Interspeech 2024
1-5 September 2024, Kos, Greece

BTS: Bridging Text and Sound Modalities for Metadata-Aided Respiratory Sound Classification

June-Woo Kim[*1,2], *Miika Toikkanen*[2], *Yera Choi*[3], *Seoung-Eun Moon*[3†], *Ho-Young Jung*[1†]

[1]Department of AI, Kyungpook National University, South Korea
[2]RSC LAB, MODULABS, South Korea
[3]NAVER Digital Healthcare LAB, NAVER AI LAB, NAVER Cloud, South Korea

{kaen2891, hoyjung}@knu.ac.kr seongeun.moon@navercorp.com

Abstract

Respiratory sound classification (RSC) is challenging due to varied acoustic signatures, primarily influenced by patient demographics and recording environments. To address this issue, we introduce a text-audio multimodal model that utilizes metadata of respiratory sounds, which provides useful complementary information for RSC. Specifically, we fine-tune a pretrained text-audio multimodal model using free-text descriptions derived from the sound samples' metadata which includes the gender and age of patients, type of recording devices, and recording location on the patient's body. Our method achieves state-of-the-art performance on the ICBHI dataset, surpassing the previous best result by a notable margin of 1.17%. This result validates the effectiveness of leveraging metadata and respiratory sound samples in enhancing RSC performance. Additionally, we investigate the model performance in the case where metadata is partially unavailable, which may occur in real-world clinical setting.

Index Terms: Respiratory Sound Classification, Pretrained Language-Audio Model, ICBHI, Metadata

1. Introduction

Identifying abnormal respiratory sounds is pivotal for diagnosing and providing timely interventions for respiratory conditions. Automated detection of abnormal respiratory sounds has great potential to improve health and quality of life for those affected by respiratory diseases by identifying risks early and expediting first aid for potentially life-threatening conditions, such as pneumonia or chronic obstructive pulmonary disease. Machine learning approaches have been regarded as a promising way for automated detection of abnormal respiratory sounds. Recently, a number of studies [1, 2, 3, 4, 5, 6, 7, 8, 9] have tackled the respiratory sound classification (RSC) task and notably increased the performance by utilizing models that have been pretrained on large non-medical datasets [10, 11], and then fine-tuned on a respiratory sound dataset [12].

Nevertheless, the inherent heterogeneity of respiratory sound data presents an obstacle to further performance improvement in RSC. The heterogeneity arises from differences in patient demographics, recording devices, and environmental conditions, which can significantly impact the acoustic properties of respiratory sounds [1]. This may lead to poor generalization on unseen data, particularly in cases underrepresented by the training data. ICBHI [12], one of the widely adopted respiratory sound datasets, provides metadata that associates the recorded audio with attributes of patients and recording environments. Such metadata may be useful for addressing difficulties caused by heterogeneity.

Some previous work has adapted the metadata associated with respiratory sound for RSC to mitigate the heterogeneity issue. For instance, incorporating demographic information of patients such as age and gender into the pretraining process provides better representations of respiratory audio samples [5]. Moreover, metadata concerning the recording environment (i.e., stethoscope) also provides useful information. SG-SCL [1] employed domain-transfer techniques to reduce the effect of heterogeneity by regarding different types of recording devices as distinct domains. Despite the potential benefits of leveraging the metadata, these previous works did not fully incorporate it as text data into the model inputs.

Recent developments in multimodal models, exemplified by Contrastive Language-Image Pretraining (CLIP) [13] for text and image data and Contrastive Language-Audio Pretraining (CLAP) [14, 15] for text and audio data, offer a flexible framework for integrating text data with non-textual data. Several studies [16, 17, 18, 19] have demonstrated the effectiveness of language-EEG multimodal models for the sentiment classification and EEG-to-text decoding tasks. Recognizing the success of multimodal models and the demonstrated benefits of multimodal data in healthcare tasks, it is compelling to consider them for RSC, where such method has not yet been explored.

In this paper, we take a step into a new direction and fully make use of the respiratory audio metadata by adapting a text-audio multimodal model, aiming not only to leverage the metadata as an additional learning signal, but to benefit from the further context during the inference stage. Building on the foundation of contrastive language-audio pretrained models, our work incorporates the respiratory audio metadata alongside the sound recordings. To this end, we format the patient's metadata into descriptions derived from key attributes including age, gender, recording device, and recording location on the body, and encode them with respiratory sound data into shared feature representation by the pretrained encoders. With these joint representations, we train a classification head for the RSC task.

Our approach, which we name the ***BTS*** (***B***ridging the ***T***ext and ***S***ound modalities), a method that leverages multimodal text-audio model to fully exploit the potential of respiratory audio metadata, achieves the state-of-the-art (SOTA) result on the ICBHI dataset, outperforming upon the previous best [4] by 1.17%. Our results reveal the capability of contrastive language-audio pretraining to improve RSC both in audio-only and multimodal settings. Moreover, we demonstrate that our method retains its performance gains in the absence of metadata during the inference. This result suggests that our approach can be adopted for practical clinical settings where additional information other than audio signals may be unavailable. Our main

---
* work done while intern at NAVER Cloud †corresponding authors
Code is available at https://github.com/kaen2891/bts.

10.21437/Interspeech.2024-492

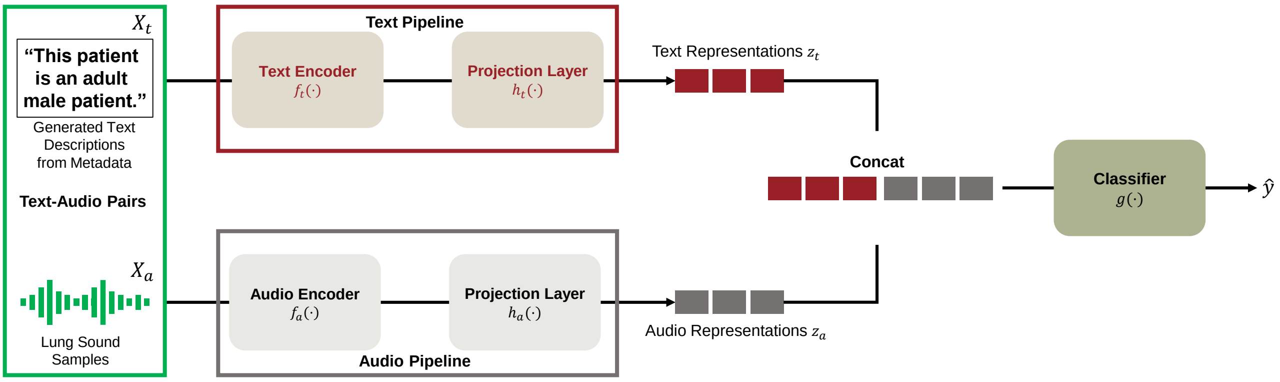

Figure 1: *An overall illustration of the proposed BTS architecture. The pretrained text and audio encoders extract feature representations of text description derived from metadata and respiratory sound samples, respectively. After the projection, the representations are integrated by a concatenation operation and used for RSC.*

contributions are as follows:

- We show that leveraging metadata of respiratory sounds improves the RSC performance. Our approach sets the new SOTA performance on the ICBHI dataset.
- We thoroughly explore ways to utilize the metadata considering a real clinical setting where the type of metadata differs from the expectation, or the metadata is partially or totally unavailable. We demonstrate that our method robustly performs RSC in such scenarios.
- We analyze how the different types of metadata affect the performance of our model. Our result shows that the information about the recording environment, such as the type of recording stethoscopes and recording locations of the human body are particularly helpful in minimizing the effect of heterogeneity in respiratory sounds.

2. Method

We introduce ***B***ridging ***T***ext and ***S***ound modalities (BTS), an approach that leverages multimodal text-audio model to fully exploit the potential of respiratory audio metadata. To mitigate the heterogeneity of respiratory sounds, we propose to explicitly utilize the metadata, which we expect to capture the significant sources of acoustic variability. By integrating this metadata, we aim to reduce the heterogeneity issue and improve the RSC performance. Toward this goal, we propose the adoption of a multimodal text-audio model for RSC, as depicted in Figure 1.

2.1. CLAP Model

While the metadata of respiratory sounds can be employed for RSC in several different ways, a free text format is flexible and easily applicable to human-produced data such as medical records. For instance, the metadata can be described by a vector of numeric values where each element indicates different metadata attributes. However, this approach is usually vulnerable to changes in input sources, such as missing data and unseen data types. In contrast, encoders for free text data are trained to understand the given input, which makes approaches utilizing input data in a free text format robust to the changes. For this reason, we use CLAP (Contrastive Language-Audio Pretraining) [15] as our starting point. The CLAP model includes both text and audio encoders, which are trained on the large-scale LAION-Audio-630K [15] dataset including diverse audio data.

Table 1: *Examples of generated text descriptions derived from metadata. 'All' is the case that includes all attributes: age, sex, recording location, and recording device.*

Metadata	Generated text descriptions
Age	This patient is an adult patient.
Sex	This patient is a male patient.
Loc	This sound was recorded from the left anterior chest.
Dev	This sound was recorded with a Meditron stethoscope.
Age-Loc-Dev	This sound was recorded from the left anterior chest of an adult patient, using a Meditron stethoscope.
....	....
All	This sound was recorded from the left anterior chest of an adult male patient, using a Meditron stethoscope.

Given the text and audio data denoted as X_i^t and X_i^a where $i \in [1, N]$ indicates the data index within a batch of size N, CLAP processes the text and audio data independently through dedicated encoders $f_t(\cdot)$ and $f_a(\cdot)$ for each modality. The embedding vectors produced by the encoders are projected onto a d-dimensional shared embedding space through projection layers $h_t(\cdot)$ and $h_a(\cdot)$.

$$\begin{aligned} z_t &= h_t(f_t(X_i^t)), \\ z_a &= h_a(f_a(X_i^a)). \end{aligned} \tag{1}$$

The CLAP model is trained to maximize the similarity between the text and audio embeddings by contrasting them with negative samples (i.e., mismatched text or audio embeddings obtained from $X^t_{j\in[1,N];j\neq i}$ or $X^a_{j\in[1,N];j\neq i}$).

2.2. Text Description Generation for Metadata

Among the metadata available in the ICBHI [12] dataset, we choose four types of data as follows: age (adult or pediatric) and gender (male or female) of patients, recording location on the chest of the patients (trachea, anterior left, anterior right, posterior left, posterior right, lateral left, or lateral right), and type of recording devices (Meditron, LittC2SE, Litt3200, or AKGC417L). Using the attributes, we construct simple text descriptions. A generated description can include any combination of the attributes, totaling 644 unique texts. Table 1 illustrates a few examples with different combinations of metadata.

2.3. Bridging Text and Sound Modalities

As shown in Figure 1, we train the text and audio encoders of CLAP for RSC by using the respiratory sound samples and generated text descriptions. For classification, we concatenate text and audio representations z_t and z_a from text and audio pipelines as described in Figure 1. Consequently, we can obtain the multimodal combined representations $z = concat(z_t, z_a)$ where $z \in \mathbb{R}^{N \times 2d}$. We then simply add a 4-dimensional linear layer for classifier $g(\cdot)$ followed by softmax function and train it with the Cross-Entropy loss $\mathcal{L}_{\text{CE}}$ (division by N is omitted):

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{n} y_i \log(\hat{y_i}), \tag{2}$$

where n is number of samples, y is the respiratory sound label $\in$ {normal, crackle, wheeze, both}, and $\hat{y}$ is the predicted probabilities obtained by the classifier.

3. Experimental Setup

3.1. Dataset

We utilized the ICBHI Respiratory dataset [12]. The dataset contains a total of approximately 5.5 hours of respiratory sound recordings with pre-defined and balanced splits for training (60%) and test (40%) without patient overlap. There are 4,142 training and 2,756 testing respiratory cycles across four classes. Table 2 illustrates the details of the ICBHI dataset. We binarize the age as the adult (over 18 years old) or pediatric (18 years old or under) for simplicity. Other than the age, we follow the metadata information of the official ICBHI records. Body mass index (BMI) data, which was provided only for adult patients, are solely employed for further analysis. For non-adults, we calculated it using their weight and height data.

3.2. Training Details

Following the data pre-processing described in [1, 3, 4, 9], we extracted the respiratory cycles from the waveform samples and standardized them to have a duration of 8 seconds. We then conducted resampling to 48kHz to match the pretraining data of CLAP. We employed the CLAP [15] model pretrained on the LAION-Audio-630K [15] dataset for all experiments. The maximum length of the text descriptions is limited to 64 tokens, which was sufficient for avoiding truncation of text. We fine-tuned the models using the Adam optimizer [20] with an initial learning rate of 5e–5. The learning rate was adjusted by cosine scheduling through a total of 50 epochs of training with a batch size of 8. To reduce the impact of random initialization, we conducted the experiments with five different random seeds.

3.3. Metrics

We adapt the *Specificity* (S_p), *Sensitivity* (S_e), and their average ($Score$) as performance metrics for RSC, following the definitions in [12]. All reported values of S_p, S_e, and $Score$ are the mean and variance from the five runs with different seeds.

3.4. Baselines

We compare the proposed method with previous studies including the current SOTA method [4], which uses Audio Spectrogram Transformer (AST) [21] as a backbone model. We also consider the result based solely on the audio embedding of CLAP (z_a in Equation (1)) as an additional baseline, which we denote as Audio-CLAP.

Table 2: *Details of the ICBHI dataset including the number of audio samples for each class and the types of metadata. L/R stands for left or right.*

	Label	Train	Test	Sum
Lung Sound	Normal	2,063	1,579	3,642
	Crackle	1,215	649	1,864
	Wheeze	501	385	886
	Both	363	143	506
	Type	**Metadata Label**		
Metadata	Age	Adult, Pediatric		
	Sex	Male, Female		
	Location	Trachea, L/R Anterior, L/R Posterior, L/R Lateral		
	Stethoscope	Meditron, LittC2SE, Litt3200, AKGC417L		
	Others	BMI (Adult only), Weight/Height (Pediatric only)		

4. Results

4.1. Main Results

Table 3 presents comprehensive ICBHI results including our method. Our method achieves a new SOTA by 1.17% improvement from the previous best without using any additional training techniques that other methods rely on, such as stethoscope-specific fine-tuning [9], co-tuning [6], Patch-Mix augmentation [4], or domain adaptation techniques [1, 3]. Particularly, it is noteworthy that our method has a considerably higher sensitivity (S_e) than the previous best model while maintaining a similar specificity (S_p). We consider that the additional context from textual descriptions enhances the model's ability to correctly identify positive cases, without increasing the false positive ratio. Additionally, the improvement of Audio-CLAP over the previous SOTA highlights the effectiveness of the contrastively language-audio pretrained encoder as a strong baseline for audio-only related tasks, where the encoder is pretrained using text descriptions as opposed to previous backbone models for RSC that employed categorical audio labels for pretraining.

Table 4 compares the performance of BTS and Audio-CLAP for each metadata category within the ICBHI test set. Although the dataset is notably imbalanced for the metadata categories, our model consistently surpasses the Audio-CLAP baseline across all classes. It is also noteworthy that a notable enhancement is observed in minority classes. Specifically, the result of the test samples, of which *Loc* is the right lateral chest, yields the Score increase of 5.77%. This underscores the value of the metadata in not only the overall performance improvement but also the effectiveness of accounting for underrepresented categories.

4.2. Influence of Metadata on Classification Performance

We analyze the impact of metadata on classification performance by comparing the results of three distinct experiment settings: (1) the full set of metadata with BTS, (2) subset with exclusion of a single attribute with BTS, and (3) audio-encoder only in the case of Audio-CLAP. The results are summarized in Table 5, which shows that more textual context results in higher performance. Specifically, using all metadata in (1) yields the highest Score, while audio-encoder only (3) scored the lowest. The results of all metadata subsets (2) fall in between them.

Furthermore, the results demonstrate that the measurement location (*Loc*) and recording device type (*Dev*) have a larger effect than the demographic attributes, i.e., age and gender of patients. The absence of *Loc* and *Dev* leads *Score* drop of 0.78% and 0.88%, respectively, compared to the all metadata case. This suggests that the type of recording device and the

Table 3: *The RSC performance on the ICBHI dataset with the official 60–40% train–test split. Here, in the Pretraining Data column, IN, AS, and LA refer to ImageNet [10], AudioSet [11], and LAION-Audio-630K [15], respectively. $*$ denotes the previous state-of-the-art ICBHI Score. The* ***Best*** *and* <u>*second best*</u> *results are highlighted by the bold characters and underlines.*

Method	Backbone	Pretraining Data	Venue	S_p (%)	S_e (%)	**Score** (%)
SE+SA [22]	ResNet18	-	*INTERSPEECH'20*	81.25	17.84	49.55
LungRN+NL [23]	ResNet-NL	-	*INTERSPEECH'20*	63.20	41.32	52.26
RespireNet [9] (CBA+BRC+FT)	ResNet34	IN	*EMBC'21*	72.30	40.10	56.20
Chang *et al.* [24]	CNN8-dilated	-	*INTERSPEECH'22*	69.92	35.85	52.89
Ren *et al.* [25]	CNN8-Pt	-	*ICASSP'22*	72.96	27.78	50.37
Wang *et al.* [8] (Splice)	ResNeSt	IN	*ICASSP'22*	70.40	40.20	55.30
Late-Fusion [7]	Inc-03 + VGG14	IN	*EMBC'22*	**85.60**	30.00	57.30
Nguyen *et al.* [6] (StochNorm)	ResNet50	IN	*TBME'22*	78.86	36.40	57.63
Nguyen *et al.* [6] (CoTuning)	ResNet50	IN	*TBME'22*	79.34	37.24	58.29
Moummad *et al.* [5]	CNN6	AS	*WASPAA'23*	70.09	40.39	55.24
Moummad *et al.* [5] (SCL)	CNN6	AS	*WASPAA'23*	75.95	39.15	57.55
Bae *et al.* [4] (Fine-tuning)	AST	IN + AS	*INTERSPEECH'23*	77.14	41.97	59.55
Bae *et al.* [4] (Patch-Mix CL)	AST	IN + AS	*INTERSPEECH'23*	81.66	43.07	62.37*
Kim *et al.* [3] (AFT on Mixed-500)	AST	IN + AS	*NeurIPSW'23*	80.72	42.86	61.79
Kim *et al.* [1] (SG-SCL)	AST	IN + AS	*ICASSP'24*	79.87	43.55	61.71
Kim *et al.* [2] (RepAugment)	AST	IN + AS	*EMBC'24*	<u>82.47</u>	40.55	61.51
Audio-CLAP [ours]	CLAP	LA	*INTERSPEECH'24*	$80.85_{\pm 3.33}$	<u>44.67</u>$_{\pm 3.77}$	<u>62.56</u>$_{\pm 0.37}$
BTS [ours]	CLAP	LA	*INTERSPEECH'24*	$81.40_{\pm 2.57}$	**45.67**$_{\pm 2.66}$	**63.54**$_{\pm 0.80}$

Table 4: *A comparison of the ICBHI Scores between the Audio-CLAP baseline and BTS. The results are shown depending on the metadata classes. Note that there is no sample of the LittC2SE in the test set. The bold characters and underlines indicate the* ***best*** *and* <u>*second best*</u> *Score improvement.*

Metadata Type	Metadata Class	Ratio (%)	Method BTS	Method Audio-CLAP	Score Difference
Age	Adult	85.70	64.53	61.67	2.86
	Pediatric	14.30	64.53	61.99	2.54
Sex	Male	78.74	64.53	62.00	2.53
	Female	21.26	64.46	61.92	2.54
Loc	Trachea	11.97	64.46	61.92	2.54
	Left Anterior	21.99	64.52	61.66	2.86
	Right Anterior	9.51	64.78	61.58	3.20
	Left Posterior	22.57	64.54	62.00	2.54
	Right Posterior	15.64	65.31	61.45	<u>3.86</u>
	Left Lateral	9.43	64.41	61.83	2.58
	Right Lateral	8.89	60.21	54.44	**5.77**
Dev	Meditron	16.65	64.54	62.00	2.54
	LittC2SE	0.0	-	-	-
	Litt3200	16.73	64.52	61.65	2.87
	AKGC417L	66.62	64.78	61.53	3.25

Table 5: *Results of the ablation study with different combinations of the metadata. The* ***best*** *result is indicated by the bold.*

Method	Setting	Metadata	S_p (%)	S_e (%)	**Score** (%)
BTS	(1)	All	$81.40_{\pm 2.57}$	$45.67_{\pm 2.66}$	**63.54**$_{\pm 0.80}$
	(2)	Age-Sex-Loc	$81.71_{\pm 4.12}$	$43.63_{\pm 3.48}$	$62.66_{\pm 0.35}$
	(2)	Age-Sex-Dev	$79.49_{\pm 3.66}$	$46.04_{\pm 2.39}$	$62.76_{\pm 1.09}$
	(2)	Age-Loc-Dev	$82.28_{\pm 5.27}$	$43.48_{\pm 4.38}$	$62.88_{\pm 0.83}$
	(2)	Sex-Loc-Dev	$84.66_{\pm 3.63}$	$41.09_{\pm 3.37}$	$62.88_{\pm 0.54}$
Audio-CLAP	(3)	-	$80.85_{\pm 3.33}$	$44.67_{\pm 3.77}$	$62.56_{\pm 0.37}$

measurement location significantly influence the acoustic properties of respiratory sounds. Therefore, the combined use of these metadata provides the meaningful context to understand the respiratory sounds.

4.3. Unknown Metadata Scenario

To probe how well the model generalizes to unseen text descriptions, we examined the model performance for unseen test data, which additionally includes a new metadata attribute that is not used for training. Specifically, we added a sentence to describe the BMI of patients to the test data. The additional sentence is written in the same style as the training descriptions, e.g., "The BMI of the patient was 20.50.". The evaluation result is denoted as BTS[BMI] in Table 6. Adding unknown metadata to text descriptions at test time shows only minor performance degradation, which suggests that the model performs reliably even with the unexpected metadata.

Table 6: *Results with variations on the metadata. The variations include the additional BMI attribute, partial metadata, and no metadata in text description. The* ***Best*** *result.*

Method	S_p (%)	S_e (%)	**Score** (%)
BTS	$81.40_{\pm 2.57}$	$45.67_{\pm 2.66}$	**63.54**$_{\pm 0.80}$
BTS[BMI]	$81.40_{\pm 2.57}$	$45.66_{\pm 2.65}$	$63.53_{\pm 0.80}$
BTS[Partial Metadata]	$81.29_{\pm 2.55}$	$45.54_{\pm 2.61}$	$63.41_{\pm 0.78}$
BTS[No Metadata]	$80.82_{\pm 3.54}$	$45.59_{\pm 2.59}$	$63.21_{\pm 1.04}$
Audio-CLAP	$80.85_{\pm 3.33}$	$44.67_{\pm 3.77}$	$62.56_{\pm 0.37}$

4.4. Missing Metadata Scenarios

To understand how metadata that is partially or entirely missing effect the model, we conducted two experiments. We first partially removed the metadata (BTS[Partial Metadata]), by randomly eliminating one of the metadata attributes from test samples and substituting "Unknown" with 10% probability. Then, we entirely removed the metadata (BTS[No Metadata]) and replace the whole description by "No description." Table 6 describes the experiment results. As expected, BTS[Partial Metadata] shows a slightly degraded Score compared to BTS, while BTS[No Metadata] results in a relatively large performance reduction. Nevertheless, the results with missing metadata maintain an edge over Audio-CLAP. The results show that the BTS model is robust to missing metadata. We conjecture that the model learns to infer certain metadata characteristics directly from the audio, thereby preserving its strong performance even in the absence of metadata during inference.

5. Conclusion

In this work, we proposed to directly utilize the metadata to improve the performance of RSC. Our experiments demonstrated that including the metadata as additional context for the classification leads to a considerable performance increase, which results in the new SOTA for RSC. In particular, the experiment results showed that our method helps minimize the performance degradation due to the acoustic variations induced by the inherent factors relating to the demographics and recording environment. Moreover, our method works reliably even when the metadata is with unexpected information, partially unavailable, or even completely unavailable. Besides, we showed that the CLAP model provides a strong baseline for audio-only tasks.

6. Acknowledgement

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

7. References

[1] J.-W. Kim, S. Bae, W.-Y. Cho, B. Lee, and H.-Y. Jung, "Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1431–1435.

[2] J.-W. Kim, M. Toikkanen, S. Bae, M. Kim, and H.-Y. Jung, "Repaugment: Input-agnostic representation-level augmentation for respiratory sound classification," *arXiv preprint arXiv:2405.02996*, 2024.

[3] J.-W. Kim, C. Yoon, M. Toikkanen, S. Bae, and H.-Y. Jung, "Adversarial fine-tuning using generated respiratory sound to address class imbalance," *arXiv preprint arXiv:2311.06480*, 2023.

[4] S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun, "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification," in *Proc. INTERSPEECH 2023*, 2023, pp. 5436–5440.

[5] I. Moummad and N. Farrugia, "Pretraining respiratory sound representations using metadata and contrastive learning," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[6] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022.

[7] L. Pham, D. Ngo, K. Tran, T. Hoang, A. Schindler, and I. McLoughlin, "An ensemble of deep learning frameworks for predicting respiratory anomalies," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022, pp. 4595–4598.

[8] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9017–9021.

[9] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[12] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, 2018, pp. 33–37.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] N. Hollenstein, C. Renggli, B. Glaus, M. Barrett, M. Troendle, N. Langer, and C. Zhang, "Decoding eeg brain activity for multimodal natural language processing," *Frontiers in Human Neuroscience*, p. 378, 2021.

[17] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 5, 2022, pp. 5350–5358.

[18] Y. Duan, C. Chau, Z. Wang, Y.-K. Wang, and C.-t. Lin, "Dewave: Discrete encoding of eeg waves for eeg to text translation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, 2023, pp. 9907–9918.

[19] J. Qiu, W. Han, J. Zhu, M. Xu, D. J. Weber, B. Li, and D. Zhao, "Can brain signals reveal inner alignment with human languages?" in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[22] Z. Yang, S. Liu, M. Song, E. Parada-Cabaleiro, and B. W. Schuller, "Adventitious respiratory classification using attentive residual neural networks," in *Interspeech*, 2020.

[23] Y. Ma, X. Xu, and Y. Li, "Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation." in *Interspeech*, 2020, pp. 2902–2906.

[24] Y. Chang, Z. Ren, T. T. Nguyen, W. Nejdl, and B. W. Schuller, "Example-based Explanations with Adversarial Attacks for Respiratory Sound Analysis," in *Proc. Interspeech 2022*, 2022, pp. 4003–4007.

[25] Z. Ren, T. T. Nguyen, and W. Nejdl, "Prototype learning for interpretable respiratory sound analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9087–9091.