Robust Multimodal Cough Detection with Optimized Out-of-Distribution Detection for Wearables

Yuhan Chen[1], *Student Member, IEEE*, Feiya Xiang[1], Michelle L. Hernandez[2], Delesha Carpenter[3], Alper Bozkurt[1], *Senior Member, IEEE*, and Edgar Lobaton[1], *Senior Member, IEEE* ,

***Abstract*— Longitudinal and continuous monitoring of cough is crucial for early and accurate diagnosis of respiratory diseases. While recent developments in wearables offer a promise for daily assessment at-home remote symptom monitoring with respect to more accurate and less frequent assessment in the clinics, important practical challenges exist such as maintaining user speech privacy and potential poor audio quality and background noise in uncontrolled real-world settings. This study addresses these challenges by developing and optimizing a compact multimodal cough detection system, enhanced with an Out-of-Distribution (OOD) detection algorithm. The cough sensing modalities include audio and Inertial Measurement Unit (IMU) signals. We optimized this multimodal cough detection system by training with an enhanced dataset and employing a weighted multi-loss approach for the ID classifier. For OOD detection, we improved the system by reconstructing the training data components. Our preliminary results indicate the robustness of the system across window sizes from 1 to 5 seconds and performs efficiently at low audio frequencies, which can protect user privacy due to illegibility or incomprehensibility at lower sampling rates. Although we found that the multimodal model is sensitive to OOD data, the final optimized robust multimodal cough detection system outperforms the single-modal model integrated with OOD detection. Specifically, the optimized system maintains 90.08% accuracy and a cough F1 score of 0.7548 at a 16 kHz audio frequency, and 87.3% accuracy and a cough F1 score of 0.7015 at 750 Hz, even with half of the data being OOD during inference. The misclassified components mainly originate from nonverbal sounds, including sneezes and groans. These issues could be further mitigated by acquiring more data on cough, speech, and other nonverbal vocalizations. In general, we observed that the Audio-IMU multimodal model incorporating OOD detection techniques significantly improved cough detection performance and could provide a tangible solution to real-world acoustic cough detection problems.**

***Index Terms*— Cough detection, audio classification, out-of-distribution, bio-signal processing, machine learning.**

I. INTRODUCTION

Respiratory diseases, such as asthma and chronic obstructive pulmonary disease (COPD), impose a significant global burden in terms of morbidity and mortality [1]. Cough is a primary symptom of these chronic conditions, whose frequency is used for both diagnosis and management [2]. However, accurate quantification of cough frequency is challenging because it often relies on patient recall, which tends to underestimate the actual frequency of coughing [3]. This underestimation can negatively impact clinical care and lead to undertreatment of these chronic conditions. To improve the assessment of chronic cough, continuous tracking of the type and frequency of cough is essential. In-home wearable devices can facilitate long-term remote symptom monitoring by incorporating machine learning models to record and analyze biosignals [4], [5], such as cough sounds. Typically, these models classify specific sounds using only audio input, assuming the data are clean. The reliability of these systems is heavily dependent on the quality of the collected audio data, which may introduce significant uncertainty during the activities of the daily life in real world environments. Therefore, incorporating a robust architecture to guard against this uncertainty and implementing multimodality to enhance detection are important.

The integration of data acquired from multiple sensing modalities offers several advantages for improving system reliability and performance. Using data from diverse sources, such as audio, visual, and textual information, machine learning (ML) models can achieve a more comprehensive understanding of the problem at hand. In clinical scenarios, biosignals, demographic information, and clinical knowledge are commonly used together as additional source of information for smart health systems [6]. In our prior work [7], an audio-Inertial Measurement Unit (IMU) multimodal model has been shown to improve the accuracy and robustness of cough detection systems during physical activities; thereby, reducing error rates and enhancing overall system resilience.

Another way to improve model robustness is to incorporate the Out-of-Distribution (OOD) detection, which is a technique employed to identify samples that deviate from the distribution of the data used to construct the system. We call these samples

*This work was supported by National Science Foundation (NSF) under awards IIS-1915599, IIS-1915169, IIS-2037328, ITE-2344423 and EEC-1160483 (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]Department of Pediatrics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA.

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

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3616945

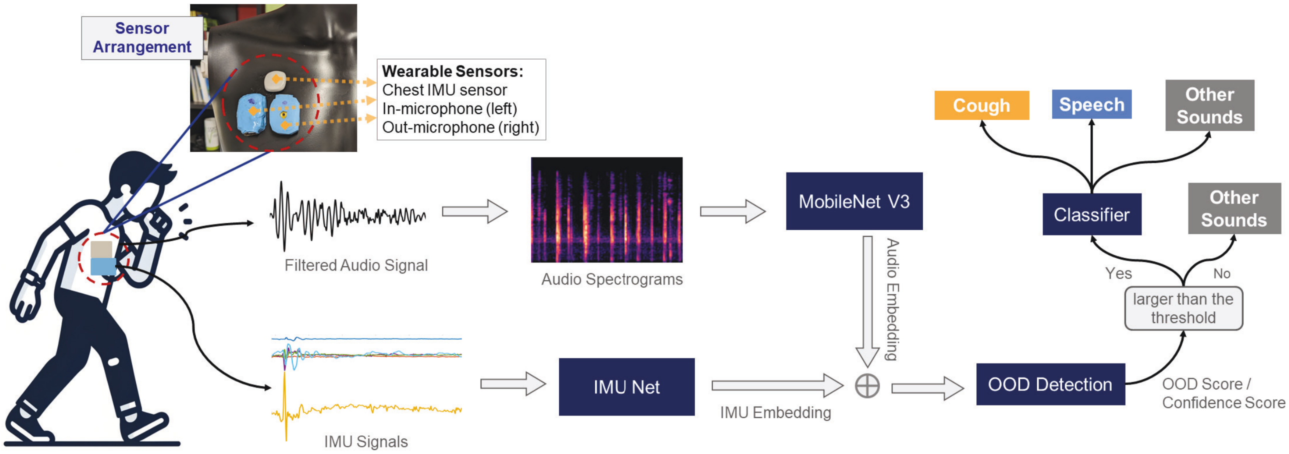


Fig. 1: Overview of the Robust Multimodal Cough Detection System Pipeline. 1) The data are collected by chest-mounted wearable sensors; 2) The filtered audio signal and IMU signals are passed through MobileNet and an IMU Net respectively to obtain audio and IMU embeddings; 3) These embeddings are then concatenated and passed through the OOD detection module to get an OOD score; 4) If the OOD score exceeds a specified threshold, the embedding features are sent to a classifier for the final prediction; otherwise, directly classified as other sounds.

as the “OOD data” and the data used for building the system are called “In-Distribution (ID) data”. Deep neural network classifiers can give high-confidence predictions to OOD inputs and lead to suboptimal results [8]–[10]. Therefore, OOD detection is essential for ensuring the robustness and reliability of machine learning and AI systems. In our preliminary work [7], [11], we have shown that incorporating OOD detection techniques improves cough detection performance by preventing unreliable predictions on unfamiliar data. Additionally, OOD detection facilitates better generalization by defining a model’s operational boundaries, enables active learning by prioritizing outliers for further training, and maintains trustworthiness by acknowledging and handling uncertainties appropriately. Thus, OOD detection is vital for the development of resilient, safe, and reliable AI systems.

To design an effective and robust multimodal cough detection system for real-world applications, we developed a comprehensive data collection protocol for this study [7]. Using data gathered through this protocol, we constructed an advanced multimodal cough and speech detection model using audio and IMU signals, which include accelerometer, gyroscope, and magnetometer data, thereby providing critical information about motion and orientation. At the end, the multimodal models demonstrated promising results, outperforming single-modal models in recognizing coughs, speech, and other vocalizations. The incorporation of OOD detection further enhanced the model’s robustness in identifying cough and speech sounds amidst OOD inputs. However, the data used in the previous work were highly unbalanced, and the analysis was not as comprehensive, allowing an opportunity for further improvement for this current study presented here.

In this paper, we present further optimization of our multimodal cough and speech detection system by enhancing both the data component and model architecture. Experiments were designed to analyze the impact of these optimizations on both ID classification task and OOD detection performances. The main contributions of this paper include:

- Adding the IMU modality significantly improves cough detection performance, with these improvements being more pronounced at lower audio frequencies. The use of an enhanced balanced dataset and a weighted multi-loss function further aids multimodal modeling, resulting in an improved cough F1 score.
- The optimized multimodal model demonstrates stability against the single-modal model’s sensitivity to window sizes and audio frequencies. Even at lower frequencies, the multimodal model remains stable with window sizes ranging from 1 to 5 seconds.
- Incorporating OOD detection addresses the multimodal model’s sensitivity to unknown OOD input. The final system can maintain a cough F1 score above 0.7 at 750 Hz, even when half of the data are OOD. This was a significant improvement over the F1 score of 0.6 obtained by a baseline, and it highlights the challenge and opportunities associated with overcoming OOD detection in this domain.
- Comprehensive analysis reveals the misclassified components in our system, demonstrating that most misclassifications involve non-verbal vocalization sounds. This insight provides a direction for further improvement.

The structure of this paper is organized as follows: Section II reviews state of the practice for cough detection, OOD detection, and our preliminary studies. Section III details data collection, model architecture, evaluation metrics, and experimental setup. Section IV presents results of four experiments, including tests on audio frequency and window size sensitivities, OOD data sensitivity, and misclassification analysis, along with a thorough discussion of these results. Finally, Section V concludes the work and suggests directions for future research.

II. Related Work

Automatic cough detection has been researched extensively [12]–[14] and the keys to improve accuracy include feature selection and modeling. Recently, there is more research using machine learning and deep learning methodologies to identify the most effective signal features and the most accurate detection methods [15], [16]. These studies have explored various features, including Short-Time Fourier Transform (STFT), Mel-frequency cepstral coefficients (MFCC), and Mel-scale filter banks (MFB), alongside classifiers such as logistic regression, feedforward artificial neural networks, support vector machines, and random forests. To further optimize detection algorithms, Lee et al. [17] further proposed advanced cough detection systems by integrating data enhancement processes. Moreover, other studies have addressed hardware device issues, specifically cross-device discrepancies, by employing ensemble classifiers [18], [19].

Besides a cough detection algorithm, techniques for improving the robustness of the system, such as OOD detection, can be used to improve the stability of the system. Recent contributions have focused on detecting OOD data within the field of computer vision [20]–[31]. Some methods utilize the maximum value of the softmax function to distinguish OOD inputs without modifying the underlying pre-trained model architecture [20], [21], while others incorporate an additional output to indicate the confidence of the results for identifying OOD inputs [22]–[24]. Furthermore, generative models, such as Variational Autoencoders (VAE) and diffusion probabilistic models (DDPM), can be employed for OOD detection by performing analyses in the latent space [25], [31].

Another approach to enhancing system robustness is the incorporation of multiple modalities, which has proven effective in both computer vision and natural language processing. In the domain of cough detection, M. Paha et al. [32] demonstrated the utility of accelerometer signals for cough detection, though it was found to be slightly less accurate than audio signals. Lara Orlandic et al. [33] further advanced this field by creating an audio-IMU dataset, illustrating the improvements achieved through multimodal integration.

In our previous work [11], we demonstrated the effectiveness of a robust pipeline that integrated cough and speech detection algorithms with OOD detection, using Mel-spectrogram inputs with various sampling frequencies and window sizes. Furthermore, we showed an improvement in the problem of detecting cough sounds during different daily activities (standing, walking, running) by transitioning from a single modality to a multimodal approach [7]. In this study, we will focus on the optimization and analysis of the multimodal cough detection system that incorporates OOD detection.

III. Proposed Methods

We divide the system design problem into **ID classification problem** and **OOD detection problem**. ID classification problem is a three-class classification that aims to accurately classify “cough”, “speech”, and other vocalization sounds including sneezing, deep breathing, groaning, laughing and speaking sounds from individuals around the subject. OOD detection is employed to recognize unknown sounds in the testing phase, simulating real-world scenarios. Our results indicate that the system effectively detects cough sounds amidst a variety of unknown sounds.

The primary challenge of this work comes from the limited and unbalanced data. In this study, our objective is to improve our robust multimodal cough detection system by optimizing both the classification model and the OOD detection algorithm by improving the quality of the data and developing a data-adaptive system. Our optimization strategy includes enhancing the dataset with an available online dataset, improving the classification model by employing a multi weighted loss function that weights different class losses based on the difficulty of detecting different classes and joins loss from both signal modality and multimodality, and improving OOD detection by refining the illness eigenvalues derived from the feature space of the multimodal model.

A. Datasets

To deal with limited data, two datasets were combined together for the system development. To better illustrate the details of these two datasets, they are denoted as dataset A and dataset B. We refer to the dataset collected for this study as dataset A [7] and the online public multimodal dataset collected by Lara Orlandic et al. [33] as dataset B.

Dataset A was built with data from a total of 13 participants and this data collection process was approved by NC State University IRB Protocol 25003 on April 13, 2023. Data from 12 participants were used in this study since the data from the last participant did not include IMU data. The participants were healthy individuals between 20 and 30 years old. In the data collection process, each participant performed a series of vocalizations under various activity levels. The participants sat (∼ 2 min), walked (∼ 2 min), ran (∼ 2 min), walked (∼ 2 min) and sat (∼ 2 min) with 30-second resting intervals in each activation transition. This activity cycle was repeated three times and each time participants were required to perform different vocalizations including pure coughing, talking, talking while coughing, and other nonverbal vocalization sounds.

Audio was recorded by two chest-mounted microphones, one facing away from the participant (out-microphone) and one facing toward the participant (in-microphone). See the sensor arrangement in Fig. 1. These microphones were repurposed from commercially available Bluetooth earbuds (Tozo model T10 [34]), with the speaker circuit disconnected. The movement of the participants was tracked using the MetaMotionS r1 sensor, mounted on the chest, which captured 9-axis IMU data (accelerometer, gyroscope, and magnetometer). IMU signals were sampled at 100 Hz and audio signals were sampled at 16 kHz.

At the beginning of each recording, participants clapped three times and this procedure is used for data synchronization

TABLE I: Summary of Datasets

	ID			OOD
	cough (∼ 28.2%)	speech (∼ 12.9%)	other * (∼ 59.6%)	unlabeled sounds
Dataset A	∼27 min	∼81 min	∼25 min	∼340 min
Dataset B	∼ 119 min		∼350 min	

* The “other” class in Dataset A includes sneezing, deep breathing, groaning, laughing and speaking sounds from individuals around the subject.
* The “other” class in Dataset B includes laughing, throat clearing, and deep breathing.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3616945

across different modalities. These three claps are distinctly observable in both the audio and the IMU signals, producing accurate synchronization. Then a low-pass FIR filter with a 3 kHz cut-off frequency was applied to attenuate high-frequency noise. This cutoff frequency was chosen based on previous findings showing stable performance over 2 kHz [11], ensuring the retention of low-frequency information to maintain the model's performance for edge-AI development, which typically supports limited computational cost.

The duration of audio recordings for each category is detailed in Tab. I. We categorize sounds into several classes: participant-generated sounds such as "cough", "speech", "sneeze", "deep breath", "groan" and "laugh"; "speech (far)", which represent speech from individuals around the subject; and "unlabeled sounds", indicating unlabeled environmental noises; including periods of silence. All participant-generated sounds except for "cough" and "speech" are treated as "other" class in ID data, and all "unlabeled sounds" are treated as OOD data. For the classification task, this is an imbalanced dataset as shown in Tab. I with the majority of the sounds falling into the "speech" category, followed by "cough" and "other". To address this imbalance, we combined dataset A with an online available dataset B [33] to create an enhanced dataset.

Dataset B was acquired by Lara Orlandic et al. [33] and this dataset was selected due to its similar format to dataset A. Both datasets have identical audio and IMU sampling rates, 16 kHz and 100 Hz respectively, and were collected under comparable conditions. Dataset B comprises data from 20 healthy subjects aged 26.5 ± 6.5 years, with data from 15 subjects available online. This dataset includes nearly 4 hours of biosignal data, featuring 4,300 annotated cough events. It combines acoustic signals from dual microphones, one facing the body and the other outward, with kinematic data from tri-axial accelerometers and gyroscopes. These multimodal biosignals encompass a variety of non-cough-related sounds and physical movements to simulate realistic environmental conditions. Each recording lasts approximately 10 seconds, during which subjects were asked to produce either a cough sound or one of three sounds that could produce similar audio or chest motion artifacts as a cough: laughing, throat clearing, and deep breathing.

Based on our observation, magnetometer signals do not exhibit a specific relationship with different sounds, and dataset B does not have magnetometer signals; therefore, only accelerometer and gyroscope signals were used as the IMU modality.

B. Data Preparation

To construct a high-quality and well-performing deep learning model, the quality of the dataset is crucial. Tab. I shows that compared to dataset B, the total recording time of dataset A is approximately four times less. To ensure that the distribution of the enhanced dataset closely matches dataset A, different methods were employed to extract data samples. For dataset A, a sliding window with a 0.5-second hop size was used to extract data points, while in dataset B, each event was extracted only once, based on the middle point of each event. In both datasets, the label of each data sample is assigned as the label corresponding to the middle time point. Consequently, the number of data points extracted from dataset A is on the same level as that from dataset B. Moreover, this strategy maintains a consistent total number of samples across different window sizes. The final component details can be found in the Supplement material, Table I.

In this work, all labeled data from nine subjects in dataset A and all data from dataset B were utilized for training. The labeled data from the remaining three subjects in dataset A were used to test the model performance in the ID classification task, and the data from unlabeled sounds from the same three subjects in dataset A were used as OOD inputs for OOD detection task testing.

To further balance the training dataset, data from Dataset B were merged into the "cough" and "other" classes of dataset A after balancing dataset B. Specifically, N samples were randomly selected from the "other" class in Dataset B, where N corresponds to the number of "cough" events in Dataset B. To simulate real-world scenarios, no data enhancement or augmentation was performed on the validation dataset.

To evaluate the improvement brought by the enhanced dataset, we compared the model built using solely dataset A with the model built using the enhanced dataset. The results in Section IV-A highlight the importance of both the quantity and quality of the dataset.

C. Architecture

The multimodal model incorporates Efficient Convolutional Neural Network (CNN), MobileNet [35]–[37] alongside a CNN-based IMUNet. MobileNet, pretrained on ImageNet [38] and AudioSet [39], is utilized to extract features from the audio modality, serving as the backbone for the single modality [40]. The multimodal model is constructed by combining MobileNet for audio embeddings with a simple 4-layer CNN model for IMU embeddings. For the training strategy, we leverage the pretrained weights of MobileNet for audio feature extraction, while IMUNet and all other network components are trained from scratch. To enhance the model's robustness against out-of-distribution (OOD) inputs, a Virtual-logit Matching (ViM) OOD detector [30] is integrated. Further details on the multimodal model architecture are provided in the Supplement. (source code will be made public after the paper is accepted.)

D. Experiment Setup

We evaluated our system using a cross-subject setting, wherein the data used for training and testing originate from different subjects. This configuration presents a greater challenge compared to an in-subject setting, where the training and testing data are derived from the same subjects. This conclusion is drawn from our preliminary investigations. Specifically, we randomly selected three subjects from dataset A as the leave-out test sets, while the data from the nine remaining subjects in dataset A were utilized as training sets.Dataset A was specifically chosen for testing due to its alignment with our experimental objectives. To establish statistical reliability, we performed 6 independent experimental iterations for each model, with 4 following a standard systematic cross-validation procedure and 2 for more consistent and robust results. The two additional iterations utilized sets selected to enhance stable results as a high performance ($\sim 94.7\%$ accuracy) observed in the original 4-folds, which can lead to internal bias in the model. The results from each fold can be found in the source code provided in Section III.C and the supplementary material Section V. This evaluation framework enables a more consistent comparison of model generalization across different subject populations.

Besides, to evaluate the stability of the system, we conducted tests using varying sampling rates (f) and window sizes (τ). When modifying the sampling rates, only the sampling rate of the audio modality was altered, while the IMU maintained the same rate. This approach takes into consideration the higher risk of user information leakage associated with audio data, whereas IMU data is considered to be more secure. Consequently, we maintained the IMU sampling rate to optimize performance, while reducing the audio sampling rate to enhance user privacy.

For the training of each model, we employed a batch size of 32 and the Adam optimizer. The learning rates were set at 3×10^{-4} for the audio modality and 2×10^{-4} for the IMU modality. These learning rates were determined to be optimal through multiple hyperparameter search iterations. To ensure effective training, we also implemented a warm-up phase followed by learning rate decay. Each model was trained for 30 epochs, and the model that achieved the highest cough F-1 score was preserved, as our primary focus in this system is cough detection.

For OOD detection, we set the dimensionality D to 512 to extract the principal subspace and the residuals. Given that data components can influence the significance of the eigenvalues for each class, we also evaluated the impact of using only cough and speech data in the training set in Section IV-C. These classes were selected due to their distinct identifiability compared to other classes, and this distinction has proven beneficial for OOD detection.

E. Evaluation Metrics

For the ID classification problem, we employed classic classification metrics to evaluate the model's performance, including accuracy, mean Precision (mP), cough F1, and speech F1 [41]. Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances. mP provides an average measure of precision across all classes, ignoring the bias caused by the imbalance issue. The F1 scores for cough and speech are particularly crucial as they offer a balanced measure of precision and recall, reflecting the model's effectiveness in identifying these specific classes.

For the OOD detection problem, we utilized the Area Under the Receiver Operating Characteristic Curve (AUROC), FPR95 and Detection Error [42] by treating ID as positive class and OOD as negative class. AUROC is essential for evaluating the trade-off between true positive and false positive rates across various thresholds, providing a comprehensive measure of the model's discriminatory power. FPR95 assesses the false positive rate when the true positive rate is fixed at 95%, offering insights into the model's performance under specific conditions and lower is better. Detection Error quantifies the proportion of instances where the model incorrectly identifies OOD samples.

TABLE II: Comparison on ID 3-Class Classification at $f = 16$ kHz and $\tau = 1.5$ s

Modality	Loss	Acc	mP	cough_f1	speech_f1
single	single-loss	0.9110 ±0.0221	0.9215 ±0.0250	0.8606 ±0.0211	0.9644 ±0.0100
single	weighted single-loss	0.9077 ±0.0253	0.9233 ±0.0224	0.8667 ±0.0242	0.9609 ±0.0101
multi	multi-loss	0.9124 ±0.0203	**0.9315** ±0.0214	0.8665 ±0.0241	0.9644 ±0.0093
multi	weighted multi-loss	**0.9140** ±0.0205	0.9247 ±0.0228	**0.8668** ±0.0253	**0.9649** ±0.0085

* Rows 1 and 3 present baseline models configured according to the settings established in our prior work [7], with slight improvements achieved through hyperparameter tuning.

Furthermore, we assessed the overall performance using standard classification metrics by treating OOD inputs as the third (other) class. This approach simulates real-world scenarios where the primary concern is the accurate detection of cough and speech sounds under the presence of known noise and unknown OOD sounds.

IV. Experiments

We set up four experiments to comprehensively evaluate our system. Section IV-A tests the improvements achieved through weighted multi-loss functions and data enhancement. Section IV-B examines the system's sensitivity to different sampling rates (f) and window sizes (τ). Section IV-C evaluates the performance of OOD detection. Section IV-D conducts error analyses. Finally, Section IV-E discusses all the results and proposes directions for future work.

A. Classification Optimization with Weighted Multi-loss and Data Enhancement

The optimization for ID classification task include weighted multi-loss and data enhancement techniques. Tab. II presents a comparison between the standard cross-entropy loss and the weighted loss. The weighted loss assigns a penalty of 10 for incorrectly detecting a cough, whereas the penalty for other errors is set to 1.

From the single modality analysis, it is evident that the weighted loss marginally enhances cough detection, thereby improving the mean Precision (mP). However, due to the imbalanced evaluation data, there is a slight decline in overall accuracy. From the multimodality analysis, the weighted loss slightly improves the performance of detecting both cough and speech, resulting in increased accuracy while sacrificing the performance of other class detections, leading to a decrease in the mP.

Compared to single and multi modalities, the multi modality overall enhances cough detection. The application of weighted loss further aids in this improvement, though its impact is more obvious in the single modality.

We maintained the weighted loss functions and further optimized the model by implementing data enhancement. In Fig. 2, we present a comparative analysis of the single-modal model against the multimodal models, with and without data enhancement, across various audio sampling rates. The multimodal model trained with enhanced data is denoted as "Multimodal+."

We observe that overall, multimodal models outperform the single-modal model, particularly in cough detection. This improvement is more pronounced at lower f. Additionally, the multimodal model trained with enhanced data further enhances cough recognition performance.

B. Sensitivity to Sampling Rates and Window Sizes

This experiment is designed to evaluate the sensitivity of three models (single-modal, multimodal, and multimodal+) across varying f and τ. We tested f ranging from 500 Hz to 16 kHz and τ ranging from 0.5 to 5 s.Results for 500 Hz and 16 kHz are presented in Fig. 2, with the full set of results available in the supplementary materials.

From Fig. 2, performance declines as the f decreases for both single-modal and multimodal models. A significant drop is observed for the single-modal model at approximately 1 kHz and a slight increase in the drop rate is observed for multimodal models. Our preliminary analysis [11] indicates that 750 Hz is the highest f capable of preserving the content of speech accurately being recognized for user privacy protection. Additionally, lower f contributes to saving computational resources, making it more suitable for resource-constrained systems. At 750 Hz, the multimodal models maintain an accuracy of approximately 0.875 and a cough F1 score of around 0.825.

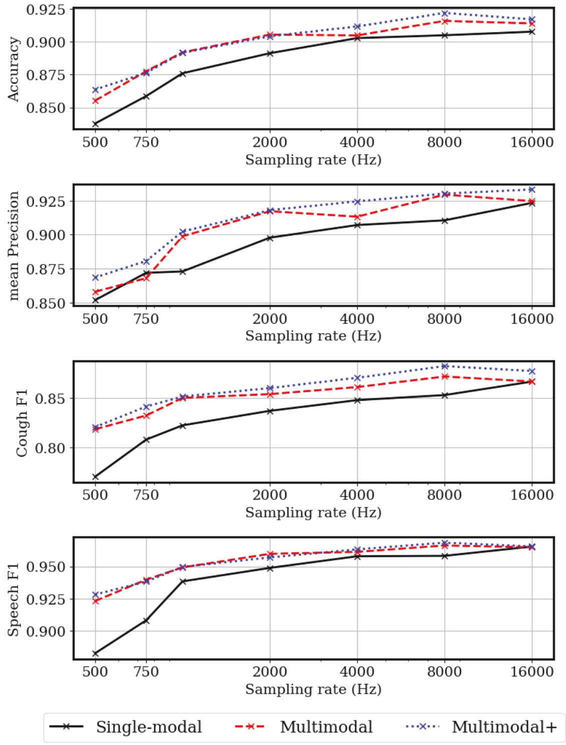


Fig. 2: Comparison of single-modal and multimodal models on classification task for ID samples using accuracy, mean precision, cough F1, and speech F1 metrics. Average results from six runs are presented. "Multimodal+" represents the multimodal model trained with data enhancement. ($\tau = 5$ s)

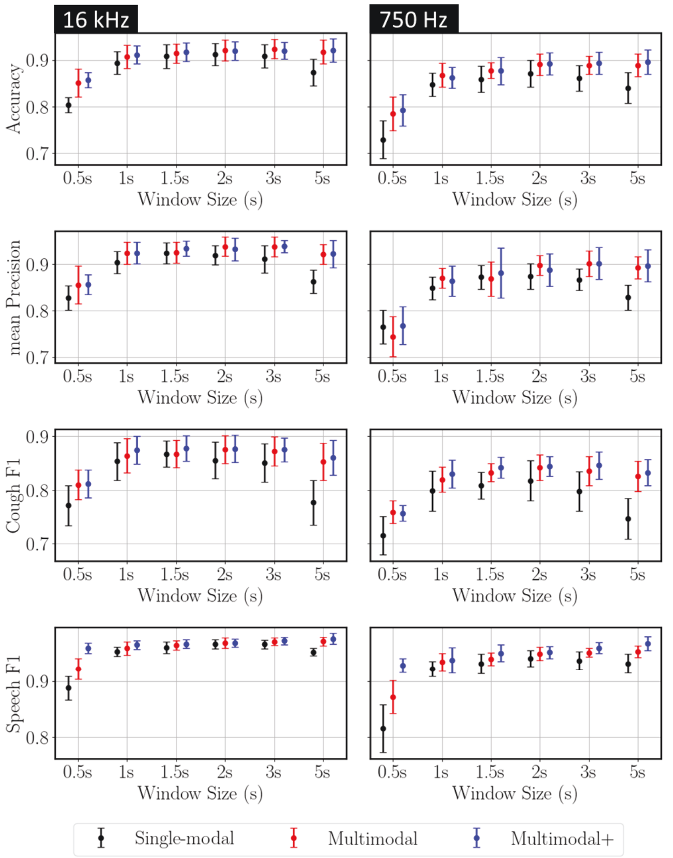


Fig. 3: Comparison of single-modal and multimodal models across varying sampling rates (f) and window sizes (τ): Means and corresponding standard deviations over six runs. "Multimodal+" denotes the multimodal model trained with data enhancement.

From Fig. 3, it is evident that the overall performance remains reliable within a τ range of 1-2 s across various f for all three models. Even at 750 Hz, multimodal models achieve an accuracy and mean precision of approximately 0.88. However, for τ between 3-5 s, there is a notable decline in the performance of single-modal model, particularly in accurately detecting coughs. When the τ is reduced to less than 1 s, performance significantly decreases, as this duration is insufficient to represent a single cough event. This inadequacy also results in a decrease in the F1 score for speech detection.

In comparison to single-modal model, multimodal models demonstrate greater stability across varying f and τ. This increased stability can be attributed to the compensatory effect of the information derived from the IMU modality, which addresses the deficiencies in the audio modality. Additionally, multimodal models that have been trained with enhanced data exhibit enhanced robustness across different τ. This robustness is particularly evident at higher f, where all information from the audio modality is preserved. This may be due to the discrepancy in feature distribution between dataset A and dataset B, which arises from the process of downsampling.

C. OOD Detection Analysis

To maintain consistency with our preliminary work, we have chosen τ as 1.5 s for the OOD detection task analysis. Since only dataset A is used as the evaluation set, we exclusively utilize the training data from dataset A for feature extraction in the OOD detection process. We also evaluated the OOD detection performance using the training data from dataset B, and the results were comparable to those obtained using only the training data from dataset A. Therefore, we present results based solely on the training data from dataset A. This approach not only ensures consistency but also reduces computational costs. Furthermore, this method provides a fair comparison with the single-modal model, as only dataset A training data is used in the single-modal OOD detection procedure.

Fig. 4 shows the comparison of single-modal and multimodal models across varying proportions of OOD data. The x-axis represents OOD data proportion ranging from 0% to 50%, where $n\%$ indicates the number of OOD samples introduced in the test set. This quantity is given by:

$$n\% = \frac{\text{The \# of OOD Samples}}{\text{The \# of Samples in the Test Data}} \cdot 100\%. \quad (1)$$

The incorporation of OOD detection does not impact the results when there is no OOD data involved at higher f. At 750 Hz, there is a decrease in accuracy by approximately 0.01 in the absence of OOD input. With the increase of OOD data involved, the models with OOD detection remain robust, and the multimodal model with OOD detection surpasses the one without the OOD detection at 10% OOD data presence. The involvement of OOD detection primarily increases the accuracy of cough detection due to the OOD detection threshold being selected based on optimal cough detection performance. We also observe that the speech F1 score improves at higher f but decreases at lower f.

Compared to the single-modal model, the multimodal model is more sensitive to OOD input at 16 kHz without the OOD detection. However, with the incorporation of the OOD detection algorithm, the multimodal model becomes stable and surpasses the single-modal model that also incorporates OOD detection. At 16 kHz, the multimodal model slightly outperforms the single-modal model, achieving 89.58% accuracy and a cough F1 score of 0.7548 with 50% of the data being OOD. At 750 Hz, the robust single-modal model shows lower performance in terms of speech F1 score, resulting in reduced accuracy. This will be discussed in detail in Section IV-D. Conversely, the multimodal model achieves 87.29% accuracy and a cough F1 score of 0.7009, sacrificing a bit of speech F1. Overall, the robust multimodal model outperforms the single-modal model in both scenarios, with and without the OOD detection.

Tab.III shows the comparison of OOD detection performance using OOD metrics with 50% OOD data. "Multimodal+*" represents the Multimodal+ model with only cough and speech sounds as ID training data. From the table, we observe that after removing the imprecise and unclear class "other," the OOD performance improves, especially at lower f. At higher f, higher AUROC and lower FPR95 indicate that using only cough and speech can enhance overall performance but requires a strict threshold. At lower f, all metrics improve. Specifically in the cough detection task, the optimized "Multimodal+*" further improves accuracy to 90.08% at 16 kHz, and to 87.30% at 750 Hz.

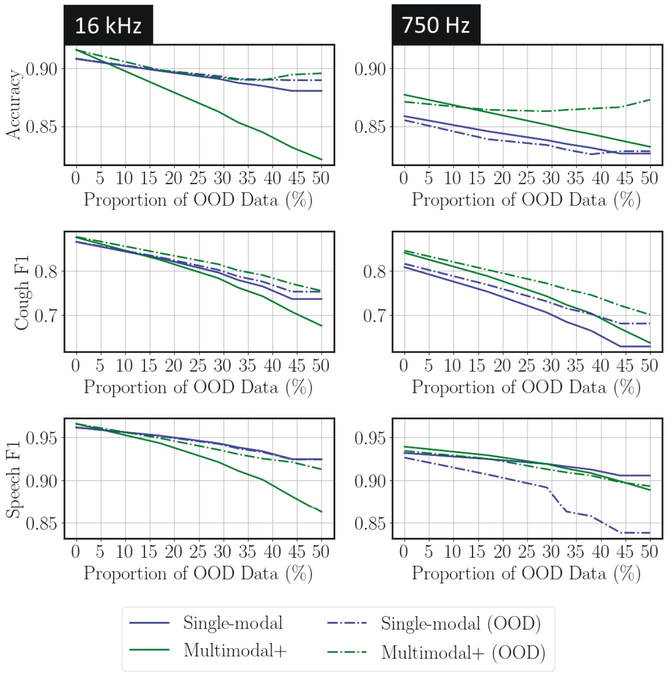


Fig. 4: Comparison of single-modal and multimodal models across varying proportions of OOD Data. Average results over six runs. The solid line represents performance without the OOD detection algorithm, while the dashed line represents performance with OOD detection algorithm implemented.

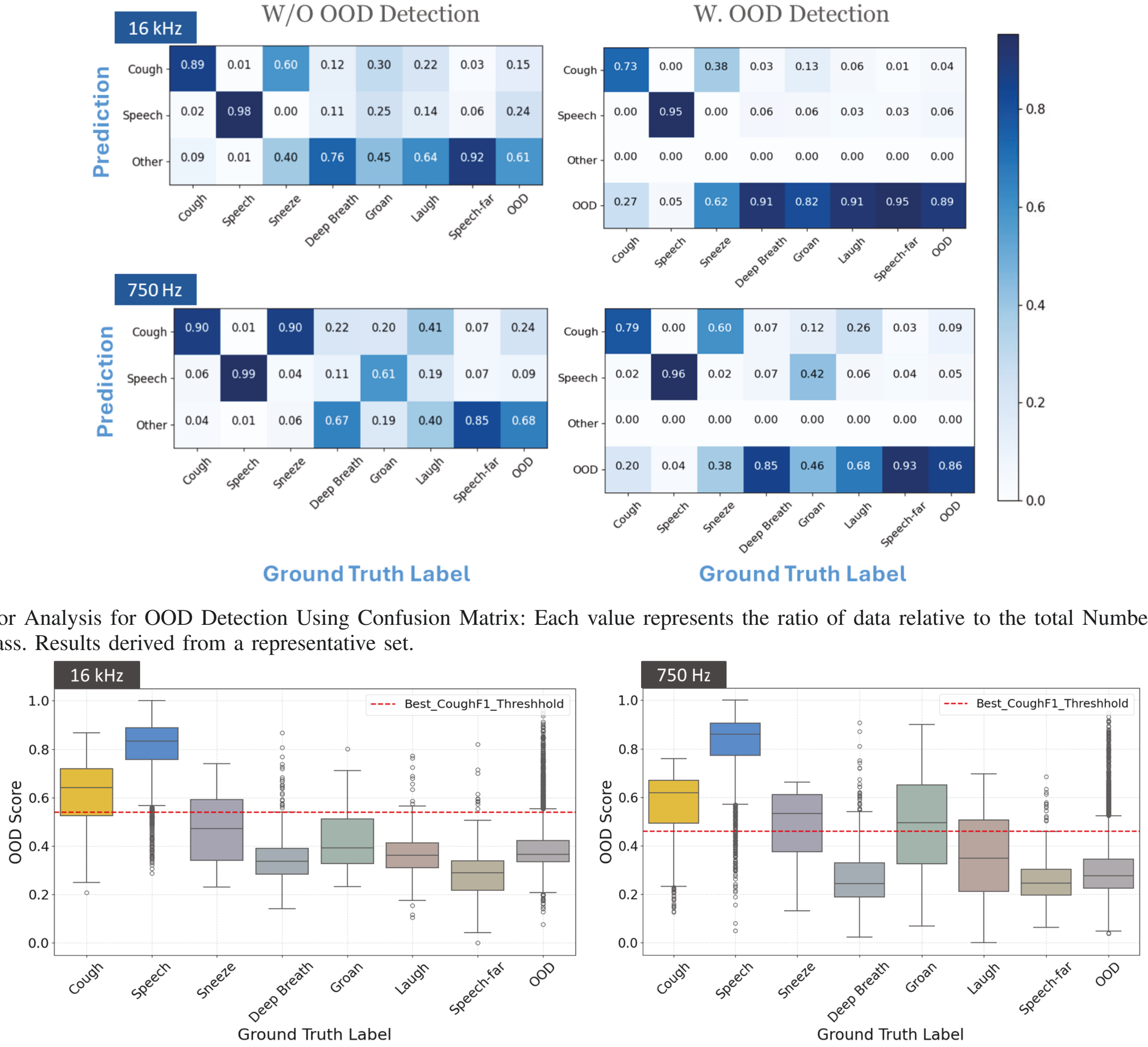
TABLE III: Comparison of OOD Performance with Different Training Data Components. * indicates that eigenvalues for ViM score computation are extracted using only cough and speech inputs.

Model	Frequency	AUROC↑	Detection Error↓	FPR95↓
Single-modal	16 kHz 750 Hz	0.7572 0.7295	0.5060 0.5261	0.8710 0.9093
Multimodal+	16 kHz 750 Hz	0.8410 0.7653	**0.4150** 0.4752	**0.7802** 0.9011
Multimodal+*	16 kHz 750 Hz	**0.8604** **0.8658**	0.4653 **0.4019**	0.8832 **0.7541**

D. Misclassified Component Analysis

In this section, we analyzed the errors in the robust multimodal cough detection system to seek ways to improve its performance. To better understand the misclassified samples, we visualized the prediction results using confusion matrices and box plots of a representative set, a single experimental run that yielded overall performance metrics closest to the average, as shown in Fig. 5a and Fig. 5b, respectively.

In Fig. 5a, we present the ratio of the number of data points to the corresponding class to highlight the sounds that are more prone to misclassification. We observe that at 16 kHz, "sneeze" is the most challenging class for the system to recognize, with 60% of these sounds being misclassified as "cough" in



(a) Error Analysis for OOD Detection Using Confusion Matrix: Each value represents the ratio of data relative to the total Number in the true class. Results derived from a representative set.

(b) Error Analysis Using Boxplots for OOD Scores Across All Classes: Results from a representative set

Fig. 5: Error Analysis Using Confusion Matrices (a) and Boxplots (b).

the system without the OOD detection. At 750 Hz, “sneeze,” “groan,” and “laugh” are frequently misclassified. Even with OOD detection, 60% of “sneeze” and 26% of “laugh” sounds are classified as “cough,” while 42% of “groan” sounds are classified as “speech.” Additionally, we observe that at 16 kHz, there is a loss of 19% of “cough” and 3% of “speech” sounds after implementing OOD detection. At 750 Hz, there is a loss of 11% of “cough” and 3% of “speech” sounds following the use of OOD detection.

Fig. 5b illustrates the OOD scores (ViM scores) for each specific class to further explain the misclassified components identified in Fig. 5a. The red dashed lines in Fig. 5b indicate the thresholds that produce the best cough F1 score, which are used to identify the critical class in our system, “cough.” At 16 kHz, there is a small overlap between the OOD scores of “cough” and “sneeze,” leading to the misclassification of sneeze as cough. At 750 Hz, the OOD score distribution for “cough” remains similar to that at 16 kHz, while the OOD scores for other sounds slightly increase and become more dispersed. These higher OOD scores result in misclassifications as either cough or speech sounds.

E. Discussion

In this study, we focus on developing a robust cough detection system for wearable devices that can efficiently detect cough sounds with minimal energy consumption. We optimized the multimodal model proposed in the preliminary work [7] and compared it with a single-modal model across various sampling rates and window sizes. Additionally, we employed OOD detection techniques to enhance robustness to unknown OOD inputs.

Considering the uncertainties present in real-world scenarios, we developed and evaluated our system using data collected from customized wearable devices (dataset A). Due to the significant imbalance in dataset A, we incorporated an online dataset B [33] to construct a higher quality dataset for system development. To ensure realistic results, only dataset A was used for testing. To further optimize the model, we employed a multi-loss approach, integrating information from both multimodal and pure audio modalities, which contain sound information that could better represent each class. We also applied weights to the multi-loss to emphasize the significant but challenging-to-predict class, “cough”. Moreover, we refined the OOD detection algorithm by reconstructing the components of the training set.

We set up the robust cough detection system design as ID classification task and OOD detection task. For ID classification, we switched from a binary classification of cough and speech detection to a three-class classification by adding an “other” class. This class includes nonverbal vocalization sounds and speech sounds from people around the subject. Involving more classes in the model development enhances the model’s robustness against non-cough and non-speech sounds. For the OOD detection task, we classified unknown sounds, such as environmental noises and silence, as OOD, while all sounds belonging to ID classes were classified as ID. The combination of three-class classification and OOD detection supports the model’s stability under varied inputs. We conducted four experiments to compare and evaluate our systems under variance circumstances.

In the first experiment, we observed an improvement in cough detection by adding the IMU modality and training with a weighted loss, as shown in Tab. II. We maintained the same IMU frequency and tested model performance under varying audio frequencies (f), as illustrated in Fig. 2. The plot indicates that the improvement due to the IMU modality is more pronounced at lower f. This is reasonable because we observed significant spikes in the accelerometer and gyroscope signals during most cough sounds, along with fluctuations for speech and other sounds. Therefore, the information from the IMU modality compensates for the missing information caused by reduced frequency. Fig. 2 also demonstrates the improvement resulting from incorporating more cough data. This provides evidence that the model can be further improved by acquiring additional data.

In the second experiment, we compared model performance over varying window sizes (τ) at f as 16 kHz, 750 Hz, and 500 Hz. 750 Hz is noteworthy as it effectively protects user privacy [11] is noteworthy as it effectively protects user privacy [11], while the 500 Hz rate is notable due to its lower consumption cost, which is advantageous for Edge AI development. In Fig. 3, we observe that the single-modal model is more sensitive to τ across different f, with the best results achieved using a τ of 1-2 s. In contrast, multimodal models demonstrate greater robustness with larger τ, suggesting their ability to learn important features even within extensive τ. For both models, there is a significant performance drop with τ smaller than 1 s, indicating that at least 1 s is necessary to capture the complete information of a cough. Additionally, we observe an improvement in cough detection for the multimodal model when trained with more data, consistent with the findings from the first experiment.

In the third experiment, we tested the model’s robustness to different proportions of OOD data during inference time and incorporated the ViM OOD detection algorithm to enhance this robustness. From Fig. 4, we observe that the multimodal model is more sensitive to OOD data without the OOD detection. However, when combined with the OOD detection algorithm, the multimodal model becomes stable and achieves the best overall results. This sensitivity may be due to the additional uncertainty introduced by IMU signals. For example, during silent periods, it is easier to distinguish using the audio modality, but the IMU signals may resemble those of cough or speech due to significant subject movement. We also observe a degradation in speech F1 scores in the single modal OOD model, primarily due to the decrease in speech recall from 0.89 to 0.76 when varying the OOD threshold from 0% to 50%. This performance drop can be attributed to two main factors: the OOD threshold being optimized for cough detection F1 scores rather than speech, and the inherent spectral similarity between speech and certain OOD classes (such as “speech-far” and “laugh”) as shown in Fig. 2 of the Supplementary Material. The spectral overlap becomes more pronounced at lower frequencies, leading to increased misclassification by the OOD detection algorithm. To further optimize OOD detection, we removed the “other” class during the fitting of eigenvalues for ViM. Tab. III shows that after removing the “other” class, OOD detection performance improves significantly, especially at lower f, while the cough recognition improves slightly. We hypothesize that this improvement is due to the unclear features in the “other” class caused by its multiple components, such as sneezes, groans, etc., indicating that data quality is also crucial for effective OOD detection.

Then, we analyzed the misclassified components of the best system we developed, which is the multimodal model trained with enhanced data integrated with ViM OOD detection using only cough and speech sounds. This analysis is illustrated in Figs. 5a and 5b. From these figures, we observe that the classes most prone to misclassification are nonverbal vocalization, including sneezes, groans, and laughs and instances from the minor classes within ”Other” are predominantly misclassified as either OOD as we excluded “Other” in ID eigenvalue extraction. To address this, we could incorporate more specific class data by training a large-scale model using publicly available online data for nonverbal vocalizations, as well as by collecting additional data on our own.

Fig. 5b clearly shows a gap between the means of the OOD score distribution for cough and speech. We use the best cough F1 score to select the threshold lead to the drop in speech detection performance, as seen in Fig. 4. However, we observe promising performance in speech detection without the OOD detection. Therefore, it is possible to improve speech detection by using an adaptive threshold based on the different classes predicted by the classifier.

V. Conclusion and Future Work

This paper presents our most recent efforts towards development of a robust cough detection system by integrating audio and IMU signals with an out-of-distribution (OOD) detection technique. This sensor system is suitable for integration into wearable devices thanks to its miniaturized size and low-cost computation, potentially facilitating at-home remote symptom monitoring of cough. We optimized the multimodal cough detection system by training with enhanced data and employing a weighted multi-loss approach for the ID classifier. We optimized OOD detection by only using the essential and high-quality classes (cough and speech). The results demonstrate the robustness of this system across window sizes from 1 to 5 seconds and performs efficiently at a lower audio frequency of 750 Hz. This is the highest down-sampling rate that muffles the speech causing it to be incomprehensible to protect user privacy. Specifically, with a 1.5-second window size, the system achieved 91.63% accuracy and a cough F1 score of 0.8754 at a 16 kHz audio frequency, and 87.72% accuracy and a cough F1 score of 0.8406 at 750 Hz on our self-collected dataset. Furthermore, while the multimodal model is sensitive to OOD input, the final optimized robust system demonstrated resilience to OOD data, achieving 90.08% accuracy and a cough F1 score of 0.7548 at a 16 kHz audio frequency, and 87.3% accuracy and a cough F1 score of 0.7015 at 750 Hz, even when half of the data was OOD during inference.

Future improvements will prioritize acquiring higher quality data, as data quality is crucial for developing both ID classifier and OOD detection, which relies on features extracted from the training data. Moreover, using online or self-collected datasets to enhance nonverbal vocalization detection is significant due to the high risk of misclassification of non-verbal sounds, especially at lower audio frequencies. We are also planning to expand our population to incorporate participants with various health conditions. Finally, incorporating additional modalities, such as heart rate and respiratory rate detection, electrocardiogram and other health symptoms of participants, can further enhance model performance. Besides system improvements, because this study was performed in healthy patients, this system will need to be further validated in patients with asthma and COPD.

References

[1] G. . 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, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith *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] V. Misra, A. Bozkurt, B. Calhoun, T. Jackson, J. S. Jur, J. Lach, B. Lee, J. Muth, Ö. Oralkan, M. Öztürk *et al.*, “Flexible technologies for self-powered wearable health and environmental sensing,” *Proceedings of the IEEE*, vol. 103, no. 4, pp. 665–681, 2015.

[5] Y. Chen, M. D. Wilkins, J. Barahona, A. J. Rosenbaum, M. Daniele, and E. Lobaton, “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.

[6] T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez, “A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom,” *Information Fusion*, p. 102040, 2023.

[7] Y. Chen, J. Barahona, I. Eldho, Y. Yu, R. Muhammad, B. Kutsche, M. L. Hernandez, D. Carpenter, A. Bozkurt, and E. Lobaton, “Robust multimodal cough and speech detection using wearables: A preliminary analysis,” in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–6.

[8] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, “Likelihood ratios for out-of-distribution detection,” in *Advances in Neural Information Processing Systems*, 2019, pp. 14 707–14 718.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[10] A. Nguyen, J. Yosinski, and J. Clune, “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.

[11] Y. Chen, P. Attri, J. Barahona, M. L. Hernandez, D. Carpenter, A. Bozkurt, and E. Lobaton, “Robust cough detection with out-of-distribution detection,” *IEEE Journal of Biomedical and Health Informatics*, 2023.

[12] M. Al-Khassaweneh and R. Bani Abdelrahman, “A signal processing approach for the diagnosis of asthma from cough sounds,” *Journal of medical engineering & technology*, vol. 37, no. 3, pp. 165–171, 2013.

[13] J.-M. Liu, M. You, Z. Wang, G.-Z. Li, X. Xu, and Z. Qiu, “Cough detection using deep neural networks,” in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2014, pp. 560–563.

[14] S. Matos, S. S. Birring, I. D. Pavord, and H. Evans, “Detection of cough signals in continuous audio recordings using hidden markov models,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1078–1083, 2006.

[15] I. D. Miranda, A. H. Diacon, and T. R. Niesler, “A comparative study of features for acoustic cough detection using deep architectures,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2601–2605.

[16] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, “Automatic cough detection in acoustic signal using spectral features,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 7153–7156.

[17] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, “Deep learning based cough detection camera using enhanced features,” *Expert Systems with Applications*, p. 117811, 2022.

[18] S. Jokić, D. Cleres, F. Rassouli, C. Steurer-Stey, M. A. Puhan, M. Brutsche, E. Fleisch, and F. Barata, “Tripletcough: Cougher identification and verification from contact-free smartphone-based audio recordings using metric learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2746–2757, 2022.

[19] F. Barata, K. Kipfer, M. Weber, P. Tinschert, E. Fleisch, and T. Kowatsch, “Towards device-agnostic mobile cough detection with convolutional neural networks,” in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2019, pp. 1–11.

[20] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *arXiv preprint arXiv:1610.02136*, 2016.

[21] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *arXiv preprint arXiv:1706.02690*, 2017.

[22] T. DeVries and G. W. Taylor, “Learning confidence for out-of-distribution detection in neural networks,” *arXiv preprint arXiv:1802.04865*, 2018.

[23] G. Shalev, Y. Adi, and J. Keshet, “Out-of-distribution detection using multiple semantic label representations,” in *Advances in Neural Information Processing Systems*, 2018, pp. 7375–7385.

[24] A. Subramanya, S. Srinivas, and R. V. Babu, “Confidence estimation in deep neural networks via density modelling,” *arXiv preprint arXiv:1707.07013*, 2017.

[25] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar, “Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance,” *arXiv preprint arXiv:1812.02765*, 2018.

[26] V. Abdelzad, K. Czarnecki, R. Salay, T. Denounden, S. Vernekar, and B. Phan, “Detecting out-of-distribution inputs in deep neural networks using an early-layer output,” *arXiv preprint arXiv:1910.10307*, 2019.

[27] K. Lee, H. Lee, K. Lee, and J. Shin, “Training confidence-calibrated classifiers for detecting out-of-distribution samples,” *arXiv preprint arXiv:1711.09325*, 2017.

[28] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure,” *arXiv preprint arXiv:1812.04606*, 2018.

[29] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, and K. Czarnecki, “Out-of-distribution detection in classifiers via generation,” *arXiv preprint arXiv:1910.04241*, 2019.

[30] H. Wang, Z. Li, L. Feng, and W. Zhang, “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.

[31] M. S. Graham, W. H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, and J. Cardoso, “Denoising diffusion models for out-of-distribution detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2947–2956.

[32] M. Pahar, I. Miranda, A. Diacon, and T. Niesler, “Automatic non-invasive cough detection based on accelerometer and audio signals,” *Journal of Signal Processing Systems*, vol. 94, no. 8, pp. 821–835, 2022.

[33] L. Orlandic, J. Thevenot, T. Teijeiro, and D. Atienza, “A multimodal dataset for automatic edge-ai cough detection,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023, pp. 1–7.

[34] TOZO, *TOZO Wireless Earbuds*, https://www.tozostore.com/.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[37] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[38] 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.

[39] J. F. Gemmeke, D. P. 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[40] F. Schmid, K. Koutini, and G. Widmer, “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.

[41] 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.

[42] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun *et al.*, “Openood: Benchmarking generalized out-of-distribution detection,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 598–32 611, 2022.