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/EMBC53108.2024.10782697

Robust Multimodal Cough and Speech Detection using Wearables: A Preliminary Analysis

Yuhan Chen[1], Jeffrey Barahona[1], Iype Eldho[1], Yanbing Yu[2], Riyaz Muhammad[2], Bill Kutsche[2], Michelle L. Hernandez[3], Delesha Carpenter[4], Alper Bozkurt[1],*Senior Member, IEEE*, and Edgar Lobaton[1],*Senior Member, IEEE*

***Abstract*—Cough detection is a crucial tool for long-term monitoring of respiratory illnesses. While clinical methods are accurate, they are not available in a home-based setting. In contrast, wearable devices offer a more accessible alternative, but face challenges in ensuring user speech privacy and detecting coughs accurately in real-world settings due to potential poor audio quality and background noise. This study addresses these challenges by developing a small-size multimodal cough and speech detection system, enhanced with an Out-of-Distribution (OOD) detection algorithm. Through our analyses, we demonstrate that combining transfer learning, a multimodal approach, and OOD detection techniques significantly improves system performance. Without OOD inputs, the system shows high accuracies of 92.59% in the in-subject setting and 90.79% in the cross-subject setting. With OOD inputs, it still maintains overall accuracies of 91.97% and 90.31% in these respective settings by incorporating OOD detection, despite the number of OOD inputs being twice that of In-Distribution (ID) inputs. This research are promising towards a more efficient, user-friendly cough and speech detection method suitable for wearable devices.**

***Index Terms*—Multimodal Cough Detection, Out of Distribution Detection, Audio Classification, Signal Processing**

I. Introduction

Lung diseases including asthma and chronic obstructive pulmonary disease (COPD) have significant global morbidity and mortality burden [1]. Long-term monitoring of these conditions is guided by assessments of coughing, a key symptom that is tracked for both the diagnosis and management of these chronic conditions [2]. However, quantifying cough frequency is often inaccurate due to inherent challenges with reliance on patient recall, with a tendency to underestimate cough [3] which can adversely impact clinical care. The earlier work in the [4] noted that the accuracy of diagnosing asthma and COPD by primary care providers can be alarmingly low, with correct diagnosis rates estimated to range between 25% and 50%. To aid with a more accurate assessment of chronic cough, long-term tracking of the type and frequency of cough is beneficial. Presently, long-term clinical monitoring in a home-based setting is expensive, not only due to the cost of specialized devices but also due to the manual labor involved in their operation and data analysis [5]. To address this diagnostic challenge and enable effective long-term monitoring, in-home wearable devices have been developed [6], [7]. These devices are embedded with machine learning models to record and analyze various biosignals, such as cough sounds. Typically, these models are trained to classify specific sounds using only audio input, assuming the data is clean. However, the reliability of these systems heavily depends on the quality of the data, a consideration that requires careful attention in system design.

Designing an effective cough detection system involves three key considerations: data collection, model architecture, and robustness. Despite the availability of various public datasets [8]–[10], there is a gap in data that accurately reflects real-world cough scenarios. Therefore, using raw data for development and testing is crucial. Architecturally, most current models rely on audio-only inputs, with limited exploration of multimodal data, like the one in [11], [12], showing obvious improvement. Inertial measurement units (IMU) have proven useful in numerous applications such as body rocking [13].

Robustness comes from accurately identifying "Out-Of-Distribution" (OOD) sounds, which are unfamiliar noises not covered during the model's training. These are contrasted with "In-Distribution" (ID) samples, which the model recognizes from its training. Models trained on specific classes often misclassify OOD sounds in practical use with high confidence, which can lead to unreliable results. Addressing OOD data handling is thus essential in our design [14]–[16]. Our previous research [17] shows that incorporating OOD detection algorithms can significantly enhance model performance, even at low sampling rates, ensuring user privacy.

To address these issues, we developed a chest worn sensor to collect microphone based audio and chest IMU based motion data to develop a multimodal model that incorporates OOD detection. Inspired by the availability of new microcontrollers with ultra-low-power neural accelerator capabilities targeting audio and sensor applications (such as the Type2DA Edge AI Module developed by Murata [18]), we envision future wearable devices capable of processing tasks like multimodal sensing (e.g., audio and inertial sensing), sounds classification, and OOD detection.

In this work, we collected a multimodal dataset of individuals performing various activities with non-verbal (e.g.,

*This work was supported by National Science Foundation (NSF) under awards IIS-1915599, IIS-1915169, IIS-2037328, and ECCS-2124002 (ERC for ASSIST).

[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA.; Corresponding author: edgar.lobaton@ncsu.edu

[2]Murata Electronics North America, Inc. 2200 Lake Park Drive, Smyrna, GA 30080-7604, USA.

[3]Department of Pediatrics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA.

[4]Division of Pharmaceutical Outcomes and Policy, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA.

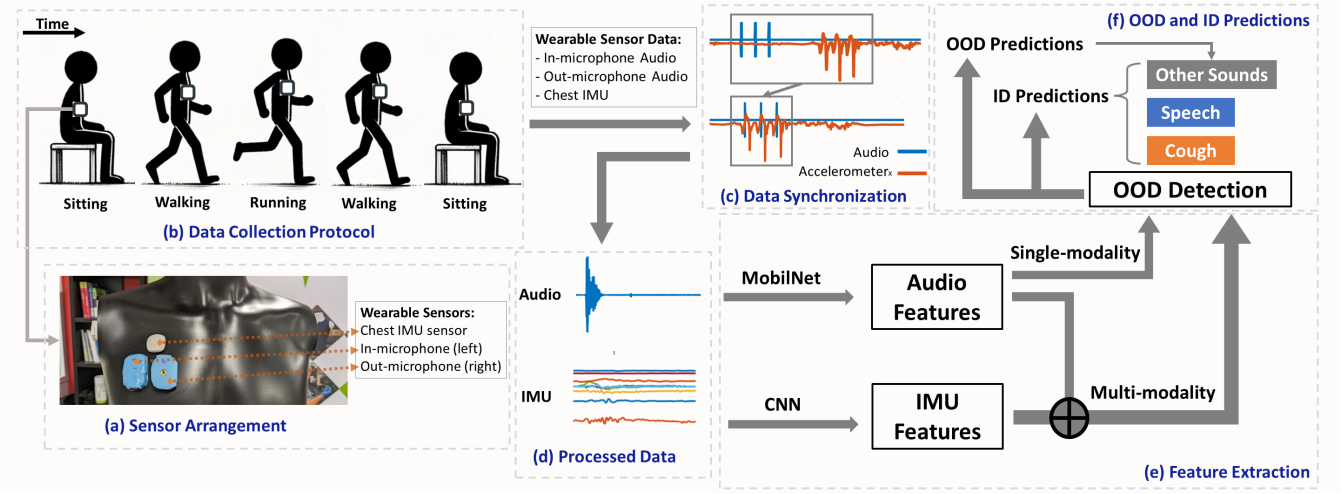

Fig. 1: Overview of Data Collection and Processing. Illustrations of (a) the sensor arrangement and (b) the data collection protocol. The data is then (c) synchronized and (d) processed using the different AI pipelines. The processing includes (e) feature extraction followed by (f) OOD and ID predictions.

cough) and verbal vocalizations (i.e., while speaking). We compared single-modal (audio-only) and multimodal (with inertial) models equipped with a simple but effective OOD detection algorithm. The overview of the data collection and processing pipelines can be found in Fig. 1. Finally, a detailed comparative evaluation of these algorithms was conducted in both in-subject and cross-subject settings. Our findings reveal that: 1) Detecting coughs and speech is more challenging when the models are tested on participants whose sounds are not part of the training set, highlighting the need for models with better generalization; 2) Transfer learning significantly boosts prediction accuracy; 3) Multimodal models show marked improvements in both in-subject and cross-subject settings, especially in cough detection; and 4) OOD detection effectively improves cough and speech sound detection when exposed to OOD inputs.

The rest of this paper is organized as follows: Section II details the process of data collection and pre-processing, and provides a summary of the dataset. Section III introduces the problems addressed in this study, describes the models employed, and outlines our experimental setups. Section IV provides detailed analyses and comparisons, covering single-modal, multimodal, and OOD-involved models, and offers an extensive discussion of these findings. Finally, Section V summarizes the entire work and explores potential future research scope. Data and the related code used in this paper are available [19].

II. Dataset

In the data collection process shown in Fig. 1(b), each participant performed a series of vocalizations under various activity levels. A total of 12 participants were involved in this study as approved by NC State University IRB Protocol 25003. The participants were healthy individuals between 20 and 30 years old. Participants sat ($\sim$ 2 min), walked ($\sim$ 2 min), ran ($\sim$ 2 min), walked ($\sim$ 2 min), and sat ($\sim$ 2 min) with 30-second resting intervals in each activation transition. This activity cycle was repeated three times, once for each one of the vocalizations specified in Table I

Audios were recorded by two chest-mounted microphones, one facing away from the participant (out-microphone) and one facing toward the participant (in-microphone) (Fig. 1 1(a)). We custom designed an enclosure and used microphones taken from commercially available Bluetooth earbuds (Tozo model T10 [20] with the speaker circuit disconnected. The participant's movement was recorded with Mbientlab's MetaMotionS r1 [21] sensor mounted on chest capturing 9-axis IMU data. The sensors were arranged on the participant as depicted in Fig. 1(a).

At the beginning of each recording, participants clapped three times and this procedure is used for data synchronization across different modalities. As depicted in Fig. 1(c), these three claps are distinctly observable in both the audio and IMU signals, producing accurate synchronization.

The data was labeled using the open-source tool Audino [22]. For the combination sounds, such as talking while coughing, we annotated both classes to the data. To make the data easy to load, we stored the annotation in an indexed format at 1 kHz, where each point is a 1×9 array with the corresponding idx labeled as 1 (i.e., using a one-hot encoding due to the potential overlap between labels).

The duration of audio recordings for each category is detailed in Table II. We categorize sounds into several classes: participant-generated sounds such as "Cough", "Speech",

TABLE I: Participant Vocalization Protocol

Abbreviation	Description
Silent	Participants remained silent except when instructed to cough every 20 seconds.
Talking	Participants read from a script or answered questions, with periodic cough instructions every 20 seconds.
Nonverbal	Participants were silent except for performing non-verbal vocalizations (coughing, laughing, sneezing, groaning, throat clearing) at 20-second intervals.

TABLE II: DATASET DESCRIPTION

Label (idx)	Cough (0)	Speech (1)	Sneeze (2)	Deep Breath (3)	Groan (4)	Laugh (5)	Speech-far (6)	Other Sounds (7)
Length (mins)	26.925	81.349	1.814	6.599	4.015	4.106	8.746	339.949
Percentage (%)	5.69	17.189	0.389	1.399	0.85	0.87	1.85	71.799

"Sneeze", "Deep Breath", "Groan", and "Laugh"; "Speech (far)", which represent speech from individuals around the subject; and "Other Sounds", indicating unlabeled environmental noises; including periods of silence. As shown in the table, this is an imbalanced dataset with the majority of the sounds falling into the "Other Sounds" category, followed by "Speech" and "Cough". For this research, we utilized all labeled sounds to train and evaluate our model on a classification task, and employed "Other Sounds" sounds as OOD inputs for testing.

Finally, Table III presents an overview of the various modalities employed in our study along with their corresponding frequencies.

III. METHODOLOGY

A. Problem Statement

We aimed to address two problems to improve model performance: enhancing prediction accuracy, and recognizing OOD inputs for the system reliability in real-world settings.

Problem Statement 1: The task was defined as a 3-class classification problem for ID inputs. The classes were "Cough", "Speech", and other vocalization sounds including "Sneeze", "Deep Breath", "Groan", "Laugh", and "Speech-far".

Problem Statement 2: The task was defined as OOD detection. In this context, all the classes defined in Problem Statement 1 were categorized as ID classes, and "Other Sounds", which were not part of the training dataset, were designated as the OOD class. The detection of OOD instances happens during the inference. To evaluate the model robustness on the main detection task (cough and speech) while OOD data was involved, we evaluated the overall performance by categorizing OOD data and other vocalization sounds in Statement 1 as the same class.

B. Metrics

For problem 1, the classic classification metrics were used for evaluation including accuracy, mAP, cough f1, and speech f1 [23]. For problem 2, besides classification metrics, OOD detection evaluation metrics were also involved including AUROC, and Detection Error [24].

C. Model Specification

In this study, we focused on three algorithms: Efficient CNN MobilNet [25]–[27], a multimodal model that combines MobilNets with IMU Net, and the Virtual-logit Matching (ViM) [28] algorithm for OOD detection. We first compared MobilNet and transferred MobilNet. Then, we examined how the multimodal model performs in comparison. Finally, we enhanced both MobilNets and the multimodal model with the ViM and compared their performances.

TABLE III: DATASET DESCRIPTION

Modality	Audio	Accelerometer	Gyroscope	Magnetometer
Frequency	16k Hz	$\approx$ 100Hz	$\approx$ 100Hz	$\approx$ 25Hz
Timestamps	✗	✓	✓	✓

MobilNet is a lightweight, efficient model designed for mobile and edge devices. It uses depth-wise separable convolutions to reduce computational load while maintaining effective performance. MobileNets [29], [30] have shown favorable performance in the audio domain. An efficient MobileNet was proposed in the literature [30] that was initially pre-trained on ImageNet and further trained on Audioset with the Knowledge Distillation technique. Using Knowledge Distillation enabled MobileNet to learn from large-scale high-performance Transformers as a student model and achieved competitive audio tagging performance.

Our multimodal model integrates an IMU network with MobileNet. The IMU network is based on a Convolutional Neural Network (CNN) with 5 convolutional layers and one linear layer. This IMU network outputs 128 features that are concatenated with the 960 features from MobileNet. The combined 1088 features are further reduced to 960 features to keep it consistent across single-modality and multimodality OOD performance comparison purposes via a linear layer. Finally, the model outputs class-dependent logits by a linear layer which was used in both ID classification and OOD detection.

ViM [28] is a state-of-the-art OOD method that combines the class-agnostic score from feature space and ID class-dependent logits. This class-agnostic score, serving as an additional logit for the virtual OOD class, is derived from the residual of the feature against the principal space. It is then scaled and matched with the original logits. A key advantage of this method is its application solely during inference but achieving promising results.

D. Experiment Setup

In our experiments, we used a 1.5 s sliding window with 0.5 s hop size [17], to extract data samples from each recording. Each data sample is labeled by the majority class in the sample. To fit into the pre-trained EfficientNet model [30], the audio was upsampled to 32 kHz. Then, the dataset was divided into two sets for training and evaluation by two different configurations, In-subject and Cross-Subject.

In-subject: Each audio was truncated such that 30% of its entire length formed the evaluation set, with the remaining 70% as the training set. We conducted 6 experimental runs for this configuration. We selected the test data using a strategy similar to K-fold cross-validation, in which the test set starts

TABLE IV: AVERAGE RESULTS OF SINGLE-MODALITY 3-CLASS CLASSIFICATION ACROSS 6 EXPERIMENTS

Model	Type	Acc	mAP	cough_f1	speech_f1
mn	In-sub	0.8518 ±0.0291	0.8627 ±0.0151	0.7564 ±0.0307	0.9244 ±0.0247
mn_as	In-sub	**0.9210** ±0.0237	**0.9363** ±0.0126	**0.8568** ±0.0263	**0.9655** ±0.0154
mn	Cross-sub	0.8135 ±0.0242	0.7752 ±0.0561	0.7011 ±0.0401	0.8986 ±0.0173
mn_as	Cross-sub	**0.8937** ±0.0216	**0.8767** ±0.0167	**0.7819** ±0.0438	**0.9573** ±0.0109

after 10%, 20%, 30%, 50%, 60%, or 70% of the length audio recording. For example, if we select a start of 10% that means that the test set consists of the audio between 10% and 40% of the recording. This is done to ensure that we have different scenarios that include different physical activities in our protocol for training and testing.

Cross-subject: We randomly selected 3 subjects as the leave-out test sets. The data from the remaining 9 subjects were utilized as training sets. We also conducted 6 experimental runs in this configuration for fair comparison purposes.

For feature extraction, we analyzed the transfer learning technique by comparing MobileNet and the MobileNet pre-trained on ImageNet [31] and AudioSet [8]. The baseline MobileNet model is denoted as **mn** and the AudioSet pre-trained MobilNet is denoted as **mn_as**.

All experimental models were trained using a batch size of 32 with 30 epochs, incorporating early stopping to prevent overfitting. In the case of the single-modal experiments, a learning rate of 5×10^{-5} was employed. For the multimodal experiments, a reduced learning rate of 4×10^{-5} was utilized, considering a greater number of parameters. For the OOD detection experiments, we selected M = 128 to split features for class-dependent logits and class-agnostic logit.

IV. RESULTS AND DISCUSSION

A. Single-Modality In-Distribution Performance

We first evaluated the single-modality effectiveness of a 3-class in-distribution classification task, conducted under both in-subject (in-sub) and cross-subject (cross-sub) settings. Table IV shows the average outcomes from six experiments as introduced in Section III-D. The AudioSet pre-trained MobileNet outperformed the standard MobileNet across all evaluated metrics with lower standard deviations indicating the effectiveness and stability of transfer learning.

In the cross-subject setting, we observed that the models were less stable and less accurate, especially in detecting cough sounds. This difficulty was mainly due to the diverse cough sounds produced by different individuals, making it harder to differentiate coughs from other non-verbal vocalizations. For speech detection, the performance drop was minor. This was because, even though speech varies among individuals, its consistent energy patterns make it easier to distinguish from other sounds.

Compared to the binary classification of cough and speech, as presented in Table V, the inclusion of additional vocalizations significantly impacted the detection of cough and speech (lower f1 scores). This was particularly evident in cough detection because a cough can be categorized as a non-verbal vocalization similar to other sounds such as groaning which increases the difficulty of the cough detection task. Similarly, the performance of speech detection was hindered by the presence of environmental speech-like sounds, despite differences in volume and fluency.

B. Multimodal In-Distribution Performance

The results of the experiments conducted on multimodal models are presented in Table VI. These results highlighted the importance of utilizing well-pretrained models for achieving both effective and stable outcomes.

Comparing Table IV and VI, it is evident that multimodal approach slightly enhanced performance when a powerful pre-trained model was employed. This improvement was particularly in the cross-subject setting, where the F1 score for cough detection increased from 0.7819 to 0.7975, and the F1 score for speech detection raised from 0.9573 to 0.9654. The diminished performance observed in single-modality and multimodal under the cross-subject setting highlighted challenges inherent in cough and speech detection when the sounds for model training and testing were from different subjects. Comparing standard deviations in different setups indicated the stability of the results was primarily influenced by the models themselves because pre-trained base models had lower standard deviations than that of baseline models.

C. OOD performance

To evaluate our model's robustness under real-world scenarios, we introduced OOD inputs at inference time and categorized them as other vocalization sounds while evaluating the accuracy. The F1 scores of cough and speech showed the detection performance in the specific class. To evaluate the model's ability on OOD task, we computed AUROC and detection error by categorizing OOD inputs as the negative class and the 3 ID classes as the positive class.

Tables VII and VIII provided a comparative analysis of models with and without OOD detection under in-subject and cross-subject settings, respectively. The comparison revealed that the integration of an OOD detection algorithm substantially improved overall accuracy. In terms of the specific class, the OOD inputs caused a big drop in cough detection resulting in a much lower F1 score of cough. This was possible because of the similarity of the cough distribution

TABLE V: AVERAGE RESULTS OF SINGLE-MODALITY COUGH AND SPEECH BINARY CLASSIFICATION ACROSS 6 EXPERIMENTS

Model	Type	Accuracy	AP	cough_f1	speech_f1
mn_as	In-sub	0.9732 ±0.0101	0.9896 ±0.0061	0.9433 ±0.0176	0.9824 ±0.0070
mn_as	Cross-sub	0.9683 ±0.0029	0.9817 ±0.0105	0.9149 ±0.0274	0.9799 ±0.0031

TABLE VI: AVERAGE RESULTS OF MULTIMODAL 3-CLASS CLASSIFICATION ACROSS 6 EXPERIMENTS

Model	Type	Acc	mAP	cough_f1	speech_f1
mn10	In-sub	0.8396 ±0.0303	0.7701 ±0.0641	0.7790 ±0.02487	0.9181 ±0.0243
mn10_as	In-sub	**0.9259** ±0.0238	**0.9431** ±0.0132	**0.8587** ±0.0294	**0.9693** ±0.0163
mn10	Cross-sub	0.8191 ±0.0343	0.7406 ±0.0469	0.6674 ±0.0330	0.9037 ±0.0263
mn10_as	Cross-sub	**0.9079** ±0.0199	**0.9016** ±0.0088	**0.7975** ±0.0332	**0.9654** ±0.0101

and OOD data distribution which were mostly contributed by non-verbal sounds. The ViM helped to relieve this effect and this enhancement was more evident in cough recognition than in speech recognition. Furthermore, the performance improvement was more obvious in the in-subject setting than in the cross-subject setting. This further indicated the challenge in model generalization which was also shown in previous tables. When comparing single-modal and multimodal models, it was observed that the multimodal models delivered a further slight improvement in overall accuracy while single-modal models can better recognize OOD inputs and produce better cough recognition.

D. Extensions

Upon reviewing Tables IV and V, it is observed that the F1 score for cough detection decreased by approximately 0.09 and 0.13 in the in-subject and cross-subject settings, respectively, after introducing a third class, vocalizations. However, the F1 score for speech detection dropped only by 0.017 and 0.023 in these settings. This discrepancy suggested the need for a deeper exploration into error sources, potentially through visualizing the confusion matrix for all 8 classes, to inform model improvement strategies.

Comparing single-modal models and multimodal models resulted in Table IV and VI, the enhancement in performance due to the integration of multiple modalities was evident, highlighting the significance of using additional related information. This indicated that incorporating more modalities (e.g. heart rate, symptoms) had the potential to further improve performance.

From Table VII and VIII, we observed that while multimodal models with OOD detection achieved the highest overall accuracy, theses were not as effective in OOD detection task as single-modal models. This phenomenon might be attributed

TABLE VII: COMPARISON OF WITH AND WITHOUT OOD DETECTION ON AVERAGE RESULTS FROM 6 EXPERIMENTS FOR IN-SUBJECT ANALYSIS (DE* IS DETECTION ERROR)

Modality	OOD	Acc	cough_f1	speech_f1	AUROC	DE*
Single	✗	0.8320	0.5025	0.8233		
Single	✓	0.9121	**0.5809**	0.8802	**0.7598**	0.2691
Multi	✗	0.8430	0.4638	0.8746		
Multi	✓	**0.9197**	0.5546	**0.9017**	0.7081	**0.2624**

TABLE VIII: COMPARISON OF WITH AND WITHOUT OOD DETECTION ON AVERAGE RESULTS FROM 6 EXPERIMENTS FOR CROSS-SUBJECT ANALYSIS (DE* IS DETECTION ERROR)

Modality	OOD	Acc	cough_f1	speech_f1	AUROC	DE*
Single	✗	0.8268	0.4562	0.8064		
Single	✓	0.8885	**0.4955**	0.8657	**0.8297**	**0.2662**
Multi	✗	0.8673	0.4342	0.9044		
Multi	✓	**0.9031**	0.4715	**0.9124**	0.7526	0.2829

to the uncertainty introduced by the IMU signals. With 9 IMU signals in total, not all contribute to sound recognition, as shown in Figure 1(d). However, the CNN model lacked the capacity to filter out irrelevant features effectively. Therefore, there was potential for further improvement in multimodal models for OOD detection, possibly through a better IMU model and appropriate OOD detection techniques.

V. CONCLUSION AND FUTURE WORK

In conclusion, our study developed a multimodal model for cough and speech detection by integrating chest based audio and IMU signals. The multimodal models demonstrated promising results, outperforming single-modal models in recognizing coughs, speech, and other vocalizations. The integration of OOD detection further enhanced the robustness of the model in identifying cough and speech sounds in the presence of OOD inputs. A detailed comparative evaluation of these algorithms conducted in both in-subject and cross-subject settings reveal that: 1) Detecting coughs and speech is more challenging when the models are tested on participants whose sounds are not part of the training set, highlighting the need for models with better generalization; 2) Transfer learning significantly boosts prediction accuracy; 3) Multimodal models show marked improvements in both in-subject and cross-subject settings, especially in cough detection; and 4) OOD detection effectively improves cough and speech sound detection when exposed to OOD inputs

In our future work, we aim to expand our data collection to include a wider range of other signals and participants' symptoms, which can be new modalities for model improvement. For preprocessing, we will add filtering and source separation techniques to remove noise. Additionally, we plan to enhance audio modalities by pre-training models on pre-existing cough datasets and explore more suitable multimodal architectures that are better fit for OOD detection. We will also investigate alternative OOD detection and uncertainty quantification methods to strengthen the model's robustness. Furthermore, considering the protection of participant privacy, we intend to evaluate the model's performance at lower sampling rates to find an optimal balance between data security and model effectiveness. Finally, we will deploy our system to the Murata Type2DA Edge AI Module and test real-time performance.

REFERENCES

[1] G. 2. C. R. D. Collaborators *et al.*, "Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: A systematic analysis for the global burden of disease

study 2015," *The Lancet. Respiratory Medicine*, vol. 5, no. 9, p. 691, 2017.
[2] K. S. Alqudaihi, N. Aslam, *et al.*, "Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities," *Ieee Access*, vol. 9, pp. 102 327–102 344, 2021.
[3] C. Yiannikas and B. T. Shahani, "Response," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 51, no. 12, p. 1600, Dec. 1988.
[4] C. Van Schayck and N. Chavannes, "Detection of asthma and chronic obstructive pulmonary disease in primary care," *European respiratory journal*, vol. 21, no. 39 suppl, 16s–22s, 2003.
[5] B. Beauvais, C. S. Kruse, *et al.*, "An exploratory analysis of the association between hospital labor costs and the quality of care," *Risk Management and Healthcare Policy*, pp. 1075–1091, 2023.
[6] V. Misra, A. Bozkurt, *et al.*, "Flexible technologies for self-powered wearable health and environmental sensing," *Proceedings of the IEEE*, vol. 103, no. 4, pp. 665–681, 2015.
[7] Y. Chen, M. D. Wilkins, *et al.*, "Toward automated analysis of fetal phonocardiograms: Comparing heartbeat detection from fetal doppler and digital stethoscope signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 975–979.
[8] J. F. Gemmeke, D. P. Ellis, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780.
[9] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.
[10] L. Orlandic, T. Teijeiro, *et al.*, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.
[11] E. Nemati, S. Zhang, *et al.*, "Coughbuddy: Multi-modal cough event detection using earbuds platform," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, IEEE, 2021, pp. 1–4.
[12] S. R. Chetupalli, P. Krishnan, *et al.*, "Multi-modal point-of-care diagnostics for covid-19 based on acoustics and symptoms," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 199–210, 2023.
[13] R. L. da Silva, B. Zhong, *et al.*, "Improving performance and quantifying uncertainty of body-rocking detection using bayesian neural networks," *Information*, vol. 13, no. 7, p. 338, 2022.
[14] J. Ren, P. J. Liu, *et al.*, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 707–14 718.
[15] I. J. Goodfellow, J. Shlens, *et al.*, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[16] A. Nguyen, J. Yosinski, *et al.*, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.
[17] Y. Chen, P. Attri, *et al.*, "Robust cough detection with out-of-distribution detection," *IEEE Journal of Biomedical and Health Informatics*, 2023.
[18] Murata, *Murata's type2da (equipped with syntiant ai chip)*, https://www.murata.com/en-eu/products/connectivitymodule/edge-ai/overview/lineup/type2da.
[19] https://github.com/ARoS-NCSU/OOD-Multimodal-CoughDet.
[20] TOZO, *Tozo wireless earbuds*, https://www.tozostore.com/.
[21] MetaMotion, *Metamotion*, https://mbientlab.com/metamotions/.
[22] M. S. Grover, P. Bamdev, *et al.*, *Audino: A modern annotation tool for audio and speech*, 2020.
[23] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.
[24] J. Yang, P. Wang, *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 598–32 611, 2022.
[25] A. G. Howard, M. Zhu, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[26] M. Sandler, A. Howard, *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[27] A. Howard, M. Sandler, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
[28] H. Wang, Z. Li, *et al.*, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4921–4930.
[29] Y. Gong, S. Khurana, *et al.*, "Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification," *arXiv preprint arXiv:2203.06760*, 2022.
[30] F. Schmid, K. Koutini, *et al.*, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
[31] J. Deng, W. Dong, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.