ISSN: 2754-4761

Journal of Pulmonology Research & Reports

SCIENTIFIC Research and Community

Research Article

Open Access

"TimBre" Pilot Study Conducted Using Multi-Country Training and Validation Data for Screening of Pulmonary Tuberculosis Using Cough (Acoustic Sounds), Clinical & Demographic Inputs

Rahul Pathri*, Shekhar Jha, Ram Samudrala and Aparna Sykam

Docturnal Private Limited

ABSTRACT

TimBre from Docturnal offers screening for multiple lung diseases – Pulmonary Tuberculosis, Pneumonia, Covid19 & COPD. Detailed studies of TimBre in the past used a third-party Microphone Array that focused on a XY arrangement that provided high fidelity cough sounds with an average length of >5 seconds and demographic data such as Height, Weight, BMI [1]. In the current study, cough sounds were collected from 7 different countries (India, Vietnam, Philippines, Uganda, Tanzania, Madagascar, and South Africa) using Mobile Phones from different manufacturers & recorded solicited coughs in a clinic for a duration of 0.5 seconds. A plethora of demographic and clinical variables were provided of which a subset was used by TimBre algorithm. Most importantly, the .WAV files were recorded in a single channel at a sampling rate of 44.1kHz & 16 bits. The study details two approaches wherein the first method was to concatenate all the 0.5 second WAV files based on a timestamp provided for each StudyID in the training & scoring set while the second method involved using the 0.5 second snippets as-is in both training and validation sets without any concatenation. Both approaches used a combination of demographic, clinical and spectral variables. The first approach on the independent test set yielded a sensitivity and specificity of 68.6% and 71.7% respectively with an AUC of 0.75 while the second approach yielded a sensitivity & specificity of 75.41% and 68.30% respectively with an AUC of 0.78.

Thus, the ML model performed better in the second approach and we anticipate it to improve with additional training data.

***Corresponding author**

Rahul Pathri, Docturnal Private Limited, India. Tel: 91 9676721717.

Received: August 21, 2023; **Accepted:** August 24, 2023, **Published:** August 25, 2023

Keywords: TimBre, Tuberculosis, COPD, Pneumonia, COVID-19

Abbreviations

UCSF R2D2	University of San Francisco Rapid Research Development & Diagnostics
COPD	Chronic Obstructive Pulmonary Disorder
AUC / ROC	Area Under Curve / Receiver Operating Characteristic Curve
.WAV	Wave Audio File Format
TB	Tuberculosis
EMR	Electronic Medical Record
XY	An arrangement of Microphone Arrays on a Microphone
BMI	Body Mass Index
HIV	Human Immunodeficiency Viruses
PDP	Partial Dependence Plot
MRMR	Maximum Relevance Minimum Redundancy
RUS	Random Under Sampling
FFT	Fast Fourier Transformation
ML / DL	Machine Learning / Deep Learning
CNN	Convolutional Neural Network
MFCC	Mel Frequency Cepstral Coefficient
Hz	Hertz
AWS	Amazon Web Services
XAI	Explainable Artificial Intelligence
AI	Artificial Intelligence
HR	Heart Rate
CI	Confidence Interval

Introduction

The CODA TB dream challenge entailed various global teams participating in the challenge to predict Pulmonary Tuberculosis that were provisioned with clinical, demographic and cough data sets for both training and testing [4]. With a wide variety of mobile phones and external microphone Arrays, cough as a biomarker has gained significant in the recent past with the advent of Machine Learning and Deep Learning techniques not only for Pulmonary Tuberculosis but also other ailments such as Asthma, COPD & Lung Cancer. This is now suggested to be a useful triage tool and also treatment monitoring tool.

TB is the leading cause of mortality from an infectious disease globally. This mortality in part is driven by a large diagnostic gap, in which only about 40% of the estimated cases are diagnosed or reported. New low-cost, non-invasive diagnostics and triage tools are needed to increase TB case detection and initiate treatment. TB a communicable disease caused by Mycobacterium tuberculosis,

is a major cause of ill health and one of the leading causes of death worldwide. Until the COVID-19 pandemic, TB was the leading cause of death from a single infectious agent, ranking even above HIV/AIDS. In 2020, an estimated 9.9 million people fell ill with TB and 1.3 million died of TB worldwide. However, approximately 40% of people with TB were not diagnosed or reported to public health authorities because of challenges in accessing health facilities or failure to be tested or treated when they do. The development of low-cost, non-invasive digital screening tools may improve some of the gaps in diagnosis. As cough is a common symptom of TB, it has the potential to be used as a biomarker for screening of disease. Several previous studies have demonstrated the potential for cough sounds to be used to screen for TB, though these were typically done in small samples or limited settings. Further development and evaluation are critical to move the field forward" – Source Coda TB dream challenge [1-4].

Materials and Methods

Here we leverage data collected from adults 18 years and older who presented to clinics across 7 countries with new or worsening cough for at least 2 weeks. Elicited coughs were recorded using the Hyfe Research app. Individuals were then comprehensively evaluated for TB with sputum-based molecular (Xpert MTB/RIF Ultra) and culture (MGIT or Lowenstein-Jensen) testing. We developed 2 models, one in which the sounds are concatenated (per participant) and then the other where the sounds were analyzed separately (per cough).

Per Participant Model

The per participant model involved concatenating the lossless .WAV files based on their timestamps and then features were extracted including Spectral Centroid, Spectral Flatness, Spectral Skewness, Coefficient of Variance, Top 10 Amplitudes, Energy, MFCC coefficients across a multitude of frequency bands.

In this model, a total of 10787 WAV files in the training set were mapped against 1269 participants in the first approach resulting in 1231 concatenated .WAV files implying that 38 participants were included in the training set using clinical and demographic variables only (without .WAV files). The clinical variables added additional clinical information that were not used in our earlier studies for lack of accurate information (EMR typically) to avoid subjectivity [1]. However, the below model included/added Prior TB due to its critical nature to contribute to the model in determining drug resistant TB and hemoptysis as a potential indicator of a more severe clinical presentation.

Clinical & Demographic Variables used

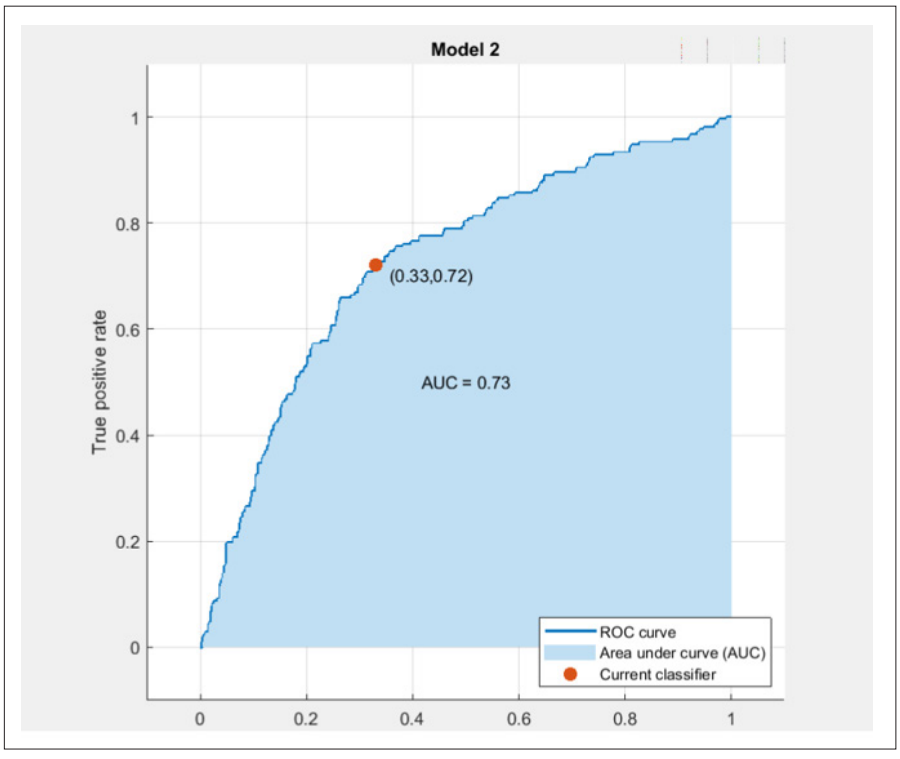
'Gender', 'Age', 'Height', 'Weight', 'BMI', 'PriorTB', 'Hemoptysis'

Per Participant Model (Table 1)

	Total Participants (Meta Data)	Participants with Cough (.WAV)	.WAV Files before Concatenation	.WAV Files Post Concatenation	Sensitivity
Train	1269	1245	10787	1231	72%
TEST (Prediction)	1276	1248	10008	1248	68.6% (65.48%, 71.61%)

Best specificity at 90% sensitivity threshold was calculated as 39.30%

ROC of the Model



Per Cough Model

In the per cough model, the lossless .WAV files were used as-is without any concatenation in that it created redundancy of clinical/demographic information that facilitated a higher model accuracy. A studyID with multiple cough file snippets was repeated across with similar demographic and clinical information & the spectral feature extraction and their inclusion in the Model remained constant across all the models.

In this model, 10,008 .WAV files were predicted using a model created out of 10787 .WAV files. The clinical variables added additional clinical information that were not used in our earlier studies for lack of accurate information (EMR typically) to avoid subjectivity [1]. However, the below model included/added Prior TB, Hemoptysis, HIV status, Night sweats, Heart rate, Fever & Temperature which are all data available in the context of routine TB care in high burden countries.

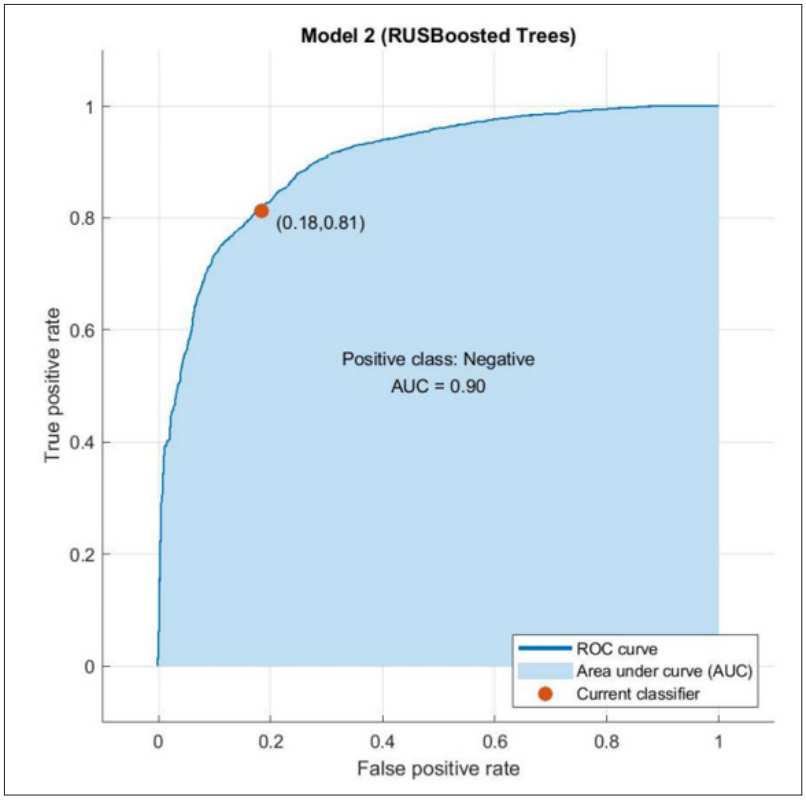
Clinical & Demographic Variables used

'Gender', 'Age', 'Height', 'Weight', 'BMI', 'HIVstatus', 'PriorTB', 'Hemoptysis', 'Fever', 'Nightsweats', 'Heartrate', 'Temperature'

Per Cough Model (Table 2)

	Total Participants (Meta Data)	Participants with cough (.WAV)	.WAV files before concatenation	.WAV files post concatenation	Sensitivity	Specificity	AUC
TRAINING	1269	1245	10383	NA	81%	81%	0.90
TEST (Prediction)	1191	1248	10008	NA	75.41% (74.40%, 76.40%)	68.30% (66.48%, 70.07%)	0.78 (0.77, 0.79)

ROC of the Model



Another Champion Ensemble Model Yielded the Following Results:

	Value	CI
Sensitivity	86.47%	85.66%, 87.25%
Specificity	46.72%	44.80%, 48.65%
AUC	0.7565	0.7456, 0.7673

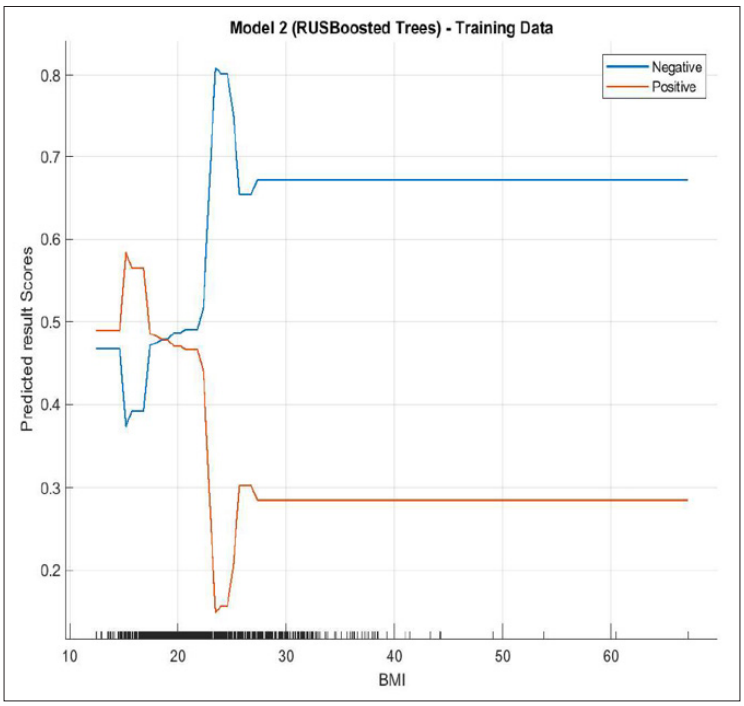
Best specificity at 90% sensitivity threshold: 37.26%

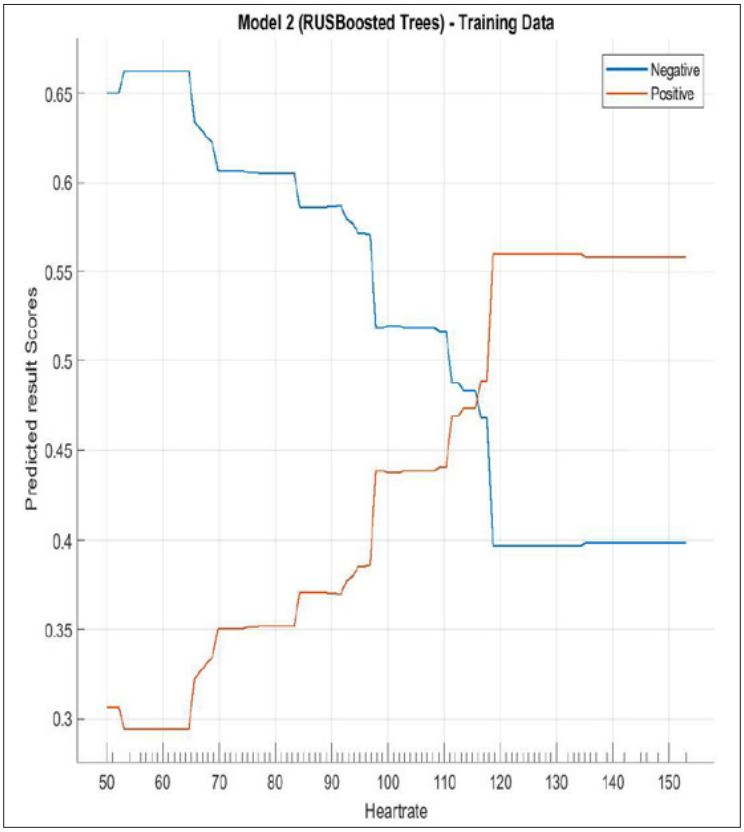
While the sensitivity is much higher than the first model that used per cough approach, specificity is on the lower side and hence we plan to add additional training data from the longitudinal data set to see how this model can be fine-tuned to an acceptable specificity of 70% and above.

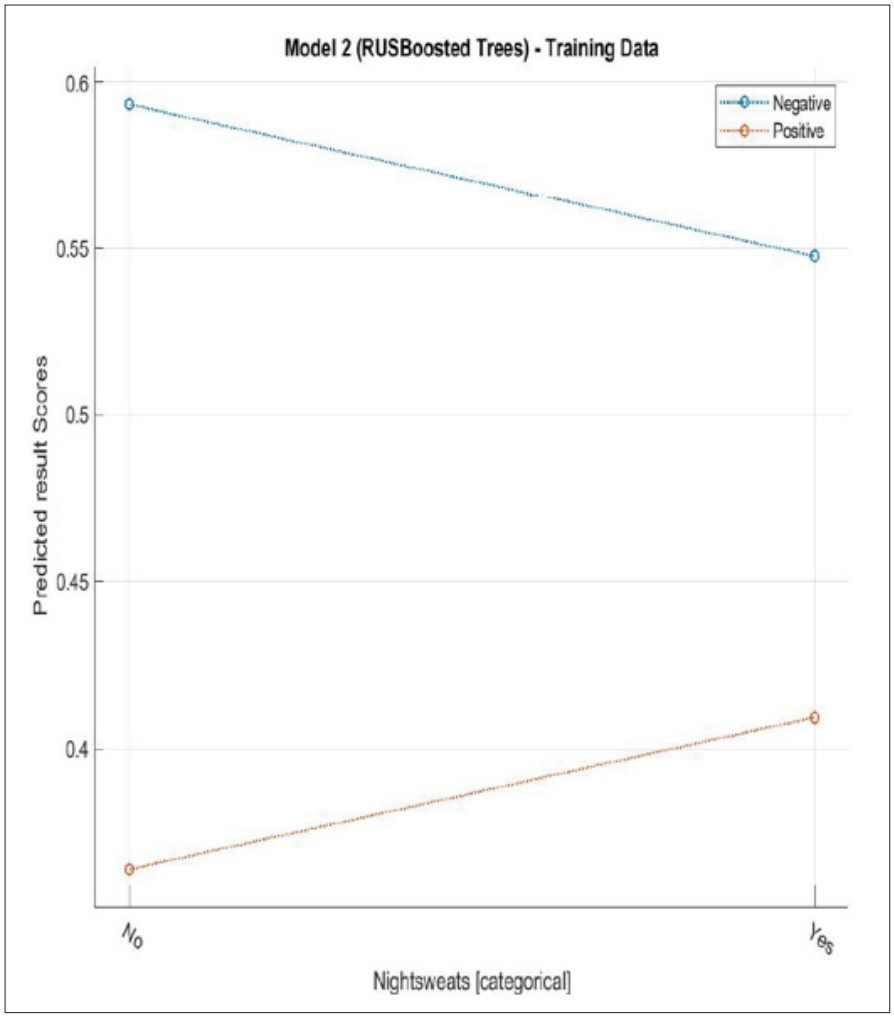
Interpretability & Explainability

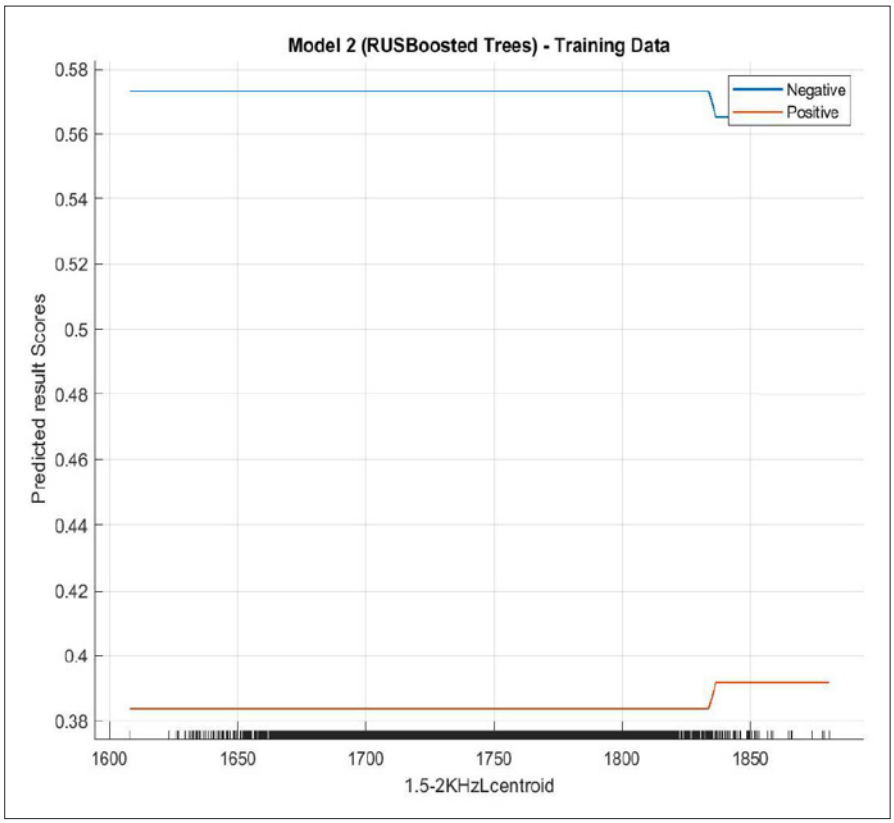
1. Both the supervised learning models used RUSBoosted Ensemble technique given the fact that the TB (Positive) and TB (Negative) were to the tune of 27% and 72% respectively in the provisioned training datasets representing an imbalanced class set. The feature selection used Kruskal-Wallis Algorithm given its distribution and outlier agnostic nature in comparison with Chi-square, Annova & MRMR. The models were built out of the box using MATLAB R2022a Classification Learner from MathWorks [6]
2. Explainability & Interpretability of the per-cough model that used 10-fold Cross Validation revealed BMI, Heart Rate, Night Sweats and Spectral Centroid to be the differentiators in segregating classes as seen below (Box1 - PDP) [10]. We presume that additional training data shall reveal more such patterns & we shall share the entire list of contributing spectral variables in Phase-2 of the study

Box1 – PDP (Matlab R2023a)






Model 2 (RUSBoosted Trees) - Training Data
Negative
Positive
Predicted result Scores
0.6
0.55
0.5
0.45
0.4
No
Yes
Nightsweats [categorical]


Model 2 (RUSBoosted Trees) - Training Data
Negative
Positive
Predicted result Scores
0.58
0.56
0.54
0.52
0.5
0.48
0.46
0.44
0.42
0.4
0.38
1600
1650
1700
1750
1800
1850
1.5-2KHzLcentroid

In the context of Pulmonary TB screening, the interpretability as depicted in the Box1-PDP (Partial Dependence Plot) above is a useful tool & can be interpreted as below:

a) **BMI:** Any value greater than 22 starts reducing the probability score from 0.8 to 0.65 for TB Negative labels as seen from the PDP. The outliers are left intentionally & distinct values are depicted as bars on the X-Axis
b) **Heart Rate:** HR of 110 and above depicts an increase in the probability score for a TB Positive label
c) **Night Sweats:** Has a slight increase in the Positive labels as seen from the probability score and inversely, a dip is observed for the Negative labels moving from No to Yes value
d) **Spectral Centroid:** Depicts a slight dip in the probability score for Negative labels and for distinct values on the X-Axis, there is a slight increase for the Positive labels

Note: PDPs are generated on the TRAINING set only & we are yet to receive labels for the TESTING set

Discussion

The approach to use Random Under Sampling Boosted Models (Ensemble – RUS) on top of features extracted post FFT has provided a great deal of explainability and interpretability [7- 8]. The fact that MFCC components were already used obviates the need for CNN models relying on Spectrograms there by retaining the explainability & interpretability [10]. Most importantly, the contrastive AI approach (repeatability) wherein one can clearly understand what variable values in the TESTING set shall alter the prediction result [9]. Example: altering the standard deviation for "spectral centroid" or a "BMI" value.

The feature extraction used variables spread across different bands ranging from 0 to 5000Hz. Early data sets received during the last quarter of 2022 observed spurious explainability for Spectral Centroid in the 200Hz band. A constant effort has been made to analyze subsequent bands while not undermining the importance of low frequency components. Models used with and without demographic & clinical variables on top of spectral features were evaluated and the inclusion of clinical & demographic variables always improved the accuracy. This has been the case from earlier pilots as well [1]. However, having this information under the guidance of a Physician or an EMR is of extreme importance instead of Subject providing this information that introduces some ambiguity as experienced in out earlier studies [1]. Since the study was blinded, we did not get enough information about how the data was collected & what to exclude to avoid subjectivity which otherwise shall perturb the algorithm. We presume that the clinical and demographic information was obtained from the participants.

It is observed in the Coda TB dream challenge, participants had access to Longitudinal cough files & used a combination of CNN and other Ensemble models [4]. Once the Longitudinal cough data is made public, we would like to append the same to existing Ensemble models & also explore CNN models.

Limitations

1. The study used Solicited coughs to build the training model. Additional 700,000 longitudinal cough files may improve the accuracy & explainability.
2. A single channel mono WAV file could be replaced with a dual channel configuration for a more breadth of information.
3. Mobile Phone make/models were unknown to determine uniformity across the results between training and testing files in that whether they were homogenous or mixed models.

Data Acknowledgement

1. "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 Merieux 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."
2. Solicited Training and Scoring data .WAV files used Hyfe.ai app - https://www.hyfe.ai/ [5].
3. We sincerely thank Dr. Devan Jaganath (University of California San Francisco) for analysis of the VALIDATION set results.
4. We sincerely thank Dr. Sophie Huddart for providing the calculated TEST results across multiple iterations based on posterior probability scores provided by TimBre algorithm.

Next Steps

1) Sage bionetworks shall open the Validation process that shall open up additional Longitudinal data for 714,922 wav files to be included in the ML/DL Model.
2) Explore the Maximum (Majority) rule for per cough model. Note that the current champion Ensemble model (RUS boosted) yielded a consistent result for each StudyId while other models had a mixed response (positive & negative) thus mandating the Majority rule
3) Explore a combination of 1: Many (Concatenated-StudyId : individual-coughs) & Many:1 (individual-coughs : Concatenated-StudyId) approach for the validation & testing sets respectively.

Funding & Support

1) "Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health through award U01AI152087"
2) For Montreal Partner Sites: "Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health through award U01AI152087, The Patrick J. McGovern Foundation and the Global Health Labs"
3) For Docturnal Private Limited, India, funding support was provided by "FICCI / Millennium Alliance Covid19 Challenge (MA/R6/SA/2020/0042)"

References

1. Pathri R, Jha S, Tandon S, Ganga Shetty S (2022) Acoustic epidemiology of pulmonary tuberculosis (TB) & Covid19 leveraging AI/ML. J Pulmonol Res Reports 4: 1-6.
2. Pahar M, Klopper M, Reeve B, Theron G, Warren R, et al. (2021) Automatic cough classification for tuberculosis screening in a real-world environment. Physiol. Meas 42: 105014.
3. Botha G H R, Theron G, Warren R M, Klopper M, Dheda K, et al. (2018) Detection of tuberculosis by automatic cough sound analysis. Physiol. Meas 39: 045005.

4. Coda TB dream challenge - CODA TB DREAM Challenge - syn31472953 - Tables (synapse.org) (sign in required as per https://sagebionetworks.org)
5. Training and Scoring data of 0.5 second snippet .WAV files used - https://www.hyfe.ai/
6. MathWorks – MATLAB Classification Learner R2022a - Train models to classify data using supervised machine learning - MATLAB - MathWorks India
7. Explainable AI (XAI) - 2102.06518.pdf (arxiv.org)
8. XAI – Explainability Vs Interpretability - presentation7.pdf (wisc.edu)
9. Contrastive AI - 1802.07623.pdf (arxiv.org)
10. AWS White Paper- Interpretability versus explainability - Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions (amazon.com)