January 27, 2025
Report: Identifying Wheezes and Crackles in Raw Respiratory Recordings

Author:
Christian Landeros, Ph.D.

Codebase:
https://www.github.com/landeros10/ICBHI_2017

1. Summary:

Few procedures match chest auscultation for its ability to provide a wide range of clinical information, while remaining quick, easy, and nearly universally accessible. In this report we outline a simple transformer-based model for the detection of ‘crackle’ and ‘wheeze’ events in auscultation recordings. These auditory events, while identifiable by most trained physicians, are often diverse and indicative of several physiological phenomena (1). Physicians must be careful to identify any unique features (i.e. timing, duration, intensity variation, etc.) that could provide insight to underlying respiratory conditions or disease progression.

In this project, the transformer-based model, wav2vec2, was finetuned to generate and classify auscultation representations, identifying the presence of crackles and wheezes in a multilabel classification framework (2). Respiratory sounds from the ICBHI database were used to train and evaluate the fine-tuned model, totaling over 5.5 hours of recordings (3). During training, 5-second durations were randomly clipped from sample. Samples were then augmented by stochastically mixing with a database of hospital ambient noise (4). The best performing model was evaluated on a test set of 381 samples from [patients], achieving a [] F1 score, and [] AUROC for crackles and wheezing respectively. Interpretability maps were also generated using importance scores derived from input gradients and attention weights extracted from the model's attention layers. These visualizations identify time segments in the recording that were most influential in the model's predictions. In identifying segments in the recording most relevant to classification, a set of representations for crackles and wheezes was compiled, enabling further analysis of the distinguishing features between them.

2. Methods

Dataset:

The dataset comprised of 920 audio samples from 126 patients. 381 were removed for model evaluation, and 539 were used for training and validation. Each sample was annotated with the following labels for each respiratory cycle in the sample: cycle start time, cycle end time, presence of crackles, and presence of wheezes. A sample or clip was considered positive if at least one respiratory cycle within it contained a positive instance of the target class.

Preprocessing:

Raw WAV files were processed using the Torchaudio library for Python. A band-pass filter ranging from 150 to 800 Hz was applied to all training and evaluation samples to reduce background noise (5). Signals were then resampled to 16,000 Hz and converted to mono.

Augmentation:

Full-duration waveforms for each audio sample were randomly clipped to 5.0 seconds, a duration that captures at least one average respiratory cycle, and has been previously used in other studies. During clipping, the ground truth annotations indicating respiratory cycle start and end times were used to match the 5-second clip to the appropriate ground-truth label.

To explore domain-specific augmentation, hospital ambient noise from the Kaggle Hospital Ambient Noise Dataset was randomly added to each sample (4). This dataset contains 562 audio chunks of 5.0 second duration. With probability 0.5, the original waveform was mixed with a randomly selected noise clip to achieve a signal-to-noise ratio (SNR) of 20.

Model:

The wav2vec2 utilizes transformer layers to generate robust and generalizable audio representations, which have been applied across various biomedical domains for downstream tasks (6). We apply this model to generate waveform embeddings, which are then passed through a classification head to produce two binary outputs indicating the presence or absence of crackles and wheezes respectively. Framing the problem as a multi-label classification task enhances generalizability, reduces overfitting risk, and allows the model to learn to identify a broad range of respiratory features.

Training:

The training set was divided into training and validation subsets using an 80/20 split ratio. During training, the transformer layers were frozen, allowing updates only to the classification head. For multi-label binary classification, the binary cross entropy loss function was used. The model was implemented using the PyTorch machine learning library and trained on Google Colab, utilizing a single NVIDIA T4 GPU. The model was trained for 100 epochs using the AdamW optimizer with a learning rate of 1e-5 and batch size of 16. Hyperparameter optimization was conducted using Weights & Biases (W&B) Sweeps with a random search strategy, optimizing for validation loss. Once an optimal set of hyperparameters was chosen, the same were applied for all model versions.

Importance Maps:

To enhance model interpretability, importance maps were generated using two complementary methods: gradient-based importance mapping and attention weight analysis. Importance maps were created by calculating the gradient of the output logits with respect to the input waveform, highlighting the most influential regions of the audio signal. Attention maps were derived using the attention rollout technique, which aggregates attention scores across layers to capture long-range dependencies (7). Additionally, attention weights from the model's attention layers were aggregated by applying a max reduction across all attention heads.

Results & Discussion:

Test results:

The best performing model in our search was applied to the test set of 381 samples. These samples were presented to the model without clipping. The F1 score for detecting crackles or wheezes were 0.58 and 0.58, respectively. Further metrics are presented in Table 1. The model's performance indicates a moderate ability to classify crackles and wheezes, with slightly better overall accuracy for wheezes. Crackles and wheezes are inherently subjective in clinical interpretation, posing challenges for model training. Wheezes, for example, are much easier to

identify, characterized by their distinctive long duration and musical quality. Crackles, on the other hand, are shorter, explosive, and more variable in presentation

Model	Acc	Precision	Recall	F1
wav2vec + hospital noise	0.55 / 0.66	0.55 / 0.51	0.53 / 0.59	0.54 / 0.55
wav2vec + random noise	0.57 / 0.65	0.63 / 0.50	0.35 / 0.62	0.45 / 0.55
wav2vec	0.59 / 0.67	0.58 / 0.52	0.59 / 0.65	0.58 / 0.58

Table 1. Model Performances

Binary classification metrics for crackles and wheezes for models tested in this study. Models were evaluated with different data augmentation techniques, including random noise and hospital ambient sound, compared to a baseline with no augmentation. Values in each cell represent performance for (crackles / wheezes).

Notably, the un-augmented model outperformed both noise-augmented variants, suggesting that the pre-trained features alone provided the most robust representation of the respiratory sounds. The results suggest that the addition of noise (ambient or Gaussian alone) may have led to the generation of waveform representations with reduced informative content, hindering the classification head's ability to distinguish meaningful features. For the hospital ambient sound specifically, its limited effectiveness could be attributed to the quality, diversity, or size of the noise dataset. Unlike random noise, which can be generated in unlimited variations, the ambient noise dataset is inherently constrained, potentially leading to insufficient exposure to diverse background conditions.

While ambient noise is domain-specific and reflective of clinical environments, it may be more effective when applied alongside other techniques. More impactful domain-specific challenges—such as patient movement, varying lung conditions, and device inconsistencies—introduce significant variability that the augmentation strategies presented above fail to address. To improve generalization, next steps should focus on building a robust augmentation pipeline that does not introduce dataloading bottlenecks. Incorporating commonly used audio augmentation methods, such as time stretching, pitch shifting, dynamic volume adjustments, and intermittent noise injections may also provide additional improvements. By incorporating a broader range of clinically relevant conditions, the model can be made more robust and better aligned with practical diagnostic applications.

A critical component in evaluating model performance is understanding which parts of the input contribute most to its final decision-making. This step is particularly essential in clinical settings, where explainability and interpretability are paramount to building clinicians' trust in AI-assisted diagnostics. Interpretable models empower physicians to integrate machine-generated insights with their own comprehensive understanding of the patient's overall condition.

To begin developing tools for model interpretability, a simple mapping technique was used to identify key segments of the auscultation recordings. My approach leveraged input gradients and transformer attention weights to highlight regions of the audio waveform deemed most informative by the model. These visualizations were then overlaid with ground truth annotations of respiratory cycles within the sample to validate model focus and ensure alignment with

clinically relevant features. Figure 1 demonstrates one such visualization.

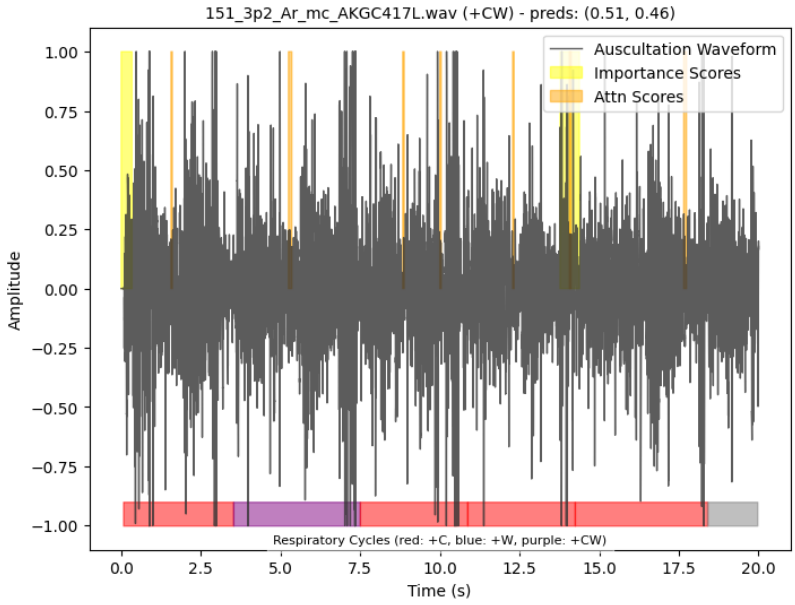

Figure 1.
The raw waveform of a test set sample is shown, with attention and importance scores highlighted above the zero axis to indicate key regions of interest. Respiratory cycle annotations are color-coded at the bottom for clarity: red represents the presence of crackles, blue indicates wheezes, purple signifies both, and gray denotes the absence of either. Model predictions are noted as (P(crackles), P(wheezes)).

In the example above, it is challenging to draw any definitive conclusions as some of the attended regions align with positive respiratory cycles while others do not. Further model refinement might be needed before these interpretability techniques can provide reliable clinical insights. Nevertheless, these visualizations could still offer value when examined alongside the corresponding audio recordings. Additionally, visualizing these regions in spectrogram transforms— which provide richer and more interpretable visual information—may facilitate deeper analysis. Finally, a comprehensive analysis of embeddings generated from the identified "important" regions could reveal underlying patterns through unsupervised clustering experiments. When combined with expert physician-guided evaluation, these approaches have the potential to enhance our understanding of the relationship between auscultatory findings and the underlying physiological processes.

Conclusions

This study demonstrates the effectiveness of leveraging pre-trained models, such as wav2vec 2.0, for respiratory auscultation. The introduction of domain-specific noise as a simple augmentation technique did not lead to any notable model performance improvements over baseline levels, highlighting the need for more robust augmentation strategies. Future efforts should focus on incorporating a diverse range of standard augmentation techniques—such as pitch shifting, time stretching, and reverb addition—while also exploring novel approaches like synthetic noise generation to better capture the complexity of real-world clinical conditions.

References

1. Bohadana, Abraham, Gabriel Izbicki, and Steve S. Kraman. "Fundamentals of lung auscultation." *New England Journal of Medicine* 370.8 (2014): 744-751.
2. Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.
3. Rocha BM et al. (2019) "An open access database for the evaluation of respiratory sound classification algorithms" Physiological Measurement 40 035001
4. Shams Nafisa Ali, and Samiul Based Shuvo. (2021). Hospital Ambient Noise Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2173743
5. Heitmann, Julien, et al. "DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries." *NPJ digital medicine* 6.1 (2023): 104.
6. Kamson, Alex Paul, et al. "Exploring Wav2vec 2.0 Model for Heart Sound Analysis." *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024.
7. Abnar, S., and W. Zuidema. "Quantifying attention flow in transformers. arXiv 2020." *arXiv preprint arXiv:2005.00928* (2022).