2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782543

Robust Anomaly Detection of Adventitious Auscultation Signals using Bayesian Belief Tracking

Annapurna Kala, Mounya Elhilali
Johns Hopkins University, Department of Electrical & Computer Engineering,
Laboratory for Computational Audio Perception

***Abstract*—Stethoscope screening serves as a primary method for diagnosing pulmonary infections, with medical professionals actively listening for signs of pathologies in breathing sounds like wheezing and crackling, which carry different clinical interpretations. Environmental conditions during auscultation recordings often share similarities with these abnormal lung sounds, and can mask or confound their presence making their detection highly sensitive to surrounding factors. To automate this process, a robust anomaly detection scheme with resilience to ambient backgrounds and high precision is essential. In this study, we propose an unsupervised framework for anomaly detection where statistics of a deep neural network embeddings are tracked using a Bayesian belief model in order to flag variations that are deemed anomalous, hence facilitating detection of adventitious auscultation events. The proposed scheme leverages two key principles: (1) learning of statistics of normal auscultation patterns using variational constraints, and (2) tracking changes in the statistics using Bayesian beliefs that interpret anomalies as deviations from normal statistics. This approach is shown to be very effective in detecting adventitious auscultations under various noise levels hence ensuring its resilience to environmental conditions.**

I. Introduction

With advances in machine learning, there is increasing interest in applying computer-aided analytics in a variety of healthcare applications [1], [2]. In the case of pulmonary diseases, lung auscultations are sounds produced by the lung and accessed through a stethoscope to provide information on the underlying pulmonary pathological conditions [3]. Advanced signal processing and machine learning tools have opened new frontiers in computer-aided auscultation techniques by analyzing stethoscope recordings to enable automatic identification of abnormal lung sounds that are indicative of underlying lung pathologies [4], [5]. Generally, a framework adopted in most studies is to train classifiers such as Convolutional Neural Networks, Support Vector Machines, Gaussian Mixture Models on different classes of auscultation sounds (normal breathing, abnormal patterns) and use this framework to analyze any incoming signal as normal or abnormal [6], [7], [8]. A recurring theme across these methodologies is the use of labeled datasets containing ample recordings of both normal and abnormal lung sounds. Access to labeled auscultation is not only a timely and costly endeavor, but it requires human expertise by trained physicians to carefully curate signals and annotate abnormal patterns. Furthermore, a challenge with supervised methods is the scarcity of adventitious lung sounds. This issue of data imbalance exacerbates the challenges associated with training deep learning architectures [9], and requires the use of various strategies, including proper sampling techniques, the utilization of appropriate evaluation metrics, as well as the use of data augmentation techniques [10].

An alternative approach to tackle these limitations is the use of unsupervised techniques, specifically anomaly detection which is well suited for the task of detection of adventitious auscultations [11]. Anomaly detection focuses primarily on identifying patterns in the data that deviate significantly from the norm. In other words, it is simply the task of identifying out of distribution examples, hence forgoing the need to label those examples. As an unsupervised method, it provides a number of advantages including flexibility and adaptability to different types of data and domain sets as well as scalability across different settings. With these advantages, identifying anomalies can be a challenging task due to variability in interpreting what is a deviation from the norm. Furthermore, noise interference or dynamic changes in the data setting or levels can alter the interpretation of normal patterns leading them to appear out of distribution and ultimately be flagged as abnormal. These limitations require careful learning strategies that leverage domain knowledge of the structure of the data in order to define meaningful representations of the underlying statistical distributions, hence yielding more robust behavior in anomaly detection [12].

The present study develops a framework for anomaly detection of adventitious auscultation signals that considers meaningful mappings that capture the characteristics of normal and abnormal instances. The proposed scheme explores features that a) best canvas the stochastic space of normal lung sounds and their underlying statistical distribution, and b) accurately reflect a change in this distribution at the onset of an adventitious sound. The proposed scheme uses a deep neural network with variational constraints in order to control the mapping of underlying statistics of auscultations. The network is trained on a dataset of normal lung sounds in order to learn a good representational space for a reference statistical distribution of auscultations. Embeddings from this model are then used as input to a Bayesian tracking model that generates statistical predictions or beliefs about changes in the underlying distribution. A simple threshold is then used to evaluate presence of any abnormality. This framework is compared against other representations and evaluated in presence of various noise levels to probe its robustness as well as timing precision to detect onset of adventitious sounds.

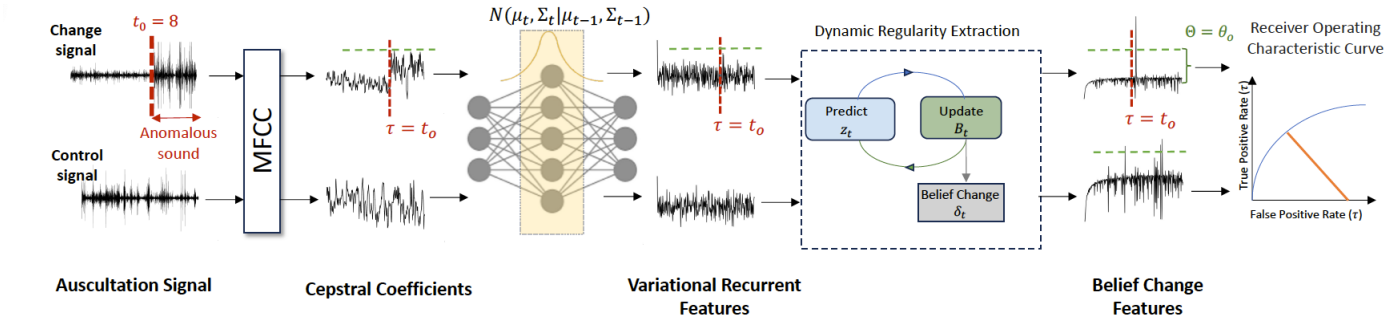


Fig. 1: Overview: Proposed pipeline for anomaly detection. A Change signal (top-left) and Control Signal (bottom-left) differ starting from $t_0 = 8$ when an anomalous sound appears in the change signal. Cepstral coefficients are extracted from lung sounds and are mapped onto a variational recurrent feature space. A belief change detection algorithm then identifies the change in this underlying Change auscultation distribution whose peak (if $\Theta > \theta_0$) marks the onset of the lung sound anomaly within tolerance. Receiver Operating Characteristics curve considering such True Positive and False Positive values and gives the final best F1-Score.

II. Data

The auscultation signals used in this study include recordings from the Pneumonia Etiology Research for Child Health (PERCH) study group [13]. Auscultations were recorded using a Thinklabs ds32a digital stethoscope at 44.1 KHz. Signals are then pre-processed using a low pass fourth-order Butterworth filter with a cutoff at 4kHz, downsampled to 8 kHz, centered to zero mean and unit variance, and denoised using an auscultation specific ambient noise-cancellation algorithm [7].

Each auscultation recording is annotated by an expert reviewer panel indicating the presence/absence of abnormal lung sounds such as wheezes and crackles [14]. Physicians also temporally localize the said abnormality by a temporal onset and offset using Audacity software. The data used in the present study consists of 13.9 hrs of data from 800 patients from 8 different chest positions where the listening panel agrees in the final arbitration of normal, wheeze, crackle or both annotation as well as their temporal placement within a chest recordings. We standardize the duration of recordings analyzed to 12 seconds. We define 'change' signals as recordings where we insert a 4-second abnormal lung sound (from the labeled dataset) at time $t_0 = 8$ sec. To evaluate different signal strengths, the abnormal signal is added at different strengths: 12 dB, 8 dB, 4 dB, and 0 dB defined over RMS (root-mean square) energy change from normal to abnormal auscultation at $t_0 = 8$ secover a 500 msec window.

The data is then divided into training and testing subsets as follows: 10.02 hrs of only normal auscultations ($Train$) of which 6.82 hrs ($Train_{Autoencoder}$) is used to obtain the data-driven statistical recurrent features and 2.2 hrs ($Train_{StatiscalModel}$) are used to learn the statistical distribution of the learnt stochastic transitional features. The remaining 1 hr ($Train_{AnomalyThreshold}$) of normal auscultations are held out to compute appropriate threshold statistics across different feature spaces to perform the final anomaly detection. The analysis is tested on a test dataset $Test$ comprising of 500 normal recordings, 530 wheezes, and 160 crackles of 12 second duration with anomaly onset for abnormal auscultations at $t_0 = 8$ sec. Note that the model training is never presented change (or abnormal) signals and as such the use of a fixed change point if simply to evaluate the accuracy of the anomaly detection. Furthermore, control signals (normal with no anomaly) are also 12 sec long.

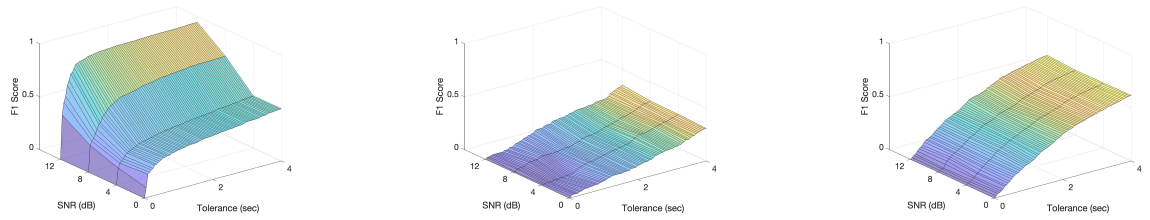
III. Methods

The proposed pipeline maps auscultation signals along different representational spaces, as outlined next.

A. Stochastic Recurrent Feature Mapping

A 12 sec auscultation signal is first analyzed to extract 13-dimensional Mel Frequency Cepstral Coefficents (MFCCs). These MFCC features over time $M(t)$ (dim: 1x13) are sequentially analyzed through a recurrent neural network. The input features are mapped through an embedding layer to obtain encoder mean and variance realizations $(\mu_{enc}(t), \Sigma_{enc}(t))$ of that time-step. We impose recurrence by conditioning the encoder embedding layer on a recurrent hidden state from previous time step $h(t-1)$. This hidden state is computed by passing the previous encoder mean $\mu_{enc}(t-1)$ through a recurrent neural network. To obtain stochasticity, we use the reparameterization trick and sample our stochastic recurrent feature $z(t)$ (dim: 1x8) from a Gaussian distribution $N(\mu_{enc}(t), \Sigma_{enc}(t))$. A decoder layer is then deployed to map back $z(t)$ at each time step to its corresponding MFCC decoded space by computing the decoder mean and variance $(\mu_{dec}(t), \Sigma_{dec}(t))$.

This Variational Recurrent Neural Network (VRNN) Autoencoder is trained *solely* on normal lung auscultations ($Train_{Autoencoder}$) in an unsupervised fashion. This feature space captures the stochastic properties of a normal signals. Such a representation space is learnt by imposing a reconstruction loss in the form of negative log likelihood between the input MFCC $M(t)$ and the decoded MFCC $\mu_{dec}(t)$. We further employ the timestep wise variational lower bound objective by backpropagating on the KL-Divergence between



(a) Input MFCC (b) Stochastic Recurrent Feature (c) Feature Belief Change

Fig. 2: 3D Visualization of F1-Score topology on performing onset detection across MFCC Features, VRNN Features and Belief Change on VRNN Features considering different tolerance duraitons and Normal-Abnormal Signal Ratios.

the encoder distribution $N(\mu_{enc}(t), \Sigma_{enc}(t))$ and decoder distribution $N(\mu_{dec}(t), \Sigma_{dec}(t))$. We perform this optimization using Adam Optimizer and tune the scaling of KL-Divergence to obtain ideal loss profile.

B. Statistical Change Point Detection

The output of the VRNN is then analyzed using a Bayesian Belief model. In this study, we employ the Dynamic Regularity Extraction (D-Rex) Model [15] that predicts the statistical distribution of the stochastic recurrent representation of auscultations at current time step (z_t) as $\psi_t = P(z_t/z_{1:t-1})$. Sufficient statistics of above distribution are then combined to continuously update the belief about the underlying feature space distribution B_t. Difference in beliefs over time is noted as Belief Change δ_t which should ideally be sensitive to the anomaly onset which further diverges the underlying auscultation statistical distribution.

C. Anomaly Onset Identification

Anomaly onset detection is done by identifying temporal instances with significant spike or amplitude changes over time $\delta(t)$. This evaluation is performed the same way over all features in the proposed pipeline (MFCC features, VRNN embeddings, Belief change) in order to compare efficacy of the different representations. The analysis is done by sweeping through a range of thresholds Θ over a set of normal and abnormal auscultation signals. Threshold values are computed based on the percentile statistics of $Train_{AnomalyThreshold}$.

D. Evaluation Metrics & Comparative Monte Carlo Analysis

An onset detection for an anomaly is considered a hit only if identified within a temporal tolerance τ. In case of normal recordings, we would expect not identifying an onset as a True Negative. We compute the Receiver Operating Characteristics (ROC) curve at each τ and identify the best threshold Θ. The final evaluation metrics is reported as an F1-score given at the best Θ.

The performance comparison across different features is done in a Monte Carlo fashion by computing the best F1-Score on test data samples of size 200 with equal number of normal and abnormal auscultations randomly sampled from $Test$ dataset. To compare the final analysis against the features themselves, we establish similar percentile thresholding to identify the onset in both MFCC Feature Space and the VRNN Feature Space. In this case, we work with absolute amplitude of the feature envelope. For comparison, we also formulate a 'random' baseline by working with a random onset detector on similar test samples.

E. Overall pipeline for Anomaly detection

Fig. 1 shows the proposed pipeline for anomaly detection. An auscultation signal without anomaly (bottom-left) or with an anomalous wheezing sound (top-left) is analyzed through the system. The bottom signal is referred to as a control signal since we expect the system to generate no detected anomalous events. Any anomaly detected in a controlled event is flagged as a false alarm. The top signal is a normal breathing signal where we introduce a wheezing anomaly at time $t_0 = 8$ sec at SNR$= 2$ dB (For reference, we evaluate the average signal-to-noise ratio at the onset of abnormalities (crackle, wheeze) relative normal breathing patterns in 57.5 minutes of normal recordings in pneumonia patients from the PERCH dataset [13]. On average, the presence of these anomalies (as annotated by expert pulmunologists) is 1.247 dB. A signal is then analyzed through a first stage to extract cepstral coefficients. The figure shows the first MFCC coefficient for the acoustic recordings shown on the left. As cepstral coefficients are mapped directly from the input signals, their variations reflect closely the overall changes in amplitude in the signal. The cepstral coefficient from the change signal (top) shows a deviation near $t_0 = 8$ sec. No such deviation is observed for the control signal. Next, all MFCC features are analyzed through a data-driven variational recurrent neural network (VRNN). In this case, we expect the network embeddings to yield some degree of alignment with the distribution of normal auscultations on which the network is trained. The output of the network is shown for both signals (top and bottom). Finally, network embeddings are analyzed through a belief tracking system where changes in their statistical structure over time are considered a benchmark against which we flag any deviation, hence leading to anomaly detection. The belief change in both signals are shown on right side of the system. The last stage considers output at different stages in this pipeline and generates a receiver operating characteristic curve considering both correct detections from change signals

and false alarms from control signals at different threshold value at a given tolerance of the original t_0.

IV. Results

A. F1-Score topology across Abnormal-Normal Signal Ratio and Tolerance

In order to examine the contribution of all 3 features in anomaly detection, we evaluate the system performance (in terms of F1-score) under different values of abnormal signal strengths (SNR in dB) and assess the results at different tolerance levels (timing precision). Given that each feature space (MFCC, VRNN and VRNN belief change) spans a number of dimensions, we evaluate the behavior of the first feature in each group before aggregating results across all features within a group. Figure 2 compares the F1-Score topology with onset detection on the first MFCC dimension, first VRNN feature dimension and the feature belief change on the first VRNN feature dimension. In this 3D visualization, we observe that across all three feature groups, the F1-Score increases with the increase in the temporal tolerance as expected. In contrast, we note a different behavior as a function of SNR. In the case of MFCC (leftmost panel), we note that the F1-score starts very high (F1=0.91) at 12dB Abnormal-Normal signal ratio, then followed by a steep decrease as the SNR drops to 0dB. This behavior is consistent with the fact that the presence of a very loud abnormal signal pattern (at 12dB) carries a clear envelope signature which in turn reflects a big change in the amplitude of the MFCC signal. This amplitude-specific effect results in a clear drop of F1 as the SNR changes from 12dB to 0dB. In contrast, the VRNN feature yields an almost flat, though relatively low, F1-Score across both tolerance duration and the SNRs. The same flatter trend is observed in the case of the belief change feature though at a higher F1 baseline.

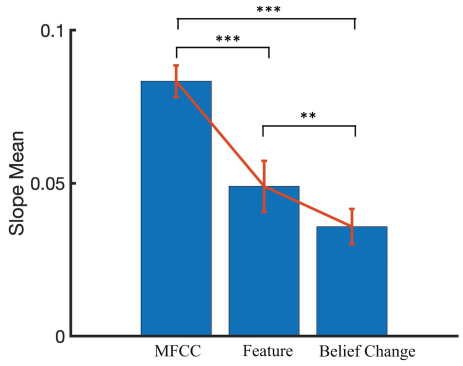


Fig. 3: Mean and standard deviation of the slope of the 3D F1-Score analysis against Signal to Noise Ratio to reflect the robustness of the feature to signal insertion intensity. The comparative analysis is done amongst the thresholding results of MFCC features and the feature belief changes of VRNN Autoencoder features. $**$ reflects statistical significance $p < 0.01$ and $***$ reflects statistical significance $p < 0.001$

While these results focus on one example feature in each feature group, they reflect a general behavior of these features in terms of effect of SNR on F1-scores.

B. Comparison of anomaly detection across features

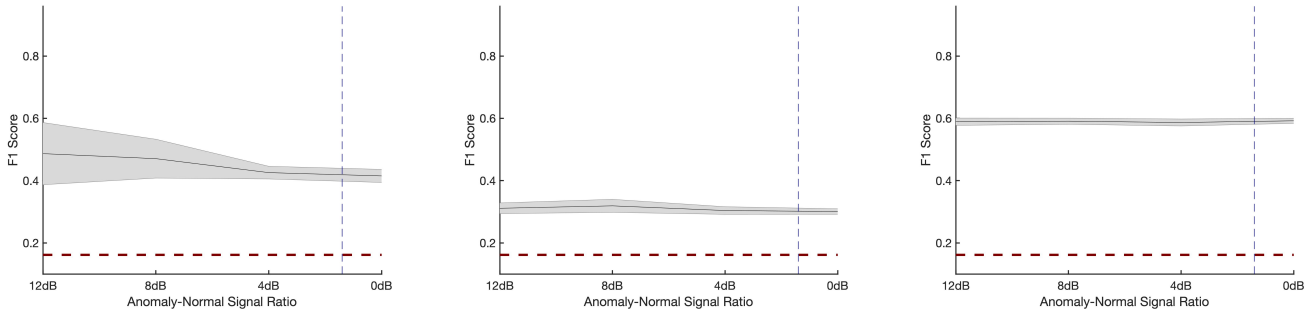
To quantify the robustness of onset detection, we compute the gradient of each individual F1-Score 3D topology (as visualized in Fig 2) across the SNR dimension. Calculating the norm of this derivative gives the slope of F1-Score performance across different Normal-Abnormal Signal Ratios. A higher slope implies a steeper decline in the performance with the decrease in SNR and thereby making the onset detection on that particular feature less robust. The robustness of onset detection based on a particular feature is represented as the mean of slope across different dimensions within that feature. We observe that the slope statistics on the random MC samples give an average of 0.083 for MFCC based onset detection, 0.049 for the statistical recurrent feature, and 0.035 for feature belief based onset detection as observed in Fig 3. We further perform Mann Whitney U Test on these slope random statistic samples to confirm statistically significance with a p-value of 1.7420e-65 on comparing MFCC Slope sample with the Belief Change slope sample and a p-value of 0.032 for the Feature-Belief change slope comparison. Comparing MFCC and VRNN features yields a p-value of 4.3e-12.

C. Robustness to signal variation

We analyze the robustness of proposed method by working the F1-Score at a tolerance of τ=4 seconds across different features as this reflects an overall soft detection of abnormality, if any. Here, the analysis considers *all* features explored in the current pipeline. Nominally, the dimensionality D of different features in this comparative analysis is as follows: 1. MFCC ($D = 13$), 2. Stochastic Recurrent Feature ($D = 8$), and 3. Feature Belief Change ($D = 8$). In Fig 4, the mean F1-Score is computed across all feature dimensions on different Abnormal-Normal Signal Ratios. The shaded curves represent the standard error across dimensions. The results show a mean of 0.61 F1-Score for the feature belief change at 0 dB as compared to the 0.431 of MFCC Envelope and 0.30 of Stochastic features. The average random baseline performs at an F1-Score 0.156 $\pm$ 0.007 and is reported as a red dashed line to evaluate the performance of a random onset detector on this task. For reference, the vertical line depicts the 'natural' Abnormal-Normal Signal RMS ratio at 1.247 dB. This value is evaluated by assessing abnormal signals in the dataset and estimating the change in signal strength from a normal breathing when an adventitious event occurs (see Methods).

V. Discussion

Overall, these results reveal three key trends. First, MFCC which are data-agnostic features seem to follow more closely the amplitude of the auscultation signal itself. At high signal levels, there is a clear intensity change which results in high F1 scores. As the normal-to-abnormal strength decreases, MFCC features decrease dramatically in performance as reliance on overall signal strength is not a robust or scalable strategy for anomaly detection [11]. As such, low-level features such as MFCC are highly susceptible to signal structure and lack a true representation of the space of normal lung sounds.



(a) Input MFCC (b) Stochastic Recurrent Feature (c) Feature Belief Change

Fig. 4: Comparative Robustness Analysis of Mean and Standard Error of F1-Score performance on onset detection across (a) MFCC, (b) VRNN Feature and (c) Feature Belief Change at tolerance = 4 seconds.

Second, data-drive features like VRNN embeddings present a general improvement in terms of overall performance across signal strength and temporal tolerance, though they operate at lower F1 scores in terms of signal strength (as shown in Fig. 3). While these features successfully capture the statistical distribution of normal auscultations, using a direct thresholding on the features themselves does not accurately leverage the learned stochastic learning. Finally, a proper tracking of statistical changes, which is achieved through a Bayesian tracking model balances both the statistical learning and temporal predictions leading to more robust anomaly detection. An important aspect of the proposed work is the clear importance of the feature space used as reference distribution to identify out of domain samples. As noted earlier, the proposed scheme leverages variational constraints in the recurrent neural network which impose constraints of Gaussianity on the learned embeddings. These constraints lend themselves well to a bayesian tracking framework (DREX) which -at its core- uses Gaussian statistics to learn past observations and make predictions about incoming observations [15].

Overall, the proposed system offers a general framework for anomaly detection for auscultation analysis. The study focuses on the appropriate choice of feature space for such effort in order to provide a suitable space to learn normal statistical behavior of the data, hence leading to robust detection of abnormal patterns under different signal levels. Such framework could be easily integrated in a clinical workflow to assist healthcare professionals in diagnosing respiratory conditions. Future research needs to examine more reliable methods of detection beyond thresholding.

ACKNOWLEDGMENT

The authors would like to thank the PERCH study group andthe patients and families enrolled in this study. We also acknowledge support from NIH R01HL163439 and ONR N00014-23-1-2086.

REFERENCES

[1] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature Communications*, vol. 11, no. 1, p. 3923, 8 2020.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," *Healthcare*, vol. 10, no. 3, p. 541, 3 2022.

[3] L. A. Geddes, "Birth of the stethoscope," *IEEE Engineering in Medicine and Biology Magazine*, vol. 24, no. 1, pp. 84–86, 1 2005.

[4] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, *Automatic adventitious respiratory sound analysis: A systematic review*, 2017, vol. 12, no. 5.

[5] J. P. Garcia-Mendez, A. Lal, S. Herasevich, A. Tekin, Y. Pinevich, K. Lipatov, H.-Y. Wang, S. Qamar, I. N. Ayala, I. Khapov, D. J. Gerberi, D. Diedrich, B. W. Pickering, and V. Herasevich, "Machine Learning for Automated Classification of Abnormal Lung Sounds Obtained from Public Databases: A Systematic Review," *Bioengineering*, vol. 10, no. 10, p. 1155, 10 2023.

[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. 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.

[8] A. M. Alqudah, S. Qazan, and Y. M. Obeidat, "Deep learning models for detecting respiratory pathologies from raw lung auscultation sounds," *Soft Computing*, vol. 26, no. 24, pp. 13 405–13 429, 12 2022.

[9] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 12 2019.

[10] A. Kala and M. Elhilali, "Constrained Synthetic Sampling for Augmentation of Crackle Lung Sounds," in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2023, p. in press.

[11] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, no. 1, p. 42, 12 2020.

[12] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 3 2022.

[13] 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.

[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.

[15] B. Skerritt-Davis and M. Elhilali, "Computational framework for investigating predictive processing in auditory perception," *Journal of Neuroscience Methods*, vol. 360, p. 109177, 8 2021.