2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871393

Implications of clinical variability on computer-aided lung auscultation classification

Annapurna Kala[1], Eric D McCollum[2], and Mounya Elhilali[1]

Abstract— **Thanks to recent advances in digital stethoscopes and rapid adoption of deep learning techniques, there has been tremendous progress in the field of Computerized Auscultation Analysis (CAA). Despite these promising leaps, the deployment of these technologies in real-world applications remains limited due to inherent challenges with properly interpreting clinical data, particularly auscultations. One of the limiting factors is the inherent ambiguity that comes with variability in clinical opinion, even from highly trained experts. The lack of unanimity in expert opinions is often ignored in developing machine learning techniques to automatically screen normal from abnormal lung signals, with most algorithms being developed and tested on highly curated datasets. To better understand the potential pitfalls this selective analysis could cause in deployment, the current work explores the impact of clinical opinion variability on algorithms to detect adventitious patterns in lung sounds when trained on gold-standard data. The study shows that uncertainty in clinical opinion introduces far more variability and performance drop than dissidence in expert judgments. The study also explores the feasibility of automatically flagging auscultation signals based on their estimated uncertainty, thereby recommending further reassessment as well as improving computer-aided analysis.**

I. INTRODUCTION

While deep learning techniques are rapidly progressing with successful use-cases in speech recognition and computer vision, their advances in the medical domain are still challenging due to the limitations of data collected, the need of precise clinical assessments and the complexity of the phenomena with a much higher real-world cost associated with the task. In the case of pulmonary infections, breathing sounds have been used for centuries as an accessible medium to evaluate lung abnormalities. More recently, computer-aided auscultation analysis has become popular with the advancement in digital stethoscopes. Several studies in the recent literature have tackled the problem of detecting pulmonary pathological indicators posing the task as a classification problem to distinguish healthy from abnormal cases [1], [2]. Despite notable successes in this area, the problem of computer-aided auscultation diagnosis is far from resolved. Applications of machine learning and classification techniques are faced with major hurdles pertaining to clinical variability and label noise. Specifically, pulmonary screening is often prone to disagreements and expert-opinion variability. This in turn creates a notable degree of inter-reviewer confusion in auscultation signals both in terms of soft and hard labels, i.e the type of lung sound present in the signal and its temporal location in case of adventitious patterns [3]. Moreover, lung sounds, especially those acquired in non-ideal clinical environments, are themselves prone to a great degree of noise and ambient maskers. The similarities in the time-frequency patterns between ambient noise and adventitious lung sounds create a second level of ambiguity that further complicates the problem of computer-aided auscultation diagnosis.

These obstacles remain a major hurdle for wider deployment of automated auscultation diagnosis techniques. Unfortunately, there is little effort in the literature to explore the true impact of these clinical variability barriers. The current state-of-the-art techniques with notable success in lung sound classification take advantage of methodologies previously developed and well-established in speech and image technologies. MFCCs (Mel-Frequency Cepstral Coefficients), features designed for Automatic Speech Recognition inspired by human auditory perception, are used in [4] for wheeze detection. Multi-resolution methods based on wavelet transform [5] were deployed for detecting abnormal lung sounds. A convolutional neural network was used as a classifier of normal and abnormal lung sounds with the auditory spectrogram as inputs which performed on par with support vector machine algorithm on MFCC features in [6]. A denoising autoencoder was used to extract unsupervised features and best suited among these for crackle and wheeze were selected correspondingly in [7]. Further, two different support vector machines were trained for the classification of wheezes and crackles from normal lung sounds separately.

In recent years, several machine learning techniques have achieved great performance metrics in terms of abnormal lung sound detection. [8] reported an accuracy of 98.62 % proposing a multi-level strategy for classifying wheezes. They positioned the breathing cycle as a preprocessing and enhancing the wheezing features within the estimated breathing cycle before passing it to a classifier. Along similar lines, [9] extracted orthogonal non-negative matrix factorization bases discriminating normal-wheeze and emphasized wheezing frames in a recursive fashion before deciding on the presence or absence of wheeze. They reported an accuracy of 98.2%. Another work [10] employed a CNN architecture for a 7-class classification of lung sounds and compared its performance with Support Vector Machine, Gaussian Mixture Models and K-Nearest Neighbour classifiers on MFCCs statistics-based hand-picked features and

[1]Johns Hopkins University, Department of Electrical and Computer Engineering
[2]Johns Hopkins School of Medicine, Global Program of Pediatric Respiratory Sciences,
Eudowood Division of Pediatric Respiratory Sciences, Department of Pediatrics

Local Binary Pattern(LBP) features extracted from the visual representation of the audio files. This work reported 100% monophonic wheezing sensitivity. [11] reported a crackle sensitivity of 95.7% designing a hilbert energy envelope algorithm without using any machine learning techniques. But the dataset validated comprised of both real and simulated crackles.

While these performance reports are extremely encouraging, there are some glaring blind spots one must be aware of in deploying any of these algorithms for real-world applications. The datasets employed for both training and testing these algorithms are often highly curated, with high quality auscultation signals that reflect unanimous clinical labels from expert reviewers. The selective composition of these datasets hugely under-simplifies the complexities of medical diagnostics, especially for pulmonary infections [3]. Even when models are trained on a gold-standard high-quality data, there is a need to broaden the diversity in test set to identify possible limitations of existing methodologies in order to identify deployment hurdles and tackle drawbacks of current classification techniques. The current work tests the hypothesis that variability in clinical assessment will have a strong impact on the performance of computer-aided auscultation diagnosis.

II. AUSCULTATION DATA

The signals used in this study were collected by the Pneumonia Etiology Research for Child Health (PERCH) study group [12]. A diverse set of 742 interpretable patient recordings from subjects of age 1-59 months across seven different countries were collected as a part of this study. A Thinklabs digital stethoscope was used for collecting lung sounds approximately 7-10 seconds long. All signals were originally sampled at 44.1KHz, pre-processed by applying a low-pass filtered with a fourth-order Butterworth filter at 4 kHz cutoff, downsampled to 8 kHz, and centered to zero mean and unit variance. The signals were further enhanced to deal with clipping distortions, mechanical or sensor artifacts, heart sound's interference, subject's intense crying and ambient noise use a noise-cancellation algorithm [13].

The data collected in this study was annotated by nine expert reviewers (pediatricians or pediatric-experienced physicians) who evaluated the signals through a listening evaluation blinded to other clinical factors pertaining to each case. The expert reviewers indicated whether the recording is a normal auscultation signal or an adventitious pattern including a wheeze or crackle. Each signal was assigned to two reviewers and further arbitrated by a third in case of disagreement. Further, each reviewer was asked to indicate a level of confidence in her/his evaluation by rating their level of certainty in the label as sure or unsure. It was noted that the inter-reviewer case disagreement w.r.t the presence or absence of abnormal lung sounds, crackles, and wheezes were 25.1%, 29.4 %, and 27.4 % respectively [14]. The data comprised of 4500 Normal, 1100 Abnormal, and 1150 non-conclusive 2 second samples . To obtain fixed duration for sample, we passed a sliding window over variable length overlapping hand-given annotations by multiple reviewers. It was ensured that the range included three degrees of agreement: complete agreement (when all reviewers agreed on the clinical assessment), partial agreement (when the majority of reviewer agreed on the clinical assessment but a minority presented an opposite opinion), and complete disagreement (where no consensus emerged among reviewers). Further, there were different levels of certainty considering how sure the reviewers were with the label they assigned. In the current study, the data with complete agreement (perfect consensus) and high reviewer surety is considered as "High Certainty" Data(HC), whereas all other cases are considered "Variable Certainty" Data(VC), which include unsure opinions and levels of disagreements among reviewers. The evaluation metrics reported consider a single true label in case of consensus and the common label given by majority of reviewers in case of partial agreement (ro minority dissent). There was no sense of a true label in case of no agreement and just the confidence values outputted by the classifier were reported.

III. AUTOMATED AUSCULTATION CLASSIFICATION

Abnormal lung sounds such as wheezes (long whistling sounds) and crackles(series of short explosive sounds) are considered a hallmark of presence of pulmonary infections and used as indicators of respiratory diseases [15]. Given the sequential nature of auscultation signals as they unfold over time as well as the temporal variability of lung sounds, the current work explores a convolutional feature extractor followed by a Long Short-term Memory network (LSTM) which is known to operate efficiently on sequential data and does not suffer from the vanishing gradient problem typical of Recurrent Neural Networks.

Auscultation audio signals were converted to mel-frequency spectrograms of dimensions 32x64 using librosa, a python library. It essentially transforms the temporal signal to a spectrogram mapped onto a bank of log-scaled asymmetrical cochlear filters. A two layer CNN was used as an feature extractor on the spectrograms of two second audio segments. A convolutional neural network when learnt on auditory spectrogram learns filters that activate when they encounter a certain auditory cue thereby producing 2-dimensional activation map. These convolutional features were then processed through a two layer Long short term memory network with 50 hidden cells followed by a fully connected layer. Finally, a sigmoid activation gave out the confidence values of predicted labels. The network was trained as a fully-supervised architecture with label "0" given to normal lung sounds and "1" given to abnormal lung sounds, including both wheezes and crackles. This two class classifier was optimized for Cross Entropy Loss using Adam optimizer in a pytorch framework. A learning rate of 0.001 was set for around 50 epochs with specified feature size. All evaluations of the classifier were performed in a 5-fold cross-validation.

IV. CLASSIFICATION AND UNCERTAINTY RESULTS

A. Classification performance

The classifier had a sigmoid activation as the last layer thereby outputting confidence values between 0 and 1, with 0 being the true label for Normal and 1 being the true label for Abnormal lung sounds. The means of specificity, sensitivity, f1 score, gmean(geometric mean of specificity and sensitivity) and area under Receiver Operating Characteristic(ROC) curve across 5-fold cross-validation along with their standard errors were reported in TABLE I. Since there is a high class imbalance between normal and abnormal lung sounds, we included the gmean metric to ensure a fair evaluation of the model across classes. The performance of the classifier trained and tested on the High-Certainty(HC) dataset across 5-folds(HC|HC) gives a reasonable geometric mean of 0.82$\pm$0.015. On evaluating the average performance of these 5-instances of HC trained model on Varied Certainty subset(HC|VC), there is a stark drop in the geometric mean to 0.67. Next, we verify whether this degradation is due to mismatch in data distributions across both datasets or size limitations of the HC subset. If these reasons were valid, retraining the classifier with the larger dataset should be able to address the data mismatch issue. To do so, we retrained the classifier with HC+VC data (with appropriate allocation of non-overlapping training and test data). As noted in TABLE I rows 3 and 4, testing on the high quality (HC) data results in a stable performance (gmean of 0.81 $\pm$0.014), whereas evaluating on the VC dataset results in only a slight improvement of gmean to 0.68. These results suggest that testing the classifier with clinical uncertainty does introduce a great deal of variability in classifier predictions, hence raising concerns as to whether such classifier can be deployed blindly without carefully considering test cases where clinical uncertainty renders the prediction of the classifier invalid. To better understand the impact of clinical uncertainty, we further analyzed the impact of various degrees of variability in expert opinions.

B. Clinical Variability

The clinical certainty was examined over two dimensions: Agreement in the true labels by different reviewers and the surety of the annotation they assigned to the auscultation signal. TABLE II reports the classifier results, where we note a steady increase in performance of the classifier as we move along agreement from partial to complete across different surety levels (0.69 to 0.82 for Sure and 0.53 to 0.64 for Unsure in geometric mean). In a similar fashion, annotations with more confidence from the reviewers had better performance than uncertain annotations across different agreement. It can be noted that the confidence of reviewers for same agreement level causes a larger drop in performance than lack of agreement. In case of complete disagreement between reviewers, there was no ground truth true label, therefore it was not possible to evaluate the accuracy of the classifier.

While the results reported in TABLE II reflect the classifier final outcome, we also explored the model predictions as reflected in the classification posteriors for both normal and abnormal classes. Based on the training constraints, we expect the normal class posterior probabilities to be concentrated around 0 and the abnormal class around 1. The empirical probability density distribution of the posterior probabilities of the normal and abnormal recordings were visualized across the two clinical certainty factors in Fig. 1. The first column, first row and third column, first row posterior distributions reflect the output of the classifier when tested on curated data with highest confidence and highest agreement among reviewers. Similar to the performance metrics reported, the most discriminable density functions were observed for the complete consensus and highly certain data (upper left panel) with posterior means of 0.21 and 0.74 for normal and abnormal respectively. The second column along the first row represents data with complete agreement and lower confidence from reviewers. Despite the agreement, there was higher overlap in the posterior densities with the mean distribution of normal and abnormal classes being 0.44, 0.64, respectively.

In the second row, the leftmost panel depicts data where majority of reviewers agreed despite some dissenting opinions for a minority, though all reviewers were highly confident of their assessment. The distribution means of normal and abnormal posteriors in this case were 0.36 and 0.72 respectively; while the case where there is only partial agreement with less certainty (second row, rightmost panel) showed far greater overlap between distributions with density means of 0.49 and 0.56, respectively. Finally, the third row shows the posterior probabilities of recordings that were not conclusively normal or abnormal (no clinical consensus) with a 0.5 posterior mean for surer data and 0.535 for unsure data.

C. Classifier Certainty

Given the fact that the classifier does reflect a degree of uncertainty based on its prediction label probability, we next explored the use of these posteriors as automatic flags of clinical uncertainty. Irrespective of the true label of each data point, if the posterior value given by the classifier is farther away from 0.5, it is more confident in its prediction. Interestingly, it was noted that at each agreement level including completely disagree, there was an increase in the density around 0.5 indicating decrease in the confidence of the classifier with the decrease in clinical certainty of each reviewer. Based on this observation, we formulated Classifier Certainty(CC) between 0 to 1 as follows where ϕ was the Auscultation soft classifier estimating the posterior probability on a test lung sound sample x_n:

$$CC(x_n) = 2 * |0.5 - \phi(x_n)|$$

The motivation behind estimating the certainty of the test sample is not to artificially improve the classification performance, but rather to propose a quantitative way to flag cases that require additional evaluation, either by collecting a new auscultation sample or by exploring other clinical markers(X-rays, blood work) in case the auscultation signals are inconclusive. To verify this particular thresholding

TABLE I: Classifier performance trained and tested on different data subsets (mean and standard error, 5-fold cross-validation)

Train‖Test	Specificity	Sensitivity	F1 Score	GMean	Area under ROC
HC‖HC	84.27±0.64	79.23±2.97	0.65±0.014	0.82±0.015	0.89±0.01
HC‖VC	62.75±1.92	72.24±2.14	0.55±0.009	0.67±0.001	0.73±0.01
HC+VC‖HC	86.13±0.88	77.34±2.11	0.66±0.021	0.81±0.014	0.89±0.001
HC+VC‖VC	65.63±0.79	71.59±1.09	0.56±0.012	0.68±0.008	0.75±0.01

TABLE II: Classifier performance tested on data subsets† with different levels of clinical variability when trained on HC

† Note that percentage of data does not sum to 100% because the classifier performance cannot be evaluated for data with no-consensus.

Agreement.	Confidence	Specificity	Sensitivity	F1 Score	GMean	Area under ROC	HC+VC Data
Agree	Sure	84.27±0.64	79.23±2.97	0.652±0.014	0.82±0.015	0.89±0.01	58.4%
	Unsure	57.29±1.65	72.17±5.2	0.512±0.013	0.64±0.02	0.68±0.02	28%
Disagree	Sure	63.47±1.78	76.39±2.74	0.62±0.03	0.69±0.017	0.79±0.02	8.4%
	Unsure	48.21±4.78	60.70±3.01	0.47±0.02	0.53±0.017	0.55±0.03	5%

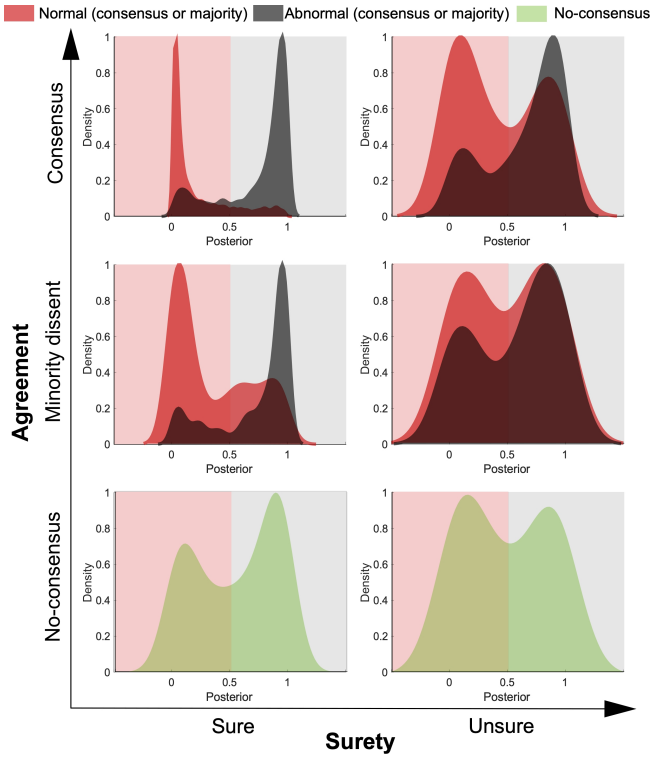

Fig. 1: Confidence Values outputted by the classifier for different clinical certainty levels

technique based on the estimation of classifier uncertainty is not trivial, we analyzed the performance of the classifier by gradually increasing the clinical certainty in Fig. 2 and compared it with a random exclusion criterion. With the increase in the minimum classifier certainty criterion, the proportion of data held out for reassessment increased in a cumulative fashion from 0% to 0.41%. This is validated by the gradual increase of gmean performance metric from 0.75 to 0.845 with the threshold. The blue marker at each threshold has a radius proportional to the percentage of data excluded. Further, mean of a 100 point gmean sample

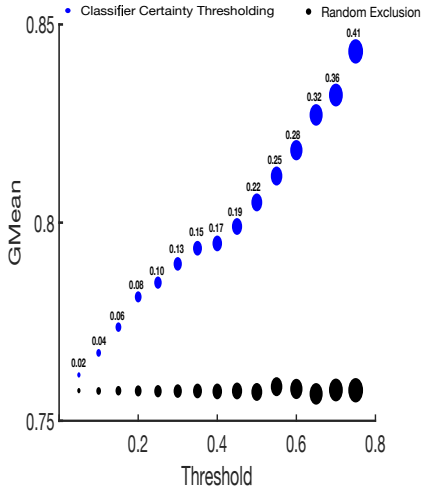

Fig. 2: GMean evaluated after refraining from misclassifying test

obtained by randomly excluding an equal percentage of test population at each threshold point is represented in gray. This random performance across different proportions of data clearly shows that exclusion of any amount of data does not inherently increase the performance on the rest of the test set. This further emphasizes the conclusion that the Classifier Certainty Thresholding technique is reassessing questionable cases and is not haphazard.

V. CONCLUSIONS

In order to make the necessary strides towards real-world deployment of any of the existing computerized auscultation analysis algorithms, we must address the potential bottlenecks which are not being accounted for in our drive to achieve better results on gold-standard data. Like with any medical application, there is bound to be an inherent bias even with expert clinical reviewers thereby giving rise to inter-reviewer variability and noisy labels. [14] established the former with the significant percentages of disagreements

reported for each abnormality. This further underscores that auscultation abnormality detection is not a traditional binary classification task, but rather a softer abnormality confidence indication. When a machine is expected to diagnose signals with very low inter-reviewer agreement or intra-reviewer certainty, one can expect it to misdiagnose. In this work, we first verified how learning the noisy labels does not necessarily solve this problem. And then proceeded to analyse the clinical interpretation dependent factors which have an impact on the performance of these CAA algorithms. We noted how the confidence with which each reviewer annotates might have a bigger impact on the predictions than ensuring if there is inter-reviewer agreement. A need for robust CAA techniques with a more comprehensive outlook in data curation especially when testing is to be realized. Despite the clinical variability or self-uncertainty in clinical assessment, each reviewer reaches some final conclusion with a certain degree of confidence. We noted how classifier tested on lung sounds annotated with low confidence from reviewers also tend have low confidence(nearer to 0.5) from the classifier. Based on this idea, a more appropriate softer diagnosis from the classifier mimicking the annotation process by the expert reviewers is proposed.

ACKNOWLEDGMENT

The authors would like to thank the PERCH study group for guidance throughout the completion of this work, and to the patients and families enrolled in this study. We also acknowledge support from NIH U01AG058532, ONR N00014-19-1-2014, and N00014-19-1-2689.

REFERENCES

[1] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. 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*, vol. 25, no. 7, pp. 2595–2603, 7 2021.

[2] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust Deep Learning Framework for Predicting Respiratory Anomalies and Diseases," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 164–167, 7 2020.

[3] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine, and W. Checkley, "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis." *Respir Med*, vol. 105, no. 9, pp. 1396–1403, 9 2011. [Online]. Available: http://dx.doi.org/10.1016/j.rmed.2011.05.007

[4] J.-C. Chien, H.-D. Wu, F.-C. Chong, and C.-I. Li, "Wheeze Detection Using Cepstral Analysis in Gaussian Mixture Models," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 3168–3171.

[5] D. Emmanouilidou, K. Patil, J. West, and M. Elhilali, "A multiresolution analysis for detection of abnormal lung sounds," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012. IEEE, 8 2012, pp. 3139–3142. [Online]. Available: http://ieeexplore.ieee.org/document/6346630/

[6] M. Aykanat, Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *Eurasip Journal on Image and Video Processing*, vol. 2017, no. 1, 2017.

[7] D. Chamberlain, D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, "Application of Semi-Supervised Deep Learning to Lung Sound Analysis Application of Semi-Supervised Deep Learning to Lung Sound Analysis," *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, no. August, pp. 804–807, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7590823/

[8] H. Chen, X. Yuan, J. Li, Z. Pei, and X. Zheng, "Automatic Multi-Level In-Exhale Segmentation and Enhanced Generalized S-Transform for wheezing detection," *Computer methods and programs in biomedicine*, vol. 178, pp. 163–173, 9 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31416545/

[9] J. De La Torre Cruz, F. J. Cañadas Quesada, J. J. Carabias Orti, P. Vera Candeas, and N. Ruiz Reyes, "Combining a recursive approach via non-negative matrix factorization and Gini index sparsity to improve reliable detection of wheezing sounds," *Expert Systems with Applications*, vol. 147, p. 113212, 6 2020.

[10] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 88, pp. 58–69, 6 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0933365717302051

[11] R. Pal and A. Barney, "Pulmonary Crackle Detection Using the Hilbert Energy Envelope," *IFMBE Proceedings*, vol. 80, pp. 994–1003, 11 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-64610-3$_1$11

[12] E. McCollum, D. Park, N. Watson, C. Focht, C. Bunthi, B. Ebruke, M. Elhilali, D. Emmnouilidou, L. Hossain, D. Moore, A. Mudaua, J. Mulindwa, J. West, K. O'Brien, D. Feikin, and L. Hammitt, "Digitally-recorded lung sounds and mortality among children 1-59 months old with pneumonia in the Pneumonia Etiology research for Child Health study," Tech. Rep., 2017. [Online]. Available: https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2017.195.1$_MeetingAbstracts.A1195$

[13] D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali, "Computerized Lung Sound Screening for Pediatric Auscultation in Noisy Field Environments," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 7, pp. 1564–1574, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/7953509/

[14] E. D. McCollum, D. E. Park, N. L. Watson, W. C. Buck, C. Bunthi, A. Devendra, B. E. Ebruke, M. Elhilali, D. Emmanouilidou, A. J. Garcia-Prats, L. Githinji, L. Hossain, S. A. Madhi, D. P. Moore, J. Mulindwa, D. Olson, J. O. Awori, W. P. Vandepitte, C. Verwey, J. E. West, M. D. Knoll, K. L. O'Brien, D. R. Feikin, and L. L. Hammit, "Listening panel agreement and characteristics of lung sounds digitally recorded from children aged 1–59 months enrolled in the Pneumonia Etiology Research for Child Health (PERCH) case–control study," *BMJ Open Respiratory Research*, vol. 4, no. 1, p. e000193, 6 2017. [Online]. Available: http://bmjopenrespres.bmj.com/lookup/doi/10.1136/bmjresp-2017-000193

[15] Z. MOUSSAVI, "Respiratory sound analysis [{Introduction} for the {Special} {Issue}]," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 1, p. 15, 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4069350