PulmoScan: A Practical Pulmonary Disease Pre-Screening System

Baixu Yan, Shijia Ge, Meizi Lu, Weixiang Zhang, Shuzhao Xie, Zhi Wang†
Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
{ybx22, gsj23, lu-mz22, zhang-wx22, xsz24}
@mails.tsinghua.edu.cn,
wangzhi@sz.tsinghua.edu.cn

***Abstract*—Automation of pulmonary disease identification has been a long-standing area of research and gained increased attention after the COVID-19 pandemic. However, existing respiratory sound classification algorithms exhibit significant limitations, including suboptimal performance, insufficient input robustness, and inadequate alignment with clinical evaluation metrics, thereby hindering their practical implementation. To address these limitations, we introduce *PulmoScan*, a practical pulmonary disease pre-screening system. *PulmoScan* comprises three fundamental modules: a *Respiratory Sound Quality Validator* that ensures the robustness of input data, a *Runtime Decision Booster* that improves performance and adapting to variating evaluation metrics, and a *Symptom Enhancement Diagnoser* that augments respiratory sound classification with comprehensive disease pre-screening capabilities. Beyond its primary function, *PulmoScan* exemplifies a methodological framework for translating theoretically limited algorithms into viable clinical applications, demonstrating essential considerations and procedural adaptations for real-world implementation.**

***Index Terms*—Application for Healthcare, Application of Large Language Models, Pulmonary Disease Pre-Screening, Runtime Category Decision Algorithm, Out-of-Distribution Detection**

I. Introduction

Respiratory diseases represent a leading cause of global mortality [1], with early diagnosis playing a pivotal role in mitigating disease transmission [2]. Stethoscope auscultation is a cost-effective and non-invasive method of diagnosing pulmonary diseases, but it presents challenges: it requires trained professionals and can lead to subjective and variable interpretations. These problems are exacerbated in resources-limited settings, especially during pandemics such as COVID-19, where the shortage of medical experts further complicates the timely and accurate diagnosis.

There has been extensive research [2]–[15] on automated respiratory sound recognition, but these efforts are still far from being used as practical disease diagnosis or screening systems, primarily due to the following limitation:

a) ***Lack of Input Robustness****:* In screening systems, users typically lack expertise in proper stethoscope operation, potentially generating suboptimal respiratory sound recordings that compromise diagnostic accuracy. Consequently, the detection of low-quality audio signals becomes imperative for reliable diagnosis. Thus, recognizing the low-quality audio is necessary.

b) ***Insufficient Input Modal Dimensions****:* Exclusive reliance on acoustic features such as crackles and wheezes in abnormal respiratory sounds proves inadequate for comprehensive diagnosis. Clinical practice demonstrates that physicians integrate multiple symptomatic indicators for thorough assessment. Therefore, a robust diagnostic model should incorporate both the patient's symptomatic profile, emulating clinical judgment, and respiratory sound analysis results to ensure comprehensive evaluation.

c) ***Inaccurate Metric Evaluation & Inadaptability to Real-World Changes****:* Most respiratory sound classification algorithms rely on fixed evaluation metrics, such as the score in ICBHI [16]. However, this approach is insufficient for practical applications. In real-world scenarios, the relative importance of sensitivity (Se) and specificity (Sp) varies. Sometimes, accurately identifying abnormal cases is more critical than correctly classifying normal ones, while in other contexts it is essential to avoid false positives. The relative importance of diagnostic factors can be modulated by multiple parameters, including resource constraints, transmission dynamics, and other contextual variables. Existing algorithms do not perform well in Se, limiting their practical implementation. Furthermore, as conditions evolve, this relative importance can change, making models optimized for specific metrics less effective over time.

To address the aforementioned challenges, we design a robust pulmonary disease screening system that can handle multimodal input, comprising three distinct modules:

1. Respiratory Sound Quality Validator. This validator determines whether the recorded audio qualifies as a test sample for the detection algorithm. We developed a manifold adapter to take advantage of CLAP's [17] general knowledge for the qualification check. Through training the adapter on a limited dataset of representative qualified and unqualified specimens

† Corresponding author. This work is supported by National Natural Science Foundation of China (Grant No. 62472249), and Shenzhen Science and Technology Program (Grant No. JCYJ20220818101014030 and KJZD20240903102300001). We thank Shenzhen HealAll Technology Co., Ltd for providing medical resources for this research. Company's website: https://healall.net/.

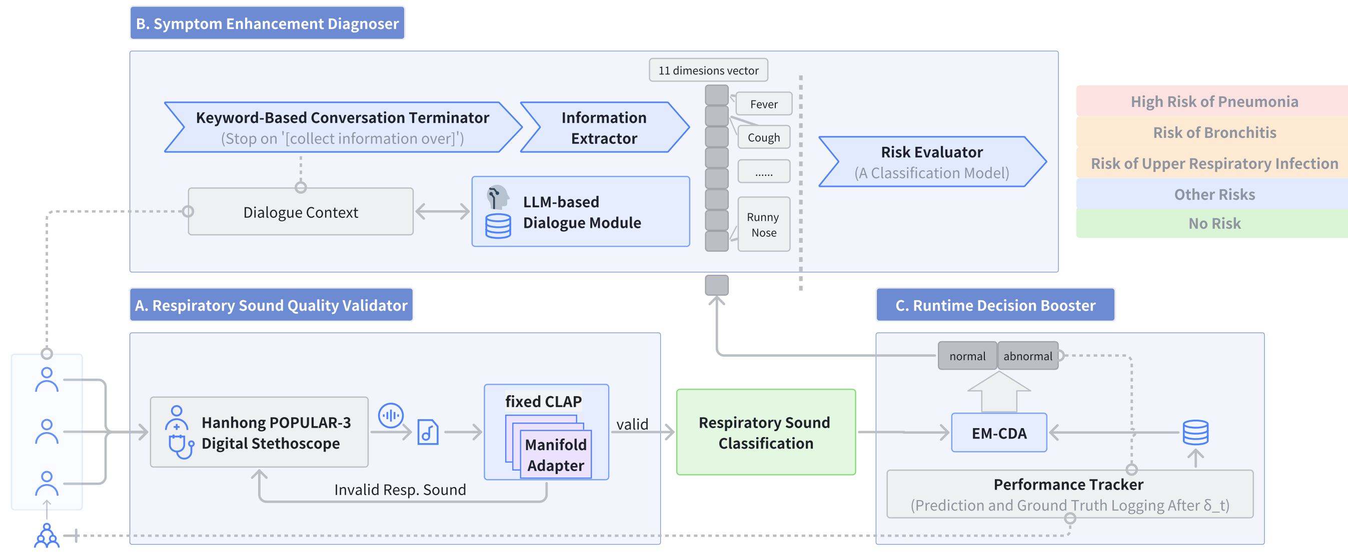


Fig. 1: The Architecture of PulmoScan. PulmoScan classifies individuals into five risk categories. We design a Respiratory Sound Validator and a Runtime Decision Booster to overcome the drawbacks of current respiratory sound classification methods. Besides, we also designs a Symptom Enhancement Diagnoser, which uses patients' other symptoms to assist in evaluating pulmonary disease risk.

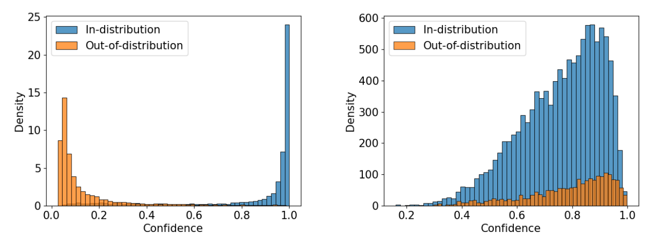
utilizing CLAP embeddings, this methodology achieves 95% precision within the established domain.

2. Symptom Enhancement Diagnoser. We select 11 additional features to enhance the final diagnosis. This Symptom Enhancement Diagnoser consists of an Auto Symptom Collector and a Risk Evaluator. It is worth noting that both the symptom inquiry and vector extraction in our system are fully automated by Large Language Models (LLM), significantly reducing the labor costs involved in the prescreening process.

3. Runtime Decision Booster. By default, the category with the highest probability is selected as the final decision. However, in disease screening contexts, false negative outcomes (misclassifying pathological cases as normal) incur substantially higher costs than false positive outcomes (misclassifying normal cases as pathological), with cost implications varying across geographical regions. To address this, we propose a Runtime Decision Booster that includes a Performance Tracker and an Expectation Maximization Category Decision Algorithm (EM-CDA). The Performance Tracker logs prediction-truth pairs after a delay, enabling real-time performance monitoring, while the EM-CDA selects the category that maximizes expected gain based on a customizable metric, which can be adjusted to reflect the relative costs of false positives and false negatives. This approach is non-intrusive to the model's training, enabling compatibility with any probabilistic classifier while optimizing decision boundaries for context-specific screening requirements, thus enhancing classification robustness and minimizing critical errors.

II. System Design and Architecture

PulmoScan, illustrated in Fig. 1. , classifies individuals into five categories: High Risk of Pneumonia, Risk of Bronchitis, Risk of URI, Other Risks, and No Risk. The classification is based on the detection of respiratory sounds and observable symptoms. The system comprises three key components: the Respiratory Sound Quality Validator (Section II-A), the Symptom Enhancement Diagnoser (Section II-B), and the Runtime Decision Booster (Section II-C), which together ensure accurate and context-aware disease screening.



(a) Image Scenario (b) Respiratory Sound Scenario

Fig. 3: The Confidence Distribution using Confidence Scaling method on ID and OOD data. We cannot find a proper detection threshold in the Respiratory Sound Classification Scenario.

A. Respiratory Sound Quality Validator

We designed a validator to ensure that the recorded audio is a qualified respiratory sound. Initial attempts using Out-of-Distribution (OOD) approaches, such as confidence-based methods from [22] (referenced in [25]), proved ineffective. As shown in TABLE I and Fig. 3, confidence scaling resulted in a detection error of 0.4969 on the best threshold, nearly equivalent to random guessing. The confidence score distributions for OOD and in-distribution (ID) samples were also nearly

TABLE I: Confidence Method Doesn't Work in Respiratory Scenario.

Scenario	Image				Respiratory Sound
Dataset	80% CIFAR-10 for training 20% CIFAR-10 [18] as a ID dataset for testing TinyImageNet[a] as an OOD dataset for testing				80% Private-LS[b] for training 20% Private-LS as a ID dataset for testing ESC50 [19] as an OOD dataset for testing
Method Used	baseline [20]	ODIN [21]	confidence [22]	conf. scaling [22]	conf. scaling [22]
TPR95↓	0.4073	0.2126	0.1784	0.1645	0.9479
Detection Error↓	0.1195	0.1029	0.0968	0.0919	0.4969
Best Threshold	0.9921	0.1007	0.3390	0.4586	0.4449
AUROC↑	0.9375	0.9569	0.9706	0.9734	0.4369
AUROC_IN↑	0.9461	0.9554	0.9737	0.9760	0.8124
AUROC_OUT↑	0.9224	0.9566	0.9694	0.9728	0.1361

[a] TinyImageNet is a subset of ImageNet [23], which contains 10000 images from 200 categories. The images will be downsampled to the size of 32 × 32 keeping the same to CIFAR's size.
[b] A Private Respiratory Sound Dataset collected by us.

TABLE II: The Performance of Manifold Adapter

			Sp	Se
Train	LungSound	20% Private-LS ICBHI [16]	98.25	94.03
	Non-LungSound	20% Private-NLS[c]		
Test	LungSound	80% Private-LS ICBHI	98.66	53.11
	Non-LungSound	AudioSet [24]		

[c] A Private Environmental Sound Dataset collected by us.

identical, making threshold-based OOD detection infeasible. This failure is likely due to the low accuracy of respiratory sound classification algorithms and limited data diversity.

Inspired by ZOC [26] and CLIPN [27], we developed a novel approach by affixing a manifold adapter to CLAP [17]. Trained on a small set of samples from each class, the manifold adapter accurately determines whether a sound is a proper respiratory sound and demonstrates OOD detection capability. During training, CLAP's weights remain fixed, enabling rapid retraining of the Manifold Adapter with updated deployment data.

We test the performance of this manifold adapter architecture, as shown in TABLE II. It turns out that the manifold adapter can achieve a decent classified accuracy with only 20% data as the train dataset. Even when facing samples from classes never occurred in training, it still shows 53.11% Se and 98.66% Sp.

B. Symptom Enhancement Diagnoser

We use an Auto Symptom Collector to collect additional obvious symptom information from individuals, combining the collected 11-dimensional symptom vector with respiratory sound detection results. This combined data is fed into a downstream classifier for disease risk classification. The system uses prompts to enable the LLM (GPT-4o) to actively inquire about symptoms and outputs an end marker `[collect information over]` once all information is gathered, terminating the conversation. The Conversation Information Extractor then extracts the gathered information using another prompt.

Prompt for Initializing Dialogue: *You are a pneumonia screening consultation expert, and you need to assume that I am a patient coming to you for a consultation. You should collect the following information from me through a conversational approach: whether I have a fever, whether I have a cough, if there is a fever, whether it has lasted for a long time, if there is a cough, whether it has persisted for more than 8 days, and whether symptoms such as sputum production, shortness of breath, facial cyanosis, tachypnea, retractions, nasal flaring, runny nose, or nasal obstruction have occurred. Please ensure that each question is not too long and guide the patient to share relevant information in a conversational manner. Once you have gathered all the necessary information, provide a closing remark [collect information over] and then exit the role-playing scenario.*

Prompt for Extracting Information: *Now, please organize your response according to the following output format, with no extra output:*{*'fever': True, 'cough': True, 'sputum production': True, 'shortness of breath': True, 'cough > 8 days': True, 'shortness of breath': True, 'facial cyanosis': True, 'tachypnea': True, 'retractions': True, 'nasal flaring': True, 'runny nose': True, 'nasal obstruction': True*}

C. Runtime Decision Booster

Given a probability vector from a probabilistic model, the category decision is typically made using $\arg\max_i \mathrm{prob}[i]$, where prob denotes a probability vector. Despite the default Category Decision Algorithm (CDA) maximizing overall expected accuracy, it may not always be the optimal solution, as accuracy is not always the primary concern. When using an ICBHI-like metric $(Sp + r \times Se)/(1 + r)$, where r represents the importance of Se, this method may not perform optimally under varying r.

We employ a Performance Tracker to log delayed $prediction - truth$ pairs and utilize EM-CDA (Algorithm 1) to improve performance on specific metrics. Testing the SOTA ICBHI algorithm [5] with EM-CDA (Fig. 4, with a delay of 8 samples), we observed a metric improvement that exceeded 6% when Se's importance was critical. This improvement grows as the importance of Se increases.

However, when Se's importance is around 1, our method may reduce the final evaluation score due to two factors:

Algorithm 1 EM-CDA
(Expectation Maximization Category Decision Algorithm)

procedure EM_CDA($prob[], past[]$, metric)
__prob[]__ is the probability vector output by upstream model; __past[]__ is a vector consists of pairs of category predicted and the ground truth where each pair represents a test sample detected before; __metric__ is a given function that takes a vector like past[] as parameter and outputs corresponding metric value.
  Declare an Array named $Expectation[]$
  for $i = 0$ **to** $\text{len}(prob[]) - 1$ **do**
    $p \leftarrow prob[i]$
    $m \leftarrow \text{metric}(past[])$
    $m_{after} \leftarrow \text{metric}(past[] + < i, i >)$
    $expectation\_vec[i] \leftarrow p * (m_{after} - m)$
    for $j = 0$ **to** $\text{len}(prob[]) - 1$ **do**
      if $i == j$ **then**
        pass
      end if
      $m_{after} \leftarrow \text{metric}(past[] + < i, j >)$
      $\delta_m \leftarrow \text{P}(truth = j | truth \neq i) * (m_{after} - m)$
      Add $(1 - p) * \delta_m$ to $Expectation[i]$
    end for
  end for
  return $\arg\max_i Expectation[i]$
end procedure

(1) EM-CDA relies on potentially inaccurate class probability vectors, and (2) selecting the class with the highest expected gain may deviate from actual outcomes in finite testing. Nonetheless, in our screening scenario, where Se's importance exceeds 1, the algorithm's benefits outweigh these biases.

III. Related Works

A. Out-of-Distribution Detection

In deep learning, the closed-world assumption [28] assumes that all test classes are observed during training. However, in our abnormal respiratory sound screening scenario, the inexperience of operators using stethoscopes breaks this assumption, requiring a module to verify if the input audio is a respiratory sound. For Out-of-Distribution (OOD) detection, a common approach is to use maximum softmax probability (MSP) as a threshold [20], with OOD samples expected to have lower MSP values than in-distribution (ID) samples. ODIN [21] enhances this method by applying temperature scaling and input perturbations to better distinguish between ID and OOD softmax score distributions. Other methods propose training models to output confidence scores [22], [29], or use ID-ness threshold-based approach like energy-based [30] and gradient-based [31] methods and so on. Additionally, generating synthetic OOD samples and adding them to the training set [32] is also a way to realize OOD detection. And there are also some zero-shot approaches [26], [27] using well-trained LLMs like CLIP or CLAP [17], [33] showing promising potential.

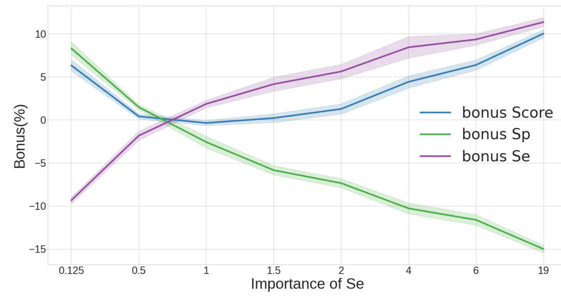


Fig. 4: Bonus gained on given importance of Se by using EM-CDA. $Score = (Sp + r \times Se)/(1 + r)$, where r denotes the importance of Se. This experiment, conducted with Patch-MixCL+AST [5] on the ICBHI [16] official train-test split, allows EM-CDA to receive feedback with an 8-sample delay. In scenarios where Se's importance is not around 1, EM-CDA significantly boosts performance.

B. Respiratory Sound Classification

The detection of abnormal respiratory sounds has been a long-standing research focus. In general, there are three main categories of methods that can accomplish this task. The first method is using the nolinear filter [13], or a tunable Q-factor wavelet transform [9] to try seperating crackle signal from the original respiratory sound. The second is the feature engineering approach. This kind of approach mainly focuses on finding better features, such as the spectral features [11], [12], the eigenvalue of singular spectrum analysis [34], MFCC [14], S-transform spectrogram [10] and so on. Third, latest researches tried deep learning methods, which take a spectrogram as the input and use a deep neural network to determine if it is an abnormal respiratory sound sample [2], [5]. These efforts primarily focus on overcoming the issue of insufficient data in respiratory sound datasets by employing data augmentation techniques or improving learning methods.

IV. Conclusion

This paper introduces PulmoScan, a system that bridges the gap between respiratory sound detection algorithms and practical pulmonary disease prescreening. Our key contributions lie in

- An accurate and cost-effective OOD detection module.
- An automated approach to collect individual data using large language models.
- A non-intrusive category decision algorithm that can make the model perform better under any given metric functions.

PulmoScan provides a robust and efficient solution for the prescreening of lung diseases. This system can also serve as a practical demonstration for other real-world healthcare scenarios.

References

[1] World Health Organization. The global impact of respiratory diseases (2nd edition). *Forum of International Respiratory Societies (FIRS)*, 2017.

[2] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 527–530. IEEE, 2021.

[3] Sandra Reichert, Raymond Gass, Christian Brandt, and Emmanuel Andrès. Analysis of respiratory sounds: State of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, page CCRPM.S530, Jan 2008.

[4] Renard Xaviero Adhi Pramono, Stuart Bowyer, and Esther Rodriguez-Villegas. Automatic adventitious respiratory sound analysis: A systematic review. *PLOS ONE*, page e0177926, May 2017.

[5] Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Seyoung Yun. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. In *24th International Speech Communication Association, Interspeech 2023*, pages 5436–5440. International Speech Communication Association, 2023.

[6] Naoki Asatani, Tohru Kamiya, Shingo Mabu, and Shoji Kido. Classification of respiratory sounds using improved convolutional recurrent neural network. *Computers & Electrical Engineering*, 94:107367, 2021.

[7] Ivan W Selesnick. Wavelet transform with tunable q-factor. *IEEE transactions on signal processing*, 59(8):3560–3575, 2011.

[8] Mohammed Bahoura. Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in biology and medicine*, 39(9):824–843, 2009.

[9] Sezer Ulukaya, Gorkem Serbes, and Yasemin P Kahya. Resonance based separation and energy based classification of lung sounds using tunable wavelet transform. *Computers in Biology and Medicine*, 131:104288, 2021.

[10] Rajkumar Palaniappan, Kenneth Sundaraj, Sebastian Sundaraj, N Huliraj, and SS Revadi. A telemedicine tool to detect pulmonary pathology using computerized pulmonary acoustic signal analysis. *Applied Soft Computing*, 37:952–959, 2015.

[11] Syed Osama Maruf, M Usama Azhar, Sajid Gul Khawaja, and M Usman Akram. Crackle separation and classification from normal respiratory sounds using gaussian mixture model. In *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, pages 267–271. IEEE, 2015.

[12] Plamen Bokov, Bruno Mahut, Patrice Flaud, and Christophe Delclaux. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. *Computers in biology and medicine*, 70:40–50, 2016.

[13] Mariko Ono, Kaoru Arakawa, Masashi Mori, Tsuneaki Sugimoto, and Hiroshi Harashima. Separation of fine crackles from vesicular sounds by a nonlinear digital filter. *IEEE transactions on biomedical engineering*, 36(2):286–291, 1989.

[14] Bor-Shing Lin and Bor-Shyh Lin. Automatic wheezing detection using speech recognition technique. *Journal of Medical and Biological Engineering*, 36:545–554, 2016.

[15] Sonia Charleston-Villalobos, Ramón González-Camarena, Georgina Chi-Lem, and Tomás Aljama-Corrales. Crackle sounds analysis by empirical mode decomposition. *IEEE Engineering in medicine and biology magazine*, 26(1):40, 2007.

[16] B. M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, R. P. Paiva, I. Chouvarda, P. Carvalho, and N. Maglaveras. A *Respiratory Sound Database for the Development of Automated Classification*, page 33–37. Jan 2018.

[17] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo 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)*, pages 1–5. IEEE, 2023.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA, 2015. Association for Computing Machinery.

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Learning,Learning*, Oct 2016.

[21] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *Cornell University - arXiv,Cornell University - arXiv*, Jun 2017.

[22] Terrance DeVries and GrahamW. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv: Machine Learning,arXiv: Machine Learning*, Feb 2018.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[25] Yuhan Chen, Pankaj Attri, Jeffrey Barahona, Michelle L. Hernandez, Delesha Carpenter, Alper Bozkurt, and Edgar Lobaton. Robust cough detection with out-of-distribution detection. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3210–3221, 2023.

[26] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576, 2022.

[27] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023.

[28] Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514, 2016.

[29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[31] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[32] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[33] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yanwen Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Cornell University - arXiv,Cornell University - arXiv*, Oct 2021.

[34] Semra İçer and Şerife Gengeç. Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds. *Digital Signal Processing*, 28:18–27, 2014.