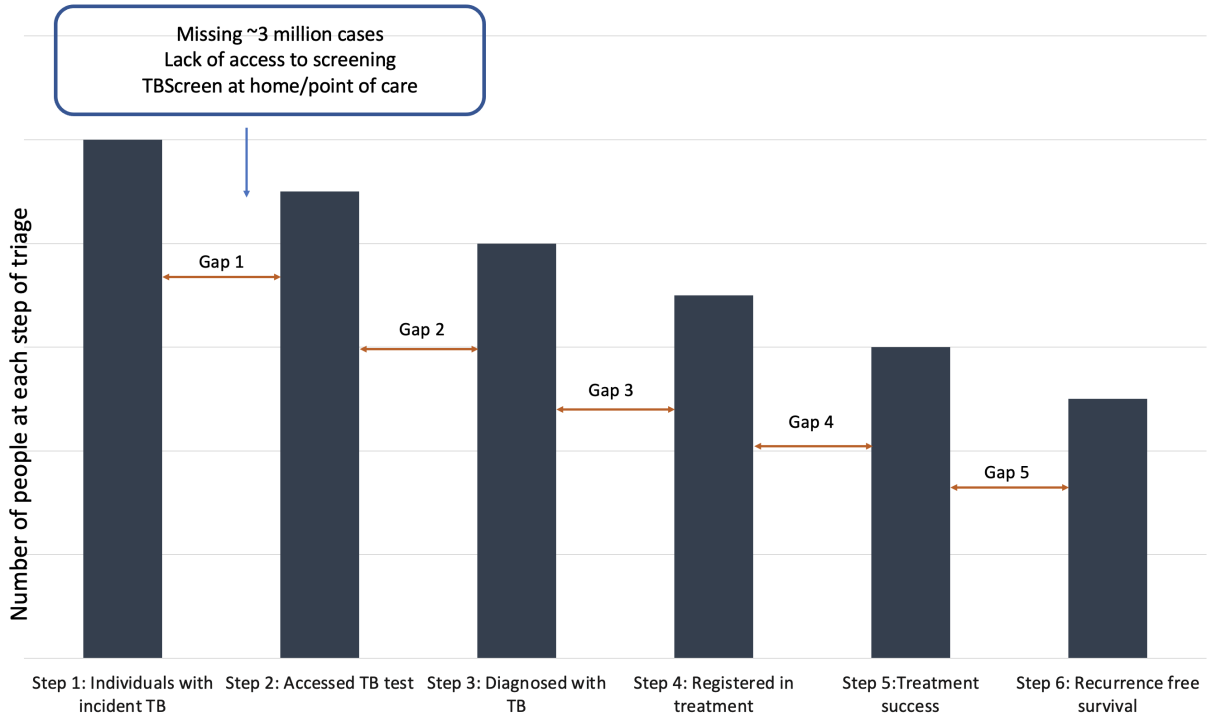


Figure 11.3: TB cascade of care. TBscreen can play an important role in decreasing gap1.

Community health workers could distribute phones in localities with high incidence rate to analyze cough sounds, helping in early screening and minimizing the transmission. Screening of passive cough in clinics could be more challenging to implement than screening based on forced cough as the test would be dependent on how often the patient coughs naturally. Alternatively, clinics could be fitted with microphones to analyze coughs in the background while the clinicians evaluate the patient. In case, a low number of coughs are recorded, patients could be sent home with the app on their phone. After COVID-19, we do understand how critical it is to identify hot spots early on and TBscreen could be used as a public health intervention to screen congregate settings for interruption of transmission events.

# BIBLIOGRAPHY

[1] Oral health in america: Advances and challenges. *National Institutes of Health*, 2021.

[2] A Aamdal-Scheie, W-M Luan, G Dahlen, and O Fejerskov. Plaque pH and microflora of dental plaque on sound and carious root surfaces. *Journal of dental research*, 75(11):1901–1908, 1996.

[3] Udantha R Abeyratne, Vinayak Swarnkar, Amalia Setyati, and Rina Triasih. Cough sound analysis can rapidly diagnose childhood pneumonia. *Annals of biomedical engineering*, 41(11):2448–2462, 2013.

[4] Jacqueline M Achkar, Stephen D Lawn, Mahomed-Yunus S Moosa, Colleen A Wright, and Victoria O Kasprowicz. Adjunctive tests for diagnosis of tuberculosis: serology, elispot for site-specific lymphocytes, urinary lipoarabinomannan, string test, and fine needle aspiration. *Journal of Infectious Diseases*, 204(suppl_4):S1130–S1141, 2011.

[5] Ali Murat Aktan, Mehmet Ata Cebe, Mehmet Ertuğrul Çiftçi, and Emine Şirin Karaarslan. A novel led-based device for occlusal caries detection. *Lasers in medical science*, 27(6):1157–1163, 2012.

[6] Luis B Almeida. The fractional fourier transform and time-frequency representations. *IEEE Transactions on signal processing*, 42(11):3084–3091, 1994.

[7] Joan M Almost, Elizabeth G VanDenKerkhof, Peter Strahlendorf, Louise Caicco Tett, Joanna Noonan, Thomas Hayes, Ryan Adam, Jeremy Holden, Tracy Kent-Hillis, Mike McDonald, et al. A study of leading indicators for occupational health and safety management systems in healthcare. *BMC health services research*, 18(1):1–7, 2018.

[8] Keith Angelino, Pratik Shah, David A Edlund, Mrinal Mohit, and Gregory Yauney. Clinical validation and assessment of a modular fluorescent imaging system and algorithm for rapid detection and quantification of dental plaque. *BMC oral health*, 17(1):162, 2017.

[9] Sharmila Baliga, Sangeeta Muglikar, and Rahul Kale. Salivary ph: A diagnostic biomarker. *Journal of Indian Society of Periodontology*, 17(4):461, 2013.

[10] Muhittin Bayram and Muhammet Ali Arserim. Analysis of epileptic ieeg data by applying convolutional neural networks to low-frequency scalograms. *IEEE Access*, 9:162520–162529, 2021.

[11] KW Becker, Cornie Scheffer, MM Blanckenberg, and AH Diacon. Analysis of adventitious lung sounds originating from pulmonary tuberculosis. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4334–4337. IEEE, 2013.

[12] Emilie Betrisey, Nicolas Rizcalla, Ivo Krejci, and Stefano Ardu. Caries diagnosis using light fluorescence devices: Vistaproof and diagnodent. *Odontology*, 102(2):330–335, 2014.

[13] Neerja Bhatla and Seema Singhal. Primary hpv screening for cervical cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 65:98–108, 2020.

[14] RG Blanks, MG Wallis, and SM Moss. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the uk national health service breast screening programme. *Journal of Medical screening*, 5(4):195–201, 1998.

[15] GHR Botha, Grant Theron, RM Warren, Marisa Klopper, Keertan Dheda, PD Van Helden, and TR Niesler. Detection of tuberculosis by automatic cough sound analysis. *Physiological measurement*, 39(4):045005, 2018.

[16] William H Bowen, Robert A Burne, Hui Wu, and Hyun Koo. Oral biofilms: pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends in microbiology*, 26(3):229–242, 2018.

[17] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[18] Alexandra E Brandimore, Michelle S Troche, Jessica E Huber, and Karen W Hegland. Respiratory kinematic and airflow differences between reflex and voluntary cough in healthy young adults. *Frontiers in Physiology*, 6:284, 2015.

[19] Rhys Brown and Lee Evans. Acoustics and the smartphone. *Proceedings of ACOUSTICS, November*, 2(4):2011, 2011.

[20] Andrew Burns, Hooisweng Ow, and Ulrich Wiesner. Fluorescent core–shell silica nanoparticles: towards "lab on a particle" architectures for nanobiotechnology. *Chemical Society Reviews*, 35(11):1028–1042, 2006.

[21] Andrew Burns, Prabuddha Sengupta, Tara Zedayko, Barbara Baird, and Ulrich Wiesner. Core/shell fluorescent silica nanoparticles for chemical sensing: towards single-particle laboratories. *Small*, 2(6):723–726, 2006.

[22] DE Caldwell, DR Korber, and JR Lawrence. Imaging of bacterial cells by fluorescence exclusion using scanning confocal laser microscopy. *Journal of Microbiological Methods*, 15(4):249–261, 1992.

[23] A Carlén, H Hassan, and P Lingström. The 'strip method': a simple method for plaque pH assessment. *Caries research*, 44(4):341–344, 2010.

[24] Jeremiah Chakaya, Mishal Khan, Francine Ntoumi, Eleni Aklillu, Razia Fatima, Peter Mwaba, Nathan Kapata, Sayoki Mfinanga, Seyed Ehtesham Hasnain, Patrick DMC Katoto, et al. Global tuberculosis report 2020–reflections on the global tb burden, treatment and prevention efforts. *International journal of infectious diseases*, 113:S7–S12, 2021.

[25] Jeremiah Chakaya, Eskild Petersen, Rebecca Nantanda, Brenda N Mungai, Giovanni Battista Migliori, Farhana Amanullah, Patrick Lungu, Francine Ntoumi, Nagalingeswaran Kumarasamy, Markus Maeurer, et al. The who global tuberculosis 2021 report–not so good news and turning the tide back to end tb. *International Journal of Infectious Diseases*, 124:S26–S29, 2022.

[26] Anne B Chang. The physiology of cough. *Paediatric respiratory reviews*, 7(1):2–8, 2006.

[27] Fabio Cocco, Maria Grazia Cagetti, Peter Lingström, Nicole Camoni, and Guglielmo Campus. The strip method and the microelectrode technique in assessing dental plaque ph. *Minerva stomatologica*, 66(6):241–247, 2017.

[28] James W Cooley, Peter AW Lewis, and Peter D Welch. The fast fourier transform and its applications. *IEEE Transactions on Education*, 12(1):27–34, 1969.

[29] Prasanta Kumar Das, Somtirtha B Ganguly, Bodhisatya Mandal, et al. Sputum smear microscopy in tuberculosis: It is still relevant in the era of molecular diagnosis when seen from the public health perspective. *Biomedical and Biotechnology Research Journal (BBRJ)*, 3(2):77, 2019.

[30] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990.

[31] J Lucian Davis, Adithya Cattamanchi, Luis E Cuevas, Philip C Hopewell, and Karen R Steingart. Diagnostic accuracy of same-day microscopy versus standard microscopy for pulmonary tuberculosis: a systematic review and meta-analysis. *The Lancet infectious diseases*, 13(2):147–154, 2013.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[33] Ashar Dhana, Yohhei Hamada, Andre P Kengne, Andrew D Kerkhoff, Molebogeng X Rangaka, Tamara Kredo, Annabel Baddeley, Cecily Miller, Satvinder Singh, Yasmeen Hanifa, et al. Tuberculosis screening among ambulatory people living with hiv: a systematic review and individual participant data meta-analysis. *The Lancet Infectious Diseases*, 22(4):507–518, 2022.

[34] Benin Dikmen. ICDAS II criteria (international caries detection and assessment system). *Journal of Istanbul University Faculty of Dentistry*, 49(3):63, 2015.

[35] Michele Baffi Diniz, George Joseph Eckert, Carlos González-Cabezas, Rita de Cássia Loiola Cordeiro, and Andrea Gonçalves Ferreira-Zandona. Caries detection around restorations using icdas and optical devices. *Journal of Esthetic and Restorative Dentistry*, 28(2):110–121, 2016.

[36] Michael Dodds, Simon Roland, Michael Edgar, and Martin Thornhill. Saliva a review of its role in maintaining oral health and preventing dental disease. *Bdj Team*, 2:15123, 2015.

[37] Y-M Dong, EIF Pearce, L Yue, MJ Larsen, X-J Gao, and J-D Wang. Plaque pH and associated parameters in relation to caries. *Caries research*, 33(6):428–436, 1999.

[38] Monika Dörfler, Roswitha Bammer, and Thomas Grill. Inside the spectrogram: Convolutional neural networks in audio processing. In *2017 international conference on sampling theory and applications (SampTA)*, pages 152–155. IEEE, 2017.

[39] Mohamed M Elsutohy, Veeren M Chauhan, Robert Markus, Mohammed Aref Kyyaly, Saul JB Tendler, and Jonathan W Aylott. Real-time measurement of the intracellular ph of yeast cells during glucose metabolism using ratiometric fluorescent nanosensors. *Nanoscale*, 9(18):5904–5911, 2017.

[40] Marie Farge. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 24(1):395–458, 1992.

[41] O Fejerskov, A AA Scheie, and F Manji. The effect of sucrose on plaque pH in the primary and permanent dentition of caries-inactive and-active kenyan children. *Journal of dental research*, 71(1):25–31, 1992.

[42] Ming-Ren S Fuh, Lloyd W Burgess, Tomas Hirschfeld, Gary D Christian, and Francis Wang. Single fibre optic fluorescence ph probe. *Analyst*, 112(8):1159–1163, 1987.

[43] Ferdia A Gallagher, Mikko I Kettunen, Sam E Day, De-En Hu, Jan Henrik Ardenkjaer-Larsen, René in 't Zandt, Pernille R Jensen, Magnus Karlsson, Klaes Golman, Mathilde H Lerche, et al. Magnetic resonance imaging of ph in vivo using hyperpolarized 13c-labelled bicarbonate. *Nature*, 453(7197):940–943, 2008.

[44] Robert J Gillies and David L Morse. In vivo magnetic resonance spectroscopy in cancer. *Annu. Rev. Biomed. Eng.*, 7:287–326, 2005.

[45] Pierre L Goupillaud, Alexander Grossmann, and Jean Morlet. A simplified view of the cycle-octave and voice representations of seismic signals. In *SEG Technical Program Expanded Abstracts 1984*, pages 379–382. Society of Exploration Geophysicists, 1984.

[46] Jasmine Y Graham, Leonard Y Nelson, and Eric J Seibel. Optical measurement of acidification of human dental plaque in vitro. In *Lasers in Dentistry XXIV*, volume 10473, pages 48–57. SPIE, 2018.

[47] Clare Green, Jim F Huggett, Elizabeth Talbot, Peter Mwaba, Klaus Reither, and Alimuddin I Zumla. Rapid diagnosis of tuberculosis through the detection of mycobacterial dna in urine by nucleic acid amplification methods. *The Lancet infectious diseases*, 9(8):505–511, 2009.

[48] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[49] Cochrane Infectious Diseases Group, Jerry S Zifodya, Jonah S Kreniske, Ian Schiller, Mikashmi Kohli, Nandini Dendukuri, Samuel G Schumacher, Eleanor A Ochodo, Frederick Haraka, Alice A Zwerling, et al. Xpert ultra versus xpert mtb/rif for pulmonary tuberculosis and rifampicin resistance in adults with presumptive pulmonary tuberculosis. *Cochrane Database of Systematic Reviews*, 2021(5), 1996.

[50] FDA-NIH Biomarker Working Group et al. Best (biomarkers, endpoints, and other tools) resource. 2016. *URL: https://www. ncbi. nlm. nih. gov/books/NBK326791*, 2016.

[51] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine*, 5(1):16, 2022.

[52] Moinuddin Hassan, Jason Riley, Victor Chernomordik, Paul Smith, Randall Pursley, Sang Bong Lee, Jacek Capala, and Gandjbakhche Amir H. Fluorescence lifetime imaging system for in vivo studies. *Molecular imaging*, 6(4):7290–2007, 2007.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[54] Karen Wheeler Hegland, Michelle S Troche, Alexandra E Brandimore, Paul W Davenport, and Michael S Okun. Comparison of voluntary and reflex cough effectiveness in parkinson's disease. *Parkinsonism & related disorders*, 20(11):1226–1230, 2014.

[55] JW Hein. Some of the variables associated with in situ plaque pH measurements. *Journal of Dental Research*, 34:695, 1955.

[56] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[57] Raimund Hibst, Robert Paulus, and Adrian Lussi. Detection of occlusal caries by laser fluorescence: basic and clinical investigations. *Medical Laser Application*, 16(3):205–213, 2001.

[58] Gabriela Hidalgo, Andrew Burns, Erik Herz, Anthony G Hay, Paul L Houston, Ulrich Wiesner, and Leonard W Lion. Functional tomographic fluorescence imaging of ph microenvironments in microbial biofilms by use of silica nanoparticle sensors. *Applied and environmental microbiology*, 75(23):7426–7435, 2009.

[59] Christopher K Hope, Karen Billingsley, Elbert de Josselin de Jong, and Susan M Higham. A preliminary study of the effects of ph upon fluorescence in suspensions of prevotella intermedia. *PLoS One*, 11(7):e0158835, 2016.

[60] Christopher K Hope and Susan M Higham. Evaluating the effect of local ph on fluorescence emissions from oral bacteria of the genus prevotella. *Journal of Biomedical Optics*, 21(8):084003–084003, 2016.

[61] CK Hope, E De Josselin De Jong, MRT Field, SP Valappil, and SM Higham. Photobleaching of red fluorescence in oral biofilms. *Journal of periodontal research*, 46(2):228–234, 2011.

[62] Md Afzal Hossan, Sheeraz Memon, and Mark A Gregory. A novel approach for mfcc feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5. IEEE, 2010.

[63] Chuqin Huang, Manuja Sharma, Lauren K Lee, Matthew D Carson, Mark E Fauver, and Eric J Seibel. Optical imaging of dental plaque ph. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, page 113152Z. International Society for Optics and Photonics, 2020.

[64] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, MD Iftikhar Hussain, and Muhammad Nabeel. Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20:100378, 2020.

[65] Rajeev Jain, Megha Mathur, Shalini Sikarwar, and Alok Mittal. Removal of the hazardous dye rhodamine b through photocatalytic and adsorption treatments. *Journal of environmental management*, 85(4):956–964, 2007.

[66] Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[67] Hyun Jeong, Rafael Salas-Montiel, Gilles Lerondel, and Mun Seok Jeong. Indium gallium nitride-based ultraviolet, blue, and green light-emitting diodes functionalized with shallow periodic hole patterns. *Scientific reports*, 7:45726, 2017.

[68] Rabia Khalaila, Miri Cohen, and Jamal Zidan. Is salivary ph a marker of depression among older spousal caregivers for cancer patients? *Behavioral Medicine*, 40(2):71–80, 2014.

[69] Sandra V Kik, Claudia M Denkinger, Martina Casenghi, Caroline Vadnais, and Madhukar Pai. Tuberculosis diagnostics: which target product profiles should be prioritised? *European Respiratory Journal*, 44(2):537–540, 2014.

[70] Sandra V Kik, Claudia M Denkinger, Pamela Chedore, and Madhukar Pai. Replacing smear microscopy for the diagnosis of tuberculosis: what is the market potential? *European Respiratory Journal*, 43(6):1793–1796, 2014.

[71] Mogens Kilian, ILC Chapple, M Hannig, PD Marsh, V Meuric, AML Pedersen, MS Tonetti, WG Wade, and E Zaura. The oral microbiome–an update for oral healthcare professionals. *British dental journal*, 221(10):657–666, 2016.

[72] Nectarios Klonis and William H Sawyer. Spectral properties of the prototropic forms of fluorescein in aqueous solution. *Journal of fluorescence*, 6:147–157, 1996.

[73] Pradeep Koppolu, Sunkara Sirisha, Soumya Penala, Pathakota Krishnajaneya Reddy, Dalal H Alotaibi, Ghadah Salim Abusalim, Amara Swapna Lingam, Areej H Mukhtar, Ali Barakat, and Ahmed A AlMokhatieb. Correlation of blood and salivary ph levels in healthy, gingivitis, and periodontitis patients before and after non-surgical periodontal therapy. *Diagnostics*, 12(1):97, 2022.

[74] J Korpáš, J Sadloňová, and M Vrabec. Analysis of the cough sound: an overview. *Pulmonary pharmacology*, 9(5-6):261–268, 1996.

[75] Mikael Kubista, Robert Sjoeback, and Bo Albinsson. Determination of equilibrium constants by chemometric analysis of spectroscopic data. *Analytical Chemistry*, 65(8):994–998, 1993.

[76] Megumi Kuribayashi, Yuichi Kitasako, Khairul Matin, Alireza Sadr, Kanako Shida, and Junji Tagami. Intraoral pH measurement of carious lesions with qpcr of cariogenic bacteria to differentiate caries activity. *Journal of dentistry*, 40(3):222–228, 2012.

[77] HJ Landau. Sampling, data transmission, and the nyquist rate. *Proceedings of the IEEE*, 55(10):1701–1706, 1967.

[78] Dan Lasserson, Kerry Mills, Ramamurthy Arunachalam, Michael Polkey, John Moxham, and Lalit Kalra. Differences in motor activation of voluntary and reflex cough in humans. *Thorax*, 61(8):699–705, 2006.

[79] Federico Lavorini, Tito Pantaleo, Pietro Geri, Donatella Mutolo, Massimo Pistolesi, and Giovanni A Fontana. Cough and ventilatory adjustments evoked by aerosolised capsaicin and distilled water (fog) in man. *Respiratory physiology & neurobiology*, 156(3):331–339, 2007.

[80] Cameron M Lee, Christoph J Engelbrecht, Timothy D Soper, Fritjof Helmchen, and Eric J Seibel. Scanning fiber endoscopy with highly flexible, 1 mm catheterscopes for wide-field, full-color imaging. *Journal of biophotonics*, 3(5-6):385–407, 2010.

[81] Juying Lei, Lingzhi Wang, and Jinlong Zhang. Ratiometric ph sensor based on mesoporous silica nanoparticles and förster resonance energy transfer. *Chemical communications*, 46(44):8445–8447, 2010.

[82] AM Lennon, W Buchalla, L Brune, O Zimmermann, U Gross, and T Attin. The ability of selected oral microorganisms to emit red fluorescence. *Caries research*, 40(1):2–5, 2006.

[83] Shiying Li, Bin Liu, Mingli Peng, Min Chen, Wenwei Yin, Hui Tang, Yuxuan Luo, Peng Hu, and Hong Ren. Diagnostic accuracy of xpert mtb/rif for tuberculosis detection in different regions with different endemic burden: A systematic review and meta-analysis. *PloS one*, 12(7):e0180725, 2017.

[84] Christopher Liner. An overview of wavelet transform concepts and applications. *University of Houston*, pages 1–17, 2010.

[85] P Lingström, D Birkhed, Y Granfeldt, and I Björck. pH measurements of human dental plaque after consumption of starchy foods using the microtouch and the sampling method. *Caries research*, 27(5):394–401, 1993.

[86] Peter Lingström, Floris OJ Van Ruyven, Johannes Van Houte, and Ralph Kent. The pH of dental plaque in its relation to early enamel caries and dental plaque flora in humans. *Journal of dental research*, 79(2):770–777, 2000.

[87] Robert Loddenkemper, Marc Lipman, and Alimuddin Zumla. Clinical aspects of adult tuberculosis. *Cold Spring Harbor perspectives in medicine*, 6(1):a017848, 2016.

[88] Robert G Loudon and Linda C Brown. Cough frequency in patients with respiratory disease. *American Review of Respiratory Disease*, 96(6):1137–1143, 1967.

[89] A Lussi, R Hibst, and R Paulus. Diagnodent: an optical method for caries detection. *Journal of dental research*, 83(1_suppl):80–83, 2004.

[90] Robert M. Stephan. Changes in hydrogen-ion concentration on tooth surfaces and in carious lesions. *Journal of the American Dental Association*, 27:718–723, 1940.

[91] Robert M. Stephan. Intra-oral hydrogen-ion concentrations associated with dental caries activity. *Journal of Dental Research*, 23(4):257–266, 1944.

[92] AK Majumder and SK Chowdhury. Recording and preliminary analysis of respiratory sounds from tuberculosis patients. *Medical and Biological Engineering and Computing*, 19:561–564, 1981.

[93] PD Marsh and Egija Zaura. Dental biofilm: ecological interactions in health and disease. *Journal of clinical periodontology*, 44:S12–S22, 2017.

[94] Christopher McCabe, Karl Claxton, and Anthony J Culyer. The nice cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics*, 26:733–744, 2008.

[95] Koki Minami, Huimin Lu, Tohru Kamiya, Shingo Mabu, and Shoji Kido. Automatic classification of respiratory sounds based on convolutional neural network with multi images. In *2020 5th International Conference on Biomedical Imaging, Signal Processing*, pages 17–21, 2020.

[96] Jessica Minion, Erika Leung, Elizabeth Talbot, Keertan Dheda, Madhukar Pai, and Dick Menzies. Diagnosing tuberculosis with urine lipoarabinomannan: systematic review and meta-analysis. *European Respiratory Journal*, 38(6):1398–1405, 2011.

[97] Zaid Nabulsi, Andrew Sellergren, Shahar Jamshy, Charles Lau, Edward Santos, Atilla P Kiraly, Wenxing Ye, Jie Yang, Rory Pilgrim, Sahar Kazemzadeh, et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. *Scientific reports*, 11(1):15523, 2021.

[98] Ruvandhi R Nathavitharana, Alberto L Garcia-Basteiro, Morten Ruhwald, Frank Cobelens, and Grant Theron. Reimagining the status quo: How close are we to rapid sputum-free tuberculosis diagnostics for all? *EBioMedicine*, page 103939, 2022.

[99] Karin Neubert and Edgar Brunner. A studentized permutation test for the nonparametric behrens–fisher problem. *Computational Statistics & Data Analysis*, 51(10):5192–5204, 2007.

[100] Nha Nguyen, An Vo, Haibin Sun, and Heng Huang. Heavy-tailed noise suppression and derivative wavelet scalogram for detecting dna copy number aberrations. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1625–1635, 2017.

[101] Nha Nguyen, An Vo, Haibin Sun, and Heng Huang. Heavy-tailed noise suppression and derivative wavelet scalogram for detecting dna copy number aberrations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5):1625–1635, 2018.

[102] Hanieh Nokhbatolfoghahaie, Marzieh Alikhasi, Nasim Chiniforush, Farzaneh Khoei, Nassimeh Safavi, and Behnoush Yaghoub Zadeh. Evaluation of accuracy of diagnodent in diagnosis of primary and secondary caries in comparison to conventional methods. *Journal of lasers in medical sciences*, 4(4):159, 2013.

[103] TC O'Haver. Teaching and learning chemometrics with matlab. *Chemometrics and Intelligent Laboratory Systems*, 6(2):95–103, 1989.

[104] World Health Organization. *Systematic screening for active tuberculosis: principles and recommendations*. World Health Organization, 2013.

[105] World Health Organization et al. Commercial serodiagnostic tests for diagnosis of tuberculosis: policy statement. Technical report, World Health Organization, 2011.

[106] World Health Organization et al. High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28-29 april 2014, geneva, switzerland. Technical report, World Health Organization, 2014.

[107] World Health Organization et al. Implementing the end tb strategy: the essentials. Technical report, World Health Organization, 2015.

[108] WORLD HEALTH ORGANIZATION et al. Who operational handbook on tuberculosis. module 2: Screening-systematic screening for tuberculosis disease [m/ol]. geneva. *World Health Organization*, pages 2021–11, 2021.

[109] Madhurananda Pahar, Marisa Klopper, Byron Reeve, Rob Warren, Grant Theron, and Thomas Niesler. Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, 42(10):105014, 2021.

[110] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. Covid-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine*, 135:104572, 2021.

[111] Madhukar Pai, Marcel A Behr, David Dowdy, Keertan Dheda, Maziar Divangahi, Catharina C Boehme, Ann Ginsberg, Soumya Swaminathan, Melvin Spigelman, Haileyesus Getahun, et al. Tuberculosis (primer). *Nature Reviews: Disease Primers*, 2(1), 2016.

[112] Pune Nina Paqué, Jenni Hjerppe, Anina N Zuercher, Ronald E Jung, and Tim Joda. Salivary biomarkers as key to monitor personalized oral healthcare and precision dentistry: A scoping review. *Frontiers in Oral Health*, 3, 2022.

[113] 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.

[114] 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.

[115] Rahul Pathri, Shekhar Jha, Samarth Tandon, and Suryakanth GangaShetty. Acoustic epidemiology of pulmonary tuberculosis (tb) & covid19 leveraging ai/ml. *medRxiv*, pages 2022–02, 2022.

[116] MP Petticrew, AJ Sowden, D Lister-Sharp, and K Wright. False-negative results in screening programmes: systematic review of impact and implications. *Health technology assessment (Winchester, England)*, 4(5):1–120, 2000.

[117] TO Peulen. Wilkinson kj environ. *Sci. Technol*, 45:3367–3373, 2011.

[118] Riccardo Piccazzo, Francesco Paparo, and Giacomo Garlaschi. Diagnostic accuracy of chest radiography for the diagnosis of tuberculosis (tb) and its role in the detection of latent tb infection: a systematic review. *The Journal of Rheumatology Supplement*, 91:32–40, 2014.

[119] AJ Preston and WM Edgar. Developments in dental plaque pH modelling. *Journal of dentistry*, 33(3):209–222, 2005.

[120] IA Pretty, WM Edgar, PW Smith, and SM Higham. Quantification of dental plaque in the research environment. *Journal of dentistry*, 33(3):193–207, 2005.

[121] Iain A Pretty. Caries detection and diagnosis: novel technologies. *Journal of dentistry*, 34(10):727–739, 2006.

[122] EM Aranibar Quiroz, Torgny Alstad, G Campus, Dowen Birkhed, and Peter Lingström. Relationship between plaque pH and different caries-associated variables in a group of adolescents with varying caries prevalence. *Caries Research*, 48(2):147–153, 2014.

[123] Md Toufiq Rahman, Andrew J Codlin, Md Mahfuzur Rahman, Ayenun Nahar, Mehdi Reja, Tariqul Islam, Zhi Zhen Qin, Md Abdus Shakur Khan, Sayera Banu, and Jacob Creswell. An evaluation of automated chest radiography reading software for tuberculosis screening among public-and private-sector patients. *European Respiratory Journal*, 49(5), 2017.

[124] Atmakuri Shanmukha Ramya, Divya Uppala, Sumit Majumdar, Ch Surekha, and KGK Deepak. Are salivary amylase and ph–prognostic indicators of cancers? *Journal of oral biology and craniofacial research*, 5(2):81–85, 2015.

[125] Elisabeth Raner, Lina Lindqvist, Sofia Johansson, Haidar Hassan, Anette Carlén, Narong Suksu-art, and Gunnar Dahlén. ph and bacterial profile of dental plaque in children and adults of a low caries population. *Anaerobe*, 27:64–70, 2014.

[126] Suriani Abdul Rani, Betsey Pitts, and Philip S Stewart. Rapid diffusion of fluorescent tracers into staphylococcus epidermidis biofilms visualized by time lapse microscopy. *Antimicrobial agents and chemotherapy*, 49(2):728–732, 2005.

[127] P Rechmann, Shasan W Liou, Beate MT Rechmann, and John DB Featherstone. Soprocare-450 nm wavelength detection tool for microbial plaque and gingival inflammation: a clinical study. In *Lasers in Dentistry XX*, volume 8929, page 892906. International Society for Optics and Photonics, 2014.

[128] RS Redman, NC Bayley, and ES Nylén. Salivary and serum biomarkers of inflammation in a man with metastatic medullary thyroid carcinoma and hyperreactive gingiva: a fourteen year odyssey. *Biotechnic & Histochemistry*, 94(6):389–397, 2019.

[129] Zhao Ren, Kun Qian, Zixing Zhang, Vedhas Pandit, Alice Baird, and Bjorn Schuller. Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica*, 5(3):662–669, 2018.

[130] Sam Rosen and Paul R Weisenstein. The effect of sugar solutions on pH of dental plaques from caries-susceptible and caries-free individuals. *Journal of dental research*, 44(5):845–849, 1965.

[131] Anne WS Rutjes, Johannes B Reitsma, Marcello Di Nisio, Nynke Smidt, Jeroen C Van Rijn, and Patrick MM Bossuyt. Evidence of bias and variation in diagnostic accuracy studies. *Cmaj*, 174(4):469–476, 2006.

[132] Alan G Ryder, Sarah Power, Thomas J Glynn, and John J Morrison. Time-domain measurement of fluorescence lifetime variation with ph. In *Biomarkers and Biological Spectral Imaging*, volume 4259, pages 102–109. SPIE, 2001.

[133] Sebastian Schlafer and Rikke L Meyer. Confocal microscopy imaging of the biofilm matrix. *Journal of microbiological methods*, 138:50–59, 2017.

[134] Luigi Monsù Scolaro, Mariangela Castriciano, Andrea Romeo, Salvatore Patane, Eugenio Cefalì, and Maria Allegrini. Aggregation behavior of protoporphyrin ix in aqueous solutions: clear evidence of vesicle formation. *The Journal of Physical Chemistry B*, 106(10):2453–2459, 2002.

[135] Eric J Seibel, Leonard Y Nelson, Manuja Sharma, and Jasmine Graham. Systems and methods for accurate optical ph sensing of biofilms, August 26 2021. US Patent App. 17/260,957.

[136] Juliet N Sekandi, Kevin Dobbin, James Oloya, Alphonse Okwera, Christopher C Whalen, and Phaedra S Corso. Cost-effectiveness analysis of community active case finding and household contact investigation for tuberculosis case detection in urban africa. *PloS one*, 10(2):e0117009, 2015.

[137] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[138] Manuja Sharma, Matthew D Carson, Jasmine Y Graham, Leonard Y Nelson, Shwetak Patel, and J Eric Seibel. Dental pH opti-wand (dpow): measuring oral acidity to guide enamel preservation. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3738–3741. IEEE, 2018.

[139] Manuja Sharma, Jasmine Y Graham, Philip A Walczak, Ryan M Nguyen, Lauren K Lee, Matthew D Carson, Leonard Y Nelson, Shwetak N Patel, Zheng Xu, and Eric J Seibel. Optical pH measurement system using a single fluorescent dye for assessing susceptibility to dental caries. *Journal of biomedical optics*, 24(1):017001, 2019.

[140] Manuja Sharma, Lauren K Lee, Matthew D Carson, David S Park, Se W An, Micah G Bovenkamp, Jess J Cayetano, Ian A Berude, Leonard Y Nelson, Zheng Xu, et al. O-ph: Optical ph monitor to measure dental biofilm acidity and assist in enamel health monitoring. *IEEE Transactions on Biomedical Engineering*, 69(9):2776–2786, 2022.

[141] Ali H. Shoeb and Gari D. Clifford. Chapter 16-wavelets ; multiscale activity in physiological signals c © 2006.

[142] T Shpitzer, Y Hamzany, G Bahar, R Feinmesser, D Savulescu, I Borovoi, M Gavish, and RM Nagler. Salivary analysis of oral cancer biomarkers. *British journal of cancer*, 101(7):1194–1198, 2009.

[143] Samiul Based Shuvo, Shams Nafisa Ali, Soham Irtiza Swapnil, Taufiq Hasan, and Mohammed Imamul Hassan Bhuiyan. A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2595–2603, 2020.

[144] R Sjôback. 1. nygren and m. kubista. *Spectrochim. Acta, Part A*, 51:L7, 1995.

[145] Robert Sjöback, Jan Nygren, and Mikael Kubista. Absorption and fluorescence properties of fluorescein. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 51(6):L7–L21, 1995.

[146] Jan Slavík. Intracellular ph of yeast cells measured with fluorescent probes. *Febs Lett*, 140(1):22–26, 1982.

[147] Bianca Sossen and Graeme Meintjes. Development of accurate non-sputum-based diagnostic tests for tuberculosis: an ongoing challenge. *The Lancet Global Health*, 11(1):e16–e17, 2023.

[148] KR Steingart, M Henry, S Laal, PC Hopewell, A Ramsay, and D Menzies. 935 j. cunningham, k. weldingh, and m. pai. 2007. a systematic review of 936 commercial serological antibody detection tests for the diagnosis of 937 extrapulmonary tuberculosis. *Thorax*, 62:911–918.

[149] Robert M. Stephan. Hydrogen ion concentration of the dental plaque. *Journal of Dental Research*, 17(3):251–256, 1938.

[150] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[151] Mari-Alina I Timoshchuk, Jeremy S Ridge, Amanda L Rugg, Leonard Y Nelson, Amy S Kim, and Eric J Seibel. Real-time porphyrin detection in plaque and caries: a case study. In *Lasers in Dentistry XXI*, volume 9306, page 93060C. International Society for Optics and Photonics, 2015.

[152] Eric J Topol. Is my cough covid-19? *The Lancet*, 396(10266):1874, 2020.

[153] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

[154] Anna NA Tosteson, Dennis G Fryback, Cristina S Hammond, Lucy G Hanna, Margaret R Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D Pisano. Consequences of false-positive screening mammograms. *JAMA internal medicine*, 174(6):954–961, 2014.

[155] Brian H Tracey, Germán Comina, Sandra Larson, Marjory Bravard, José W López, and Robert H Gilman. Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In *2011 Annual international conference of the IEEE engineering in medicine and biology society*, pages 6017–6020. IEEE, 2011.

[156] Samuel Turesky, Neville D Gilmore, and Irving Glickman. Reduced plaque formation by the chloromethyl analogue of victamine c. *Journal of periodontology*, 41(1):41–43, 1970.

[157] Srikanth Vasudevan, Anindita Saha, Michelle E Tarver, and Bakul Patel. Digital biomarkers: Convergence of digital health technologies and biomarkers. *NPJ digital medicine*, 5(1):36, 2022.

[158] Amy L Vavere, Gráinne B Biddlecombe, William M Spees, Joel R Garbow, Dayanjali Wijesinghe, Oleg A Andreev, Donald M Engelman, Yana K Reshetnyak, and Jason S Lewis. A novel technology for the imaging of acidic prostate tumors by positron emission tomography. *Cancer research*, 69(10):4510–4516, 2009.

[159] Catherine MC Volgenant, Michel A Hoogenkamp, Mark J Buijs, Egija Zaura, Jacob (Bob) M ten Cate, and Monique H van der Veen. Red fluorescent biofilm: the thick, the old, and the cariogenic. *Journal of oral microbiology*, 8(1):30346, 2016.

[160] Catherine MC Volgenant, Monique H van der Veen, Johannes J de Soet, and Jacob M ten Cate. Effect of metalloporphyrins on red autofluorescence from oral bacteria. *European Journal of Oral Sciences*, 121(3pt1):156–161, 2013.

[161] Bryan Vonasek, Alexander Kay, Tara Devezin, Jason M Bacha, Peter Kazembe, Dilsher Dhillon, Sandile Dlamini, Heather Haq, Lineo Thahane, Katie Simon, et al. Tuberculosis symptom screening for children and adolescents living with hiv in six high hiv/tb burden countries in africa. *Aids (london, England)*, 35(1):73, 2021.

[162] James Maxwell Glover Wilson, Gunnar Jungner, World Health Organization, et al. Principles and practice of screening for disease. 1968.

[163] Eric L Wisotzky, Benjamin Kossack, Florian C Uecker, Philipp Arens, Steffen Dommerich, Anna Hilsmann, and Peter Eisert. Validation of two techniques for intraoperative hyperspectral human tissue determination. In *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10951, page 109511Z. International Society for Optics and Photonics, 2019.

[164] Allen Wong, Paul E Subar, and Douglas A Young. Dental caries: an update on dental trends and therapy. *Advances in Pediatrics*, 64(1):307–330, 2017.

[165] SR Wood, J Kirkham, PD Marsh, RC Shore, B Nattress, and C Robinson. Architecture of intact natural human plaque biofilms studied by confocal laser scanning microscopy. *Journal of Dental Research*, 79(1):21–27, 2000.

[166] Hao Xue and Flora D Salim. Exploring self-supervised representation ensembles for covid-19 cough classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1944–1952, 2021.

[167] Christina Yoon, Fred C Semitala, Elly Atuhumuza, Jane Katende, Sandra Mwebe, Lucy Asege, Derek T Armstrong, Alfred O Andama, David W Dowdy, J Luke Davis, et al. Point-of-care c-reactive protein-based tuberculosis screening for people living with hiv: a diagnostic accuracy study. *The Lancet Infectious Diseases*, 17(12):1285–1292, 2017.

[168] Takuma Yoshitani, Masa Ogata, and Koji Yatani. Lumio: a plaque-aware toothbrush. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 605–615, 2016.

[169] Karen K Yuen. The two-sample trimmed t for unequal population variances. *Biometrika*, 61(1):165–170, 1974.

[170] Liqiang Zhang, Fengyu Su, Xiangxing Kong, Fred Lee, Kevin Day, Weimin Gao, Mary E Vecera, Jeremy M Sohr, Sean Buizer, Yanqing Tian, et al. Ratiometric fluorescent ph-sensitive polymers for high-throughput monitoring of extracellular ph. *RSC advances*, 6(52):46134–46142, 2016.

[171] Alexandra J Zimmer, César Ugarte-Gil, Rahul Pathri, Puneet Dewan, Devan Jaganath, Adithya Cattamanchi, Madhukar Pai, and Simon Grandjean Lapierre. Making cough count in tuberculosis care. *Communications medicine*, 2(1):83, 2022.

[172] Victor Zue and Ronald Cole. Experiments on spectrogram reading. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 116–119. IEEE, 1979.

# Appendix A

# TBSCREEN ADDITIONAL RESULTS

Table A.1: Linear regression model of clinical variables and cough counts. A linear regression model was used to examine whether the indicated variables were associated with cough counts. The relationship between GeneXpert semi-quantitative scaling (1-5) and cough count was significant, for every increase in test level, cough count increases by 22.

| **Variable** | **Coeff** | **Standard Error** | **z** | **P>z** | **95% Confidence Interval** | **Number of Observations** |
|---|---|---|---|---|---|---|
| Age | -0.32 | 0.93 | -0.34 | 0.73 | -2.14, 1.51 | 103 |
| Gender | -37.57 | 24.41 | -1.54 | 0.12 | -85.4, 10.3 | 103 |
| HIV Infection | -3.31 | 34.58 | -0.1 | 0.92 | -71.1, 64.5 | 99 |
| GeneXpert Semiquantitative grading | 22.80 | 9.69 | 2.35 | 0.02 | 3.8, 41.8 | 102 |
| Sputum smear | 17.64 | 10.83 | 1.63 | 0.10 | -3.6, 38.9 | 96 |
| Cavity Chest X-ray | 38.69 | 21.81 | 1.77 | 0.08 | -4.1, 81.4 | 103 |

Table A.2: De-Long test results for performance across datasets using different recording devices: $P < 0.05$ marked with * indicate significant difference between the ROC curves of models using Delong test.

| | **Model Training Parameters** | **Test Set** | **P - value** |
|---|---|---|---|
| TBscreen<br><br>Trained/Evaluated on coughs from one recording device | T1<br>Scalogram: 10 Hz – 4 KHz,<br>Sampling rate: 44.1 KHz | Smartphone\| Boundary Microphone | 2e-14* |
| | | Smartphone\| Condenser Microphone | 2e-16* |
| | | Condenser Microphone \| Boundary Microphone | 4e-11* |
| TBscreen<br><br>Trained/Evaluated on coughs from one recording device | T2<br>Scalogram: 10 Hz – 4 KHz,<br>Sampling rate: 44.1 KHz | Smartphone\| Boundary Microphone | 2e-16* |
| | | Smartphone\| Condenser Microphone | 2e-16* |
| | | Condenser Microphone \| Boundary Microphone | 5e-14* |
| TBscreen<br><br>Trained/Evaluated on coughs from one recording device | T3<br>Scalogram: 10 Hz – 4 KHz,<br>Sampling rate: 44.1 KHz | Smartphone\| Boundary Microphone | 0.2 |
| | | Smartphone\| Condenser Microphone | 1 |
| | | Condenser Microphone \| Boundary Microphone | 0.2 |

Table A.3: Statistical significance in comparing model performance: $P < 0.05$ marked with * indicate significant difference between the models using Delong test.

| | **Model Training Parameters** | **Test Set** | **P value** |
|---|---|---|---|
| Frequency range of scalogram<br><br>Device: All,<br><br>Sampling rate: 44.1 KHz | 10 Hz – 4 KHz \| 10 KHz – 16 KHz | T1: Subject Balanced CV | 0.8 |
| | 10 Hz – 4 KHz \| 4 KHz – 8 KHz | T1: Subject Balanced CV | 2e-16* |
| | 4 KHz – 8 KHz\|<br>10 KHz – 16 KHz | T1: Subject Balanced CV | 2e-16* |
| Sampling rate<br><br>Device: All<br><br>Scalogram: 10 Hz – 4KHz | 8 KHz\|<br>44.1 KHz | T1: Subject Balanced CV | 2e-16* |
| Recording device<br><br>Scalogram: 10 Hz -4KHz,<br><br>Sampling rate: 44.1 KHz | Smartphone \| Boundary Mic. | T1 subset: Subject Balanced CV (Smartphone\| Boundary Mic. coughs) | 2e-14* |
| | Smartphone \| Condenser Mic. | T1 subset: Subject Balanced CV (Smartphone\| Condenser Mic. coughs) | 2e-16* |
| | Boundary Mic.\| Condenser Mic. | T1 subset: Subject Balanced CV (Boundary Mic. coughs \| Condenser Mic. coughs) | 4e-11* |
| Alternative Frequency Representation<br><br>Mel-spectrogram<br><br>Device: All<br><br>Baseline Model | VGGish \| TBscreen | T1: Subject Balanced CV | 2e-16* |
| | ResNet18 \| TBscreen | T1: Subject Balanced CV | 2e-16* |
| | VGGish \| ResNet18 | T1: Subject Balanced CV | 6e-07* |

Table A.4: Statistical significance in comparing smartphone model performance for various sub-categories: $P < 0.05$ marked with * indicate significant difference between the models using Delong test.

| | **Category** | **Sub-Category** | **P-score** | **Coughs Control vs Case** | **Subjects Control vs Case** |
|---|---|---|---|---|---|
| Demographic Bias | Gender | Male \| Female | P = 2e-16* | (2930 vs. 2786) \| (1068 vs. 1270) | (27 vs 27) \| (17 vs 18) |
| | Age Group | [18,40] \| (40,60] | P = 0.01* | (2069 vs. 2394) \| (1068 vs. 1270) | (22 vs 22) \| (17 vs 17) |
| Clinical Bias | HIV History | No HIV history \| With HIV history | P = 8e-16* | (2429 vs. 3322) \| (1434 vs. 551) | (24 vs 38) \| (16 vs 5) |
| | Smoking History | No Smoking History \| With Smoking History | P = 2e-16* | (3376 vs. 2434) \| (512 vs. 1581) | (37 vs 32) \| (5 vs 12) |
| TB Presentation | GeneXpert | Low Bacterial Load \| High Bacterial Load | P = 2e-16* | (3998 vs. 858) \| (3998 vs. 3198) | (44 vs 16) \| (44 vs 29) |
| | Sputum Smear | Low Bacterial Load \| High Bacterial Load | P = 0.2 | (3998 vs. 631) \| (3998 vs. 3173) | (44 vs 13) \| (44 vs 28) |
| | Lung Cavity | No Cavity \| With Cavity | P = 2e-16* | (3998 vs. 905) \| (3998 vs. 3151) | (44 vs 18) \| (44 vs 27) |

# VITA

Manuja Sharma is a PhD candidate at University of Washington in the depeartment of Electrical and computer Engineering and co-advised by Dr. Eric J. Seibel and Dr. Shwetak N. Patel.