Comprehensive cough data analysis on *CODA TB*

Jyoti Yadav[1], Aparna S. Varde[1], Lei Xie[2,3]
1. School of Computing, Montclair State University, NJ 2. CS Faculty, CUNY Hunter, NY 3. Adjunct, Weill Cornell Medicine, NY
(yadavj2@montclair.edu, vardea@montclair.edu, lxi0003@hunter.cuny.edu), ORCID ID: 0000-0002-3170-2510 (Varde)

***Abstract*—This work leverages CODA TB, a groundbreaking dataset for a novel comprehensive method of early TB detection from medical big data. Departing from the erstwhile, we find mere cough duration less effective in TB prediction. We discover key demographic and clinical factors (e.g. heart rate, presenting symptoms) to be crucial in distinguishing TB cases, motivating comprehensive cough data analysis with enhanced screening.**

Keywords: Audio-visual analysis, medical big data, TB detection

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386805

I. INTRODUCTION

Tuberculosis (TB) poses a significant global health threat: around 10·6 million people became ill with TB in 2021, and around 1.5 million deaths occurred in 2020 [1]. Its impact extends beyond health, affecting economic development and disproportionately impacting vulnerable populations. Rise of drug-resistant strains adds complexity, emphasizing urgent need for TB eradication to save lives, reduce poverty, protect the vulnerable, and prevent the spread of drug-resistant forms [2]. A challenge in TB eradication is difficulty in identifying cases (~40% TB-affected people undiagnosed or unreported due to obstacles in accessing healthcare or a lack of testing / treatment [3]). Addressing this challenge needs affordable, non-invasive digital screening tools. Traditionally, cough has been a marker for TB cases & treatment. Recent advances in acoustic AI offer a scope to passively detect / monitor cough. Prior studies are limited in sample size & settings, motivating a need for more development of AI algorithms to accurately distinguish tubercular from non-tubercular coughs.

The CODA TB DREAM Challenge [4] presents a notable opportunity to advance cough-based TB diagnosis. It gathers data from individuals across 7 countries presenting with new or worsening cough for at least 2 weeks. Recorded coughs are collected using the *Hyfe Research App*, and participants undergo comprehensive TB evaluations. CODA TB releases the data to the public, inviting AI experts to develop and test algorithms predicting TB status from features extracted from elicited coughs. In our research, we use a substantial dataset to investigate the acoustic attributes of cough sounds for TB prediction. Our journey commences with the essential data operations, encompassing clinical & audio metadata loading, providing a solid foundation for analysis. With much focus on demographics and presenting symptoms, we uncover noteworthy differences in TB+ & TB- subjects. Notably, we observe variances in symptoms such as weight loss, fever, night sweats & hemoptysis. Our study highlights a potential utility of combining audio data & demographics for accurate TB detection, promising major contributions to the field.

II. MEDICAL DATA AND EXPLORATORY ANALYSIS

The CODA TB data is from health centers across 7 countries: India, Philippines, South Africa, Uganda, Vietnam, Tanzania, Madagascar). Clinical investigation encompasses individuals 18 or older who seek assistance at outpatient health centers; specifically, those with a new or worsening cough persisting for at least 2 weeks. In this process, a survey is administered during the 1st visit to obtain basic demographic & clinical data from participants. Simultaneously, sputum samples are collected for TB testing. As an integral aspect of the study protocol, participants are prompted to cough, and the ensuing cough sounds are recorded. Cough sounds identified by the *Hyfe* cough prediction algorithm are included for subsequent analysis. It is crucial that the number of solicited coughs vary for each participant based on how frequently they cough in each 5-second recording interval. Additionally, the act of producing a solicited cough can trigger more coughing, giving data with a blend of solicited and spontaneous coughs.

TABLE I. STATISTICAL OVERVIEW OF CODATB SOLICITED DATASET

Features	*TB+*	*TB-*	*Total*
Participants	297	808	1105
Total coughs	2930	6842	9772
Avg. no. of coughs / participant	10.06	8.65	9.03
Min. no, of coughs / participant	3	3	-
Max. no of coughs / participant	50	37	-
Total duration of coughs (minutes)	24.41	57.01	81.43

Our analytical models encompassing visual analytics and descriptive statistics include demographic variables, e.g. age & sex, adhering to microbiological standards in TB detection. See Table I for some statistics. It entails cough audio features and clinical data. Further, Table II showcases a compilation of demographic & clinical metadata employed as features in a Cough+Metadata experiment, i.e. BMI: body mass index, P-TB: (pulmonary TB), EP-TB: (extrapulmonary TB), Solicited Coughs - recorded at clinic, Longitudinal Coughs - subjects given phones to self-record coughs for 2 weeks.

III. RESULTS AND BIOMEDICAL SIGNIFICANCE

We introduce a holistic methodology that encompasses a TB detection pipeline. It involves: extraction of cough signals; construction of a Mel spectrogram; implementation of sound event detection; extraction of pertinent features; and finally, assignment of cough classification. It is illustrated in Fig. 1.

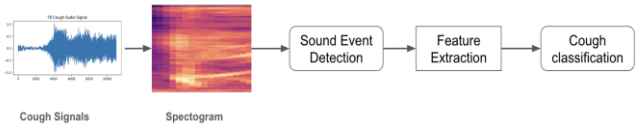

Fig. 1. Proposed Method: Comprehensive Pipeline for TB Detection

Our initial analysis reveals significant insights as follows: *Comparable Demographics:* TB and non-TB cough audio can be recorded by generating Mel spectrograms in decibels and displaying the associated raw audio signals (Fig. 2). *Distinctive Symptoms:* Notably, key presenting symptoms (weight loss, fever, night sweats, and hemoptysis) show differences in cohorts, hence offering diagnostic potential. *Heart Rate Indicator* ($p < 0.05$): TB-positive subjects display higher heart rates, making it a biomarker for TB (See Fig 3). *Body Temperature:* This remains similar across both groups.

Subject-reported cough duration: This poses problems - high standard deviation *(p>0.05).* infer that cough duration alone may not be a reliable factor to detect TB. Our findings (e.g. Fig. 3, 4) illuminate major traits of TB+ & TB- subjects.

TABLE II. DEMOGRAPHIC FEATURES IN COUGH+METADATA EXPERIMENT

Participants Demographics	**TB Negative**	**TB Positive**
Age in Years		
Mean±SD	42.06 ±15.28	37.55 ±14.85
Range	18-85	18-83
Sex		
Male	393(49%)	195(49%)
Female	415(51%)	202(51%)
Anthropometrics		
Height (CM)	160.99±8.79	163.80 ±8.49
Weight (KG)	59.84±14.41	51.84±9.24
BMI (KG/M^2)	23.1	19.3
Heart Rate	82.94±14.27	94.95±19.61
Temperature ©	36.64±0.46	36.96±0.66
Prior Illness		
Prior TB Exposure	151(19%)	48(16%)
P-TB Diagnosis	136(17%)	44(15%)
EP-TB Diagnosis	13(2%)	4(1%)
Presenting Symptoms		
Weight Loss	397(49%)	228(77%)
Fever	298(37%)	199(67%)
Night Sweats	295(37%)	189(62%)
Hemoptysis	84(10%)	64(22%)
Cough Duration / Day (SD)		
Reported at presentation	44.73±56.74	53.29±49.51
Cough Audio(n)		
Solicited Cough	6,842	2,930
Longitudinal Cough	274,145	440,777

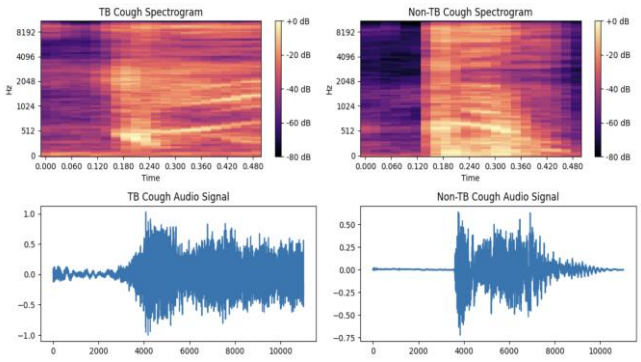

Fig 2. Audio: Mel-Frequency Cepstral Coefficients; cough sounds (TB+/-)

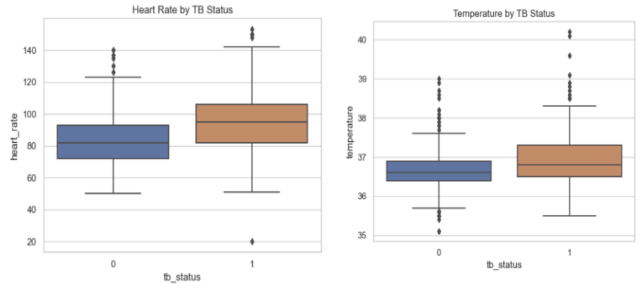

Fig 3. Box plots depicting correlations of clinical factors with TB status

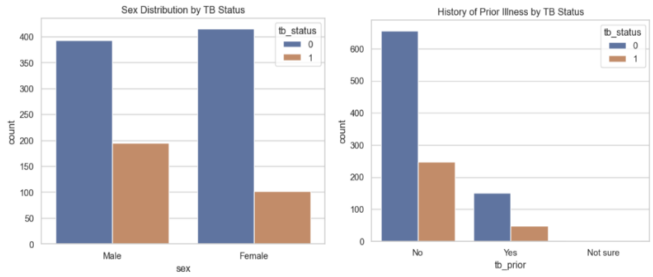

Fig 4. Bar charts portraying correlations of demographic factors and TB

We display a 2x2 grid of subplots (e.g. Fig. 5) with Log-Mel Spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) for a TB patient versus healthy individual. These visualizations help fathom the frequency content and acoustic characteristics of cough sounds, aiding analysis of potential differences. They also extract mean and standard deviation features from TB+, TB- spectrograms. Conventional machine learning models that necessitate vectorized inputs require a series of statistical operations that we implement on Low-Level Descriptors (LLD) to condense features derived from all frames of the audio signal. This is described next.

Let *m={x1, x2, x3, ⋯, xN }* be a sample LLD of N values. Hence, we have:

1. *Mean:* measures average value of LLD

$$\hat{x} = \frac{1}{N}\sum_{N=1}^{N} \quad X_n$$

2. *Standard Deviation:* measures amount of variation or dispersion of LLD

$$s = \sqrt{\frac{1}{N-1}\sum_{N=1}^{N} \quad (x_m - \hat{x})^{\ 2}}$$

3. *Skewness:* measures asymmetry of sample distribution of LLD & mean

$$b_1 = \frac{\frac{1}{N}\sum_{n=1}^{N} \quad (x_n - \hat{x})^{\ 3}}{\left[\frac{1}{N-1}\sum_{N=1}^{n} \quad (x_n - \hat{x})^{\ 2}\right]^{\frac{3}{2}}}$$

These initial findings illuminate unique traits of TB+ and TB- subjects. Hence, they offer valuable insights to enhance TB diagnosis and advance biomedical research.

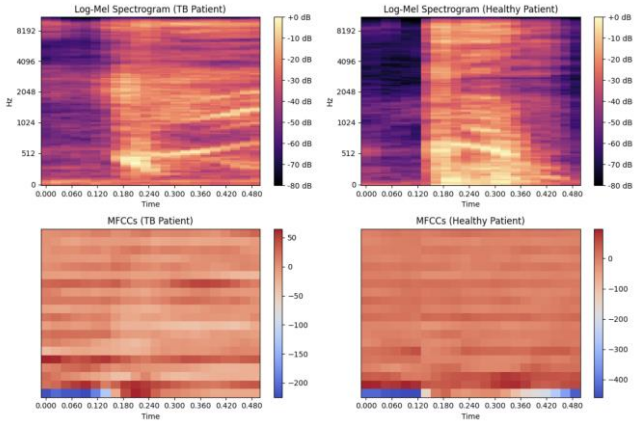

Fig 5. Log-Mel Spectrograms and MFCCs for TB+ and TB- cases

IV. Related Work

As per research in medical data analysis for TB diagnosis, significant strides have been made via prior studies e.g. [1-5]. This is line with some of our own work on medical big data, e.g. [6, 7] pertaining to Covid-19, and [8, 9] for other health informatics areas. Some research using big data in TB studies [10] explores audio analysis using Support Vector Machines (SVM); while other work [11] aims to classify audio features with clinical data. Such studies offer very good results, yet they present a potential for further enhancement.

Thriving on many success stories, including some by our own research groups, we propose a comprehensive method in this paper identifying key demographic and clinical factors for automation of early TB detection, thus challenging traditional reliance on mere cough duration. Out initial work hereby demonstrates much scope to enhance TB detection in medical big data analytics on the whole. Addressing limitations and pushing integration boundaries, our work takes a small step towards a possible new direction in TB diagnostics. Beyond numerical accuracy per se, it offers a holistic understanding, thus contributing to more practical TB diagnostic methods.

V. Conclusions and Roadmap

A pivotal revelation from our study is the inadequacy of relying solely on subject-reported cough-duration as a major determinant in TB detection, due to high standard deviation. Our initial findings advocate for a more comprehensive approach integrating clinical, demographic & audio data, thus helping to offer more effective diagnosis. To advance these insights, our roadmap entails the use of Convolutional Neural Networks (CNN) or transformer-based models for deeper analysis, as well as the utilization of explainable models, e.g. decision trees, for better interpretability.

Moving forward, we aim to validate the findings on diverse datasets for wider applicability & reliability. Future work in this area can entail the development of remote TB screening applications, leveraging insights gained form studies such as ours in this paper. In earlier work, we have developed apps to help in the broad realm of Covid-19 and related work, e.g. [12, 13] and conducted research in scientific data analysis with broader impacts on health and well-being, e.g. [14-17]. Motivated by that, the step we take here aims to translate the research into practical solutions for more accessible and widespread TB screening, especially in regions with limited healthcare infrastructure. Through such strategic initiatives, the big data community can make contributions to the evolution of TB diagnostics and global efforts in combating TB via more comprehensive medical big data analysis.

Acknowledgments

Jyoti Yadav is a Graduate Assistant (GA) at Montclair State University, NJ. Dr. Aparna Varde acknowledges NSF grant 2018575. She is an Associate Director, Clean Energy & Sustainability Analytics Center, Montclair, NJ. Dr. Lei Xie heads the Precision Drug Discovery Lab at CUNY, Hunter, NY, as Full Professor. He is also an Adjunct Professor, Neuroscience, at Weill Cornell Medical College, Cornell University, NY. The datasets used for the analyses described were contributed by Dr. Adithya Cattamanchi at UCSF and Dr. Simon Grandjean Lapierre at University of Montreal and were generated in collaboration with researchers at Stellenbosch University (PI Grant Theron), Walimu (PIs William Worodria and Alfred Andama); De La Salle Medical and Health Sciences Institute (PI Charles Yu), Vietnam National Tuberculosis Program (PI Nguyen Viet Nhung), Christian Medical College (PI DJ Christopher), Centre Infectiologie Charles Mérieux Madagascar (PIs Mihaja Raberahona & Rivonirina Rakotoarivelo), and Ifakara Health Institute (PIs Issa Lyimo & Omar Lweno) with funding from the U.S. National Institutes of Health (U01 AI152087), The Patrick J. McGovern Foundation and Global Health Labs. They were obtained as part of the COugh Diagnostic Algorithm for Tuberculosis (CODA TB) DREAM Challenge DREAM Challenge through Synapse [syn31472953].

References

[1] Ss. Bagcchi, (2023). WHO's global tuberculosis report 2022. The Lancet Microbe, 4(1), e20.

[2] A. Matteelli, A Rendon, S. Tiberi, S. Al-Abri, C. Voniatis, A. Carvalho,& G. B. Migliori (2018). Tuberculosis elimination: where are we now?. *European Respiratory Review*, 27(148).

[3] R. G. Loudon, & S. K. Spohn, (1969). Cough frequency and infectivity in patients with pulmonary tuberculosis. *American Review of Respiratory* Disease, 99(1), 109-111.

[4] G. P. Kafentzis, S. Tetsing, J. Brew, L. Jover, M. Galvosas, C. Chaccour, C., & P. Small (2023). Predicting Tuberculosis from Real-World Cough Audio Recordings and Metadata. *arXiv preprint arXiv:2307.04842*.

[5] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, & T. Niesler (2021). Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, 42(10), 105014.

[6] A. S. Varde, D. Karthikeyan, & W. Wang (2023). Facilitating COVID recognition from X-rays with computer vision models and transfer learning. *Multimedia Tools and Applications (MTAP) Journal, Springer,* pp. 1-32.

[7] M. Puri, Z. Dau, & A.S. Varde (2021). COVID and social media: Analysis of COVID-19 and social media trends for smart living and healthcare. ACM SIGWEB, (2021 Autumn), Article 5, pp. 1-20.

[8] R. Hidalgo, A. DeVito, N. Salah, A.S. Varde, A. S., R.W. Meredith (2022). Inferring Phylogenetic Relationships using the Smith-Waterman Algorithm and Hierarchical Clustering. *IEEE Big Data*, pp. 5910-5914.

[9] X. Du, O. Emebo, A. Varde, N. Tandon, S. N. Chowdhury & G. Weikum, (2016). Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning. *IEEE ICDE (W),* pp. 54-59.

[10] L. Xie, E. Draizen, & P. Bourne, (2017). Harnessing big data for systems pharmacology. *Annual Review of Pharmacology & Toxicology journal*, 57, 245-262.

[11] Y. Liu, Y. Wu, X. Shen, L. Xie (2021). COVID-19 multi-targeted drug repurposing using few-shot learning. *Frontiers in Bioinformatics*, 1, 69317

[12] J. Torres, V. Anu, A. S. Varde and C. Duran (2021), My-Covid-Safe-Town: A mobile application to support post-Covid recovery of small local businesses. *IEEE IEMTRONICS* pp. 1-7.

[13] C. Varghese, D. Pathak & A. S. Varde (2021). SeVa: A Food Donation App for Smart Living, *IEEE Computing and Communication Workshop and Conf. (CCWC)*, pp. 408-413, doi: 10.1109/CCWC51732.2021.9375945.

[14] F. M. Suchanek, A. S. Varde, R. Nayak, & P. Senellart, P. (2011). The hidden Web, XML and the semantic Web: Scientific data management perspectives. *ACM EDBT Intl Conf. Extending Database Tech.* pp. 534-537.

[15] M. J. Pawlish, & A. S. Varde (2010). A decision support system for green data centers. *PIKM (workshop on Ph. D. students) in ACM CIKM (Conference on Information and Knowledge Management),* pp. 47-56.

[16] J. Tancer & A. S. Varde (2011). The Deployment of MML for Data Analytics over the Cloud. *IEEE 11th International Conference on Data Mining,* workshops, pp. 188-195.

[17] A. Singh, J. Yadav, S. Shrestha, & A.S. Varde (2023). Linking Alternative Fuel Vehicles Adoption with Socioeconomic Status and Air Quality Index. *AAAI Conference*, AISG workshop, arXiv:2303.08286.