Sound-Dr: Reliable Sound Dataset and Baseline Artificial Intelligence System for Respiratory Illnesses

Truong V. Hoang[1], Quang H. Nguyen[2], Cuong Q. Nguyen[3], Phong X. Nguyen[4], and Hoang D. Nguyen[5]

[1,3,4] *AI Center, FPT Software Company Limited, Vietnam*
{*truonghv1, cuongnq1, phongnx1*}*@fsoft.com.vn*
[2] *Reliable Machine Learning Group, Vietnam*
nh.quang313@gmail.com
[5] *School of Computer Science and Information Technology, University College Cork, Ireland*
hn@cs.ucc.ie

ABSTRACT

As the burden of respiratory diseases continues to fall on society worldwide, this paper proposes a high-quality and reliable dataset of human sounds for studying respiratory illnesses, including pneumonia and COVID-19. It consists of coughing, mouth breathing, and nose breathing sounds together with metadata on related clinical characteristics. We also develop a proof-of-concept system for establishing baselines and benchmarking against multiple datasets, such as Coswara and COUGHVID. Our comprehensive experiments show that the Sound-Dr dataset has richer features, better performance, and is more robust to dataset shifts in various machine learning tasks. It is promising for a wide range of real-time applications on mobile devices. The proposed dataset and system will serve as practical tools to support healthcare professionals in diagnosing respiratory disorders. The dataset and code are publicly available here: https://github.com/ReML-AI/Sound-Dr/.

1. INTRODUCTION

Abnormalities can be discovered in the respiratory sounds of individuals with fever, asthma, tuberculosis, pneumonia, and COVID-19 compared to the sound of those without these conditions. A solid body of literature has shown the effectiveness of respiratory sounds in disease detection with the use of artificial intelligence (AI) (Song, 2015; Sakkatos et al., 2019; Yang et al., 2022). Furthermore, AI systems and data can be periodically updated, thereby improving accuracy and reliability. In real-world situations, sound-based medical screening tools can be widely deployed in multiple locations, such as airports, factories and supermarkets.

To date, there are several respiratory sound datasets to detect diseases, such as the Internal Conference on Biomedical Health Informatics (ICBHI) data (Rocha et al., 2019) in which each audio recording identifies the patients in terms of being healthy or exhibiting one of the following respiratory diseases or conditions including COPD, Bronchiectasis, Asthma, Upper and Lower respiratory tract infection, Pneumonia, and Bronchiolitis. To detect COVID-19, there are also two well-known datasets from New York (*NYU Breathing Sounds for COVID-19*, 2020) and Cambridge (Brown et al., 2020) universities. These respiratory datasets, however, are likely prone to reliability issues, as shown in our later experiments with the use of a dataset shift detection method (Rabanser, Günnemann, & Lipton, 2019).

In order to build a high-quality and reliable dataset, we developed a system to collect respiratory sound data in an efficient manner, such as recording each sound multiple times to reduce the impact of unwanted noises and capture the average sample's duration longer for better reliability. As a result, our dataset, named Sound-Dr, is collected under many different diseases, such as fever, asthma, and COVID-19, to enable researchers to solve various machine-learning problems related to respiratory diseases, including disease classification and anomaly detection.

The Sound-Dr dataset contains three types of respiration sounds, including nose breathing, mouth breathing, and coughing, with extensive lengths to foster more possibilities in machine learning algorithms and model deployments. Besides the audio recordings, metadata and health-related characteristics (e.g., smoking, insomnia) are included with high quality and data richness, which can be useful for multiple machine learning tasks on medical diseases related to respiratory systems.

Compared to existing datasets, the Sound-Dr dataset has multiple advantages with the following contributions:

- In collaboration with medical experts (*Woolcock Institute of Medical Research Vietnam*, 1981), we contribute a publicly available, high-quality, and reliable dataset with a sampling rate of 48,000 Hz and an average duration of 23s for respiration sounds, including breathing (nose & mouth) and coughing.

- Besides the audio recordings, the dataset also provides metadata and health-related characteristics for various tasks in machine learning, including but not limited to disease classification, anomaly detection, and symptom recognition for respiratory illnesses. It is suitable for large-scale adoption and deployment via smart devices for homes or businesses.

- This paper establishes an open baseline framework to facilitate benchmarking the Sound-Dr dataset and other datasets in terms of performance and robustness, as well as the efficiency of the data collection.

2. BACKGROUND

2.1. Studies of Human Sounds for Medical Screening

In medicine, human sounds have been well-studied as viable inputs for identifying vocal fold pathology, which involves either subjective or objective assessments. In subjective approaches, a skilled medical professional hears the sound signal and determines whether it is diseased or normal based on their prior training and experience. Nevertheless, depending on the level of experience, this type of evaluation may differ from doctor to doctor (Kreiman, Gerratt, & Precoda, 1990). As a result, both medical and engineering professionals are paying more and more attention to objective approaches to voice pathology.

Many medical conditions can be accurately identified using computer-aided voice pathology classification tools and deep learning techniques. For example, a recent study (Deb et al., 2022) proposed a deep learning-based classification model that can accurately predict whether a person has a cold or not based on their speech on URTIC dataset. This dataset is highly imbalanced, with only 2876 samples having Cold class. On the other hand, No Cold class contains 25776 samples. Moreover, Mo et al. (Mo, Gui, & Fletcher, 2022) proposes a computer-based deep learning algorithm that could enable rapid screening of the most common pulmonary diseases (COPD, Asthma, and respiratory infection (COVID-19)) using voluntary cough sounds alone.

During the global pandemic, Sharma et al. (N. K. Sharma et al., 2022) presented a challenge aimed at accelerating the research in acoustics-based detection of COVID-19, a valuable topic at the intersection of acoustics, signal processing, machine learning, and healthcare. This was an open call with great interest from the research community.

2.2. Machine learning datasets for respiratory diseases

The use of machine learning has become increasingly promising for the detection and monitoring of respiratory illnesses. A recent work (Pham et al., 2022) presented an exploration of various deep-learning models for detecting respiratory anomalies from auditory recordings. Authors used the ICBHI 2017 (Rocha et al., 2019), each audio recording contains one or different types of respiratory cycles, labeled as Crackle, Wheeze, Both Crackle & Wheeze, or Normal. Using a late fusion of inception based and transfer learning frameworks outperforms state of the art, recording the best score of 57.3%. Another study (Islam, Abdel-Raheem, & Tarique, 2022) applied a CNN for discriminating pathological voices from normal voices with the speech samples from Saarbrücken Voice Database, which is a collection of speech and electroglottography signals of more than 2000 speakers. Resulting in F1 scores of 78.7%.

There are also recent datasets for respiratory diseases, such as COUGHVID (Orlandic, Teijeiro, & Atienza, 2021) and Coswara (N. Sharma et al., 2020). COUGHVID is a global cough signal recordings dataset for COVID-19 detection with some clinical information and metadata. And Coswara is another dataset composed of voice samples from healthy individuals, including breathing sounds (fast and slow), cough sounds (deep and shallow), phonation of sustained vowels, and counting numbers. These datasets are large-scale and regularly updated; nevertheless, they are susceptible to reliability issues due to their intrinsic properties and distributional characteristics. We use these two datasets for performance benchmarking and evaluate the dataset shift problem.

3. SOUND-DR DATASET AND TASK DEFINITION

3.1. Sound-Dr Dataset Collection

Sound-Dr dataset is a project[1] of AI Center of FPT Software Company Limited (FPT, 1999), an entity in charge of AI research and development with consultation from the Woolcock Institute of Medical Research, Vietnam (*Woolcock Institute of Medical Research Vietnam*, 1981). The application collects users' demographic information, medical history, and records of voices and respiratory sounds. The Sound-Dr dataset was solely collected by FPT for community purposes during the peak season of the COVID-19 pandemic in Vietnam from August 2021 to October 2021. We treat ethical issues as important, and users were prompted to read about our terms and conditions and give us their consent solely for development and community purposes. Therefore, the dataset has the agreement and consent of all users.

In the data collection, we developed web-based and mobile-based applications for users to easily interact and record three

[1] https://www.fpt-software.com/fpt-softwares-ai-app-listens-out-for-respiratory-diseases-amid-looming-covid-19

different sounds: (1) mouth breathing, (2) nose breathing, and (3) coughing. With the involvement of medical experts in the field, for each audio type, users are requested to record at least three times with a minimum duration of 5 seconds in each turn. The sample rate of 48,000 Hz is set to be the default, and no noise reduction method is used in web-based or mobile-based applications to collect the true nature of the data. Additionally, some metadata of users is also collected via a survey form which includes personal information (e.g., age and gender), related respiratory illness symptoms, smoking status, and COVID-19 diagnosis, as shown in Table 1.

Table 1. Metadata fields of the Sound-Dr dataset.

Categories	Fields	Details
Demographics	Sex options	Gender: Male, Female
	Ages	Age
	Current city	The current city living
Related-Flu Symptoms	Symptoms status choice	Symptoms since last 14 days: Fever, Chills, Sore throat, Dry cough, Wet cough, Stuffy nose, Snivel, Headache, Difficulty breathing or feeling short of breath, Muscle aches, Dizziness, Confusion or vertigo, Tightness in your chest, Loss of taste and smell, None
Medical Conditions	Condition choice	The medical conditions of the subject: Asthma, Diabetes, Cystic fibrosis, COPD/Emphysema, Pulmonary fibrosis, Other lung diseases, Angina, Previous stroke or Transient ischaemic attack, Cancer, Previous heart attack, Valvular heart disease, HIV or impaired immune system, Other long-term condition, Other heart diseases, Previous organ transplant, None
Insomnia Symptoms	Insomnia status choice	How often the subject suffers from insomnia: Never, Once in the last 2 weeks, Once a week, 2 to 3 days a week, 4 days a week or more
Smoking Status	Smoking status options	Never smoked, Ex-smoker, Current smoker (less than once a day), Current smoker (1-5 cigarettes per day), Current smoker (11-20 cigarettes per day), Current smoker (21+ cigarettes per day),
COVID-19 Status	Cov19 status choice	How long has had a positive test for COVID-19: Never, In the last 14 days, More than 14 days ago.
	F condition choice	Status with COVID-19 of a subject

3.2. Descriptive Statistics of Sound-Dr Dataset

We obtained a dataset of 3,930 sound recordings; the distribution of coughing, mouth breathing, and nose breathing is presented in Figure 1. There are 1,310 in total subjects with gender distribution shown in Figure 2 and age distribution presented in Figure 4. It can be seen that more males (e.g. 60%) than females (e.g. 40%) participated in our program. In terms of age groups, subjects between 20 and 40 years old are dominant. Regarding the smoking status, Figure 3 indicates that 90% of the subjects are non-smokers. From 346 COVID-19 positive subjects, there are 293 objects with symptoms and 193 objects without symptoms.

Statistics of the duration, shown in Figure 5, reveal several interesting characteristics in quantifying above mentioned datasets. Some lengths of audio in Coswara and COUGHVID samples are less than 1s, which might lead to being unqualified for the training model and errors in the reading input data process. Nevertheless, the Sound-Dr dataset has longer durations to ensure that training data is split into even parts and does not require padding when the sound length is unsatisfactory. Moreover, regarding statistics of the sampling rate, the Coswara and COUGHVID datasets have audio sampling rates of 44100Hz and 22050Hz, respectively. The Sound-Dr dataset has a higher sampling rate; therefore, our training data can be easily converted to different sampling rates for benchmarking. Our data collection system was built skillfully to ensure the best quality for machine learning tasks.

3.3. Task Definition

Given the Sound-Dr dataset, we propose three main tasks: (I) Detect negative or positive COVID-19 subjects, (II) Detect subjects with and without related respiratory symptoms, and (III) Detect normal subjects and anomaly subjects (i.e., anomaly subjects are positive COVID-19 or present related respiratory symptoms). For each task, the audio input of coughing, mouth breathing mouth, and nose breathing are evaluated independently.

Based on the metadata as shown in Table 1, the total of 1,310 subjects are separated into: COVID-19 negative and COVID-19 positive subjects, subjects with and without symptoms, and normal and anomaly subjects for task I, II, and III respectively as shown in Figure 6.

To evaluate the Sound-Dr dataset for each defined task, we apply a 5-fold cross-validation method where the final result is an average of all folds. Random seeds are used to ensure that results are reproducible and the data is divided the same way into all methods for benchmarking. Every experiment's result is an average of the 5 different seeds. The evaluation metrics in use are Accuracy (Acc), F1 score (Sasaki, 2007), and AUC (Bradley, 1997). All the model training, benchmarking, and evaluation tasks have been executed in a system with Ubuntu 18.04, 12 GB RAM, and NVIDIA GTX 1080 GPU.

4. Baseline System

Given the Sound-Dr dataset, we develop a deep-learning-based framework to explore, which is referred to as the baseline. Generally, the baseline framework can be separated into two main steps: Feature extraction and Classification.

4.1. Feature Extraction

The raw audio from one channel (mono) is firstly re-sampled with a sample rate of 16000 Hz using Librosa toolkit (McFee

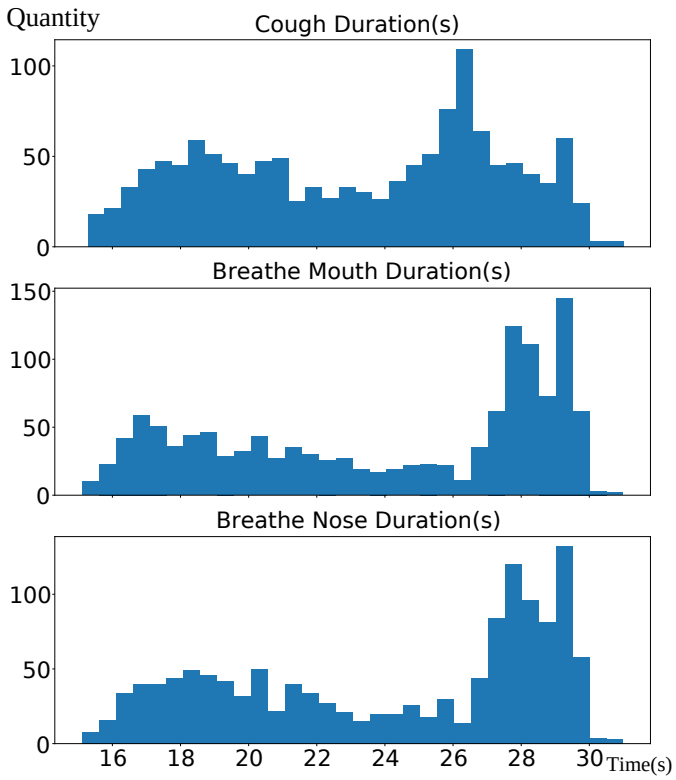

Figure 1. Histograms of three types of audio recording duration.

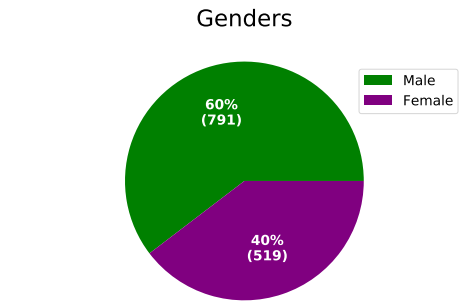

Figure 2. Gender Data Distribution.

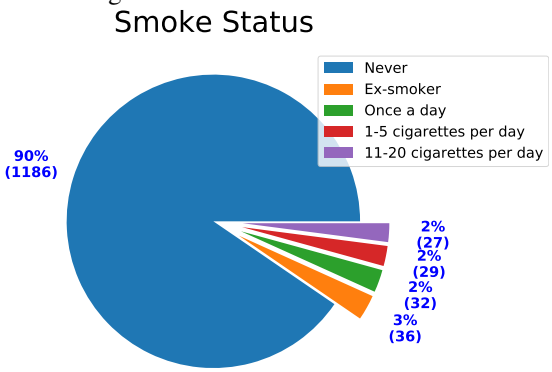

Figure 3. Smoke Status Distribution.

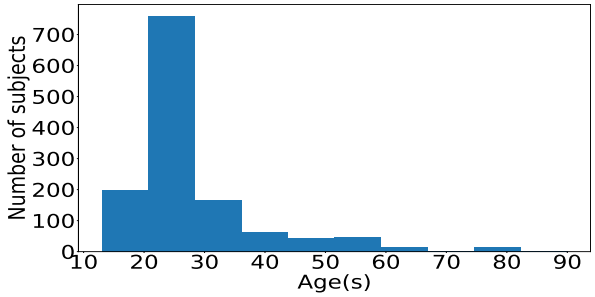

Figure 4. Age Group Data Distribution.

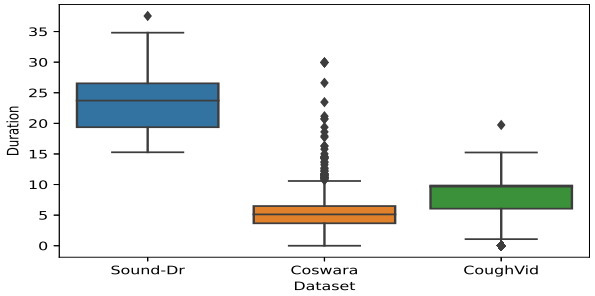

Figure 5. The distribution of the duration of datasets.

et al., 2015). Then, re-sampled audio recordings are fed into a pre-trained model to extract embedding features. In this paper, the pre-trained model is from both TRILL (Shor et al., 2020) and FRILL (Peplinski, Shor, Joglekar, Garrison, & Patel, 2021), which is recommended for downstream tasks on non-semantic speech signals. Using TRILL to extract features from Cough sounds for detecting COVID-19 has also been proven effective (Hoang, Pham, Ngo, & Nguyen, 2022).

While the pre-trained TRILL model is based on ResNet architecture presenting a large footprint, the pre-trained FRILL model is built on MobileNet architecture, leveraging knowledge distillation from the pre-trained TRILL model. As a result, the pre-trained FRILL model is suitable for real-time application on edge devices (i.e., FRILL pre-trained model is 32 times faster on a Pixel 1 smartphone and equals to 40% of TRILL size, but still competitive to TRILL model with an average decrease of only 2% in terms of accuracy).

The outputs of both pre-trained models are a time series of one embedding. This means we obtain one embedding (i.e., 2048-dimensional vector) from every second when feeding the audio recordings with different lengths from the Sound-Dr dataset into the models. Hence, we obtain multiple embeddings representing one audio recording. Consequently, we conduct two statistical features of mean and standard deviation across the time axis. We then concatenate these features to create the final embedding (4096-dimensional vector).

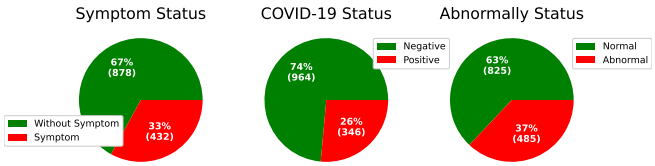

Figure 6. The number of subjects for each task defined.

Table 2. The setting parameters of Classification

Tasks	Models	Setting Parameters
Task I: COVID-19 Subject Detection	XGB	max_depth = 6, learning_rate = 0.07, scale_pos_weight = 2.78, n_estimators = 200, subsample = 1, colsample_bytree = 1, eta = 1, objective = 'binary:logistic', eval_metric = 'auc'
Task II: Symptom Subject Detection, **Task III:** Anomaly Subject Detection	XGB, XGBOD	max_depth = 7, learning_rate = 0.3, scale_pos_weight = 1.7, n_estimators = 200, subsample = 1, colsample_bytree = 1, nthread = -1, eval_metric = 'logloss'
	Isolation Forest	n_estimators = 500, max_samples = 'auto', contamination = 0.1

4.2. Classification

We conduct experiments on the Support Vector Machine, Random Forest, Multilayer Perceptron, ExtraTrees Classifier, LightGBM, and XGB Classifier. Anomaly detection mostly focuses on unsupervised or semi-supervised settings, we use Isolation Forest (Liu, Ting, & Zhou, 2008), and XGBOD (Zhao & Hryniewicki, 2018) for actually seeing the usage of this dataset for anomaly detection for recognizing the outliers. However, we only achieved a good score on XGB Classifier for both Coswara and COUGHVID. Therefore, to build a baseline system and classify extracted embedding features into certain groups defined in Section 3.3, we use XGB Classifier (Friedman, 2000). To fine-tune hyper-parameters of this classifier, shown in Table 2, we make use of the Optuna framework (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) with the Grid Search algorithm. All these classification models are implemented by using XGBoost library (Chen & Guestrin, 2016) for XGB Classifier, Python Outlier Detection library (Zhao, Nasrullah, & Li, 2019) for XGBOD, and Scikit-Learn toolkit (Pedregosa et al., 2011) for the others.

4.3. Experimental Results and Discussion

We experimented with the task of COVID-19 Detection based on the three collected sound types: Cough, Breathing mouth, and Breathing nose, as illustrated in Table 3. The performance using Breathe Mouth and Breathe Nose is lower compared to the Cough sound data. The best performance using Cough sound scores 88.44 AUC, 73.13 F1, 86.06 Accuracy. Although TRILL outperforms FRILL on Accuracy by about 0.2% (86.06-86.26 Acc), on F1 and AUC metrics, FRILL performs better for 2% (73.13-71.34, 88.44-86.56 AUC). Therefore, we use FRILL for our baseline model as it is satisfactory for the real environment that needs fast, accurate detection, especially on mobile devices.

Table 3. The experimented results on Sound-Dr dataset over five runs. Results in **bold font** mark the best results given the same (fair) task.

Data Type	Feature - Classifier	Detection Task	Acc Mean	F1 Mean	AUC Mean (Std)
Cough	TRILL - XGB	Symptom	78.78	67.22	80.86 (00.58)
		COVID-19	**86.56**	71.34	86.56 (01.09)
		Abnormal	**77.25**	67.46	79.20 (00.66)
	FRILL - XGB	Symptom	**79.05**	**67.67**	**81.23 (00.33)**
		COVID-19	86.06	**73.13**	**88.44 (00.38)**
		Abnormal	77.18	**68.12**	**81.16 (01.11)**
Breathing Mouth	TRILL - XGB	Symptom	74.43	62.65	78.45 (01.03)
		COVID-19	82.75	67.44	85.55 (00.79)
		Abnormal	75.64	65.66	78.03 (01.28)
	FRILL - XGB	Symptom	79.31	65.65	80.57 (00.65)
		COVID-19	86.18	70.76	87.23 (01.14)
		Abnormal	75.04	66.04	78.79 (00.62)
Breathing Nose	TRILL - XGB	Symptom	78.47	63.75	78.75 (01.08)
		COVID-19	86.11	70.16	85.04 (00.72)
		Abnormal	76.34	64.37	78.24 (00.71)
	FRILL - XGB	Symptom	79.54	65.19	79.80 (00.86)
		COVID-19	84.50	70.28	85.79 (00.97)
		Abnormal	77.10	67.74	80.75 (00.93)

Abnormal: Symptom + COVID-19.

In addition, we also experiment with the Abnormal Detection in respiratory sound by adjusting the label which we combine the COVID-19 Positive and Symptomatic status into Abnormal labels. Using XGB Classifier with hyper-parameters shown in Table 2, we achieve promising results of 81.16 AUC, 68.12 F1, and 77.18 Accuracy. On XGBOD, results of 82.95 AUC, 70.02 F1, and 79.77 Accuracy show that the unsupervised settings can be used on this dataset for anomaly detection. The performance comparison is described in Table 4. This shows that our dataset has potential for more reliable outcomes on multiple tasks, such as Outlier Detection and Anomaly Detection in Respiration Sound. We hope that models based on the Sound-Dr dataset could be built to support the doctor's diagnosis of disease faster and more accurately in the future.

5. Benchmarks

5.1. Dataset Shift Detection

Besides evaluating the effectiveness of the models applied on 3 datasets, we parallelly consider the dataset shift problem that contributes to measuring the dataset's robustness. It happens due to the different distributional characteristics of data between train and test set (Quionero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009).

Table 4. The benchmark results of unsupervised methods on other datasets over five runs for Abnormal Detection task (Symptom + COVID-19). Results in **bold font** mark the best results given the same (fair) dataset.

Data Type	Feature - Classifier	Acc Mean	F1 Mean	AUC Mean (Std)
Coswara	FRILL - IsolationForest	76.63	16.37	49.82 (02.99)
	FRILL - XGBOD	76.23	41.47	67.44 (01.25)
COUGHVID	FRILL - IsolationForest	74.99	15.38	49.45 (00.68)
	FRILL - XGBOD	49.62	33.96	58.84 (01.62)
Sound-Dr (Ours)	FRILL - IsolationForest	60.31	15.86	53.64 (00.54)
	FRILL - XGBOD	**79.77**	**70.02**	**82.95 (01.33)**

Table 5. Detecting Dataset Shift Using Failing Loudly

Data type	Number of samples from test				
	50	100	500	1000	10000
Coswara	0.28	0.44	0.43	0.43	0.42
COUGHVID	0.39	0.44	0.40	0.80	0.41
Sound-Dr (Ours)	**0.25**	**0.29**	**0.39**	**0.38**	**0.38**

Many machine learning algorithms are based on the assumption that the training and test data are drawn from the same distribution; thus, dataset shift might lead to the model's tremendous performance degradation. We qualify the robustness of the dataset between Sound-Dr, Coswara, and COUGHVID by detecting accuracy shifts indicating the degree of distribution shifting in the dataset.

We conducted several experiments based on a pipeline for detecting dataset shift by a two-sample-testing-based approach, using pre-trained classifiers for dimensionality reduction (Rabanser et al., 2019). Specifically, the train and test set is reduced in dimension and subsequently analyzed via statistical hypothesis testing. We investigate the equivalence of the source distribution (from which training data is sampled) and target distribution (from which real-world data is sampled). The shifting of datasets is evaluated with various amounts of samples including 50,100,500,1000,10000 accordingly. Table 5 shows that the Sound-Dr dataset exhibits less shifting in train-test distribution. Over samples, our result reached better values of about 15% and 26% with respect to Coswara and COUGHVID; thereby leading to lesser risk of drifting and more reliability for real-world deployment.

5.2. Task Performance

The paper aims to establish performance benchmarks for multiple machine-learning tasks. We exploit extracted features through SVMs with linear kernels for classification tasks. Specifically, we use several extraction methods including FRILL, OpenSmile (Eyben, Wöllmer, & Schuller, 2010), OpenXBOW (Schmitt & Schuller, 2017) and Deep Spectrum (Amiriparian et al., 2017) to extract feature representations from preprocessed raw audio data. Acquired representations were scaled to zero mean and unit standard deviation following the parameters from the respective training set. These normalized features were applied to the SVM model employed by the Scikit-Learn toolkit (Pedregosa et al., 2011) with its class LINEARSVC with the optimized complexity parameter C. We conduct experiments in these settings and unify them to a result in Table 6.

We utilise the same feature extraction process and classifier (SVM) for COVID-19 and abnormal detection tasks on datasets. The experiment results on the Sound-Dr dataset are better than the two other datasets in Table 6. The task per-

Table 6. The benchmark results of supervised methods on other datasets over five runs. Results in **bold font** mark the best results given the same (fair) feature.

Data Type	Feature - Classifier	Detection Task	Acc	F1	AUC
			Mean	Mean	Mean (Std)
Coswara	FRILL - SVM	COVID-19	74.97	40.94	65.12 (01.42)
		Abnormal	70.78	39.16	58.71 (00.92)
	OpenSmile - SVM	COVID-19	74.83	43.58	67.61 (01.69)
		Abnormal	**68.24**	32.85	58.70 (01.33)
	DeepSpectrum -	COVID-19	**75.19**	17.52	51.48 (01.71)
	SVM	Abnormal	**72.38**	19.36	51.31 (01.66)
	OpenXBoW 1000 -	COVID-19	76.92	41.07	64.97 (01.16)
SVM		Abnormal	**73.10**	29.52	56.91 (02.10)
	OpenXBoW 2000 -	COVID-19	80.28	42.29	65.15 (01.40)
SVM		Abnormal	**73.74**	29.51	56.99 (01.27)
	OpenXBoW 3000 -	COVID-19	78.01	43.98	65.01 (02.21)
SVM		Abnormal	**76.87**	30.56	56.00 (01.98)
COUGHVID	FRILL - SVM	COVID-19	81.49	16.89	54.60 (00.63)
		Abnormal	64.69	23.55	50.89 (00.43)
	OpenSmile - SVM	COVID-19	**80.80**	15.68	53.66 (01.35)
		Abnormal	65.34	25.01	51.89 (01.12)
	DeepSpectrum -	COVID-19	49.78	14.15	49.02 (01.32)
	SVM	Abnormal	49.00	27.55	49.53 (00.58)
	OpenXBoW 1000 -	COVID-19	**79.47**	15.77	53.67 (01.87)
SVM		Abnormal	72.12	21.62	52.26 (01.40)
	OpenXBoW 2000 -	COVID-19	**86.20**	11.45	51.86 (01.67)
SVM		Abnormal	69.84	21.40	51.37 (00.94)
	OpenXBoW 3000 -	COVID-19	**83.04**	15.00	55.35 (00.84)
SVM		Abnormal	67.67	21.59	50.72 (00.78)
Sound-Dr	FRILL - SVM	COVID-19	**82.53**	**70.48**	**81.37 (00.85)**
(Ours)		Abnormal	**76.11**	**69.45**	**75.54 (00.26)**
	OpenSmile - SVM	COVID-19	69.31	**56.96**	**71.75 (01.41)**
		Abnormal	65.04	**59.33**	**65.83 (01.92)**
	DeepSpectrum -	COVID-19	64.58	**38.62**	**57.41 (01.51)**
	SVM	Abnormal	58.02	**45.22**	**55.71 (00.95)**
	OpenXBoW 1000 -	COVID-19	75.40	**59.56**	**72.77 (01.96)**
SVM		Abnormal	64.43	**55.19**	**63.35 (01.65)**
	OpenXBoW 2000 -	COVID-19	75.25	**60.27**	**73.42 (02.10)**
SVM		Abnormal	66.26	**59.60**	**66.46 (02.00)**
	OpenXBoW 3000 -	COVID-19	75.56	**59.84**	**72.99 (01.32)**
SVM		Abnormal	68.02	**58.31**	**66.45 (01.69)**

Abnormal: Symptom + COVID-19.

formance improvements are statistically significant for both COVID-19 Detection and Abnormal Detection on both the Coswara and COUGHVID datasets respectively.

It demonstrates that the Sound-Dr dataset might provide potential features for detecting anomalies in respiratory sounds such as cough and breath. In addition, better results of the Sound-Dr dataset indicate that our dataset was processed well to obtain high-quality samples during data collection.

6. Conclusion

High-quality respiratory sound data, which can be used to detect patient symptoms, is in demand; thus, the Sound-Dr dataset is essential for researchers to build health applications. We also build a system to evaluate multiple datasets and create the first baseline system for future research and benchmarking. Based on our comprehensive experimental results, the Sound-Dr dataset is better than multiple existing datasets in terms of both unsupervised and supervised methods. Therefore, the Sound-Dr dataset is effectively collected with extensive lengths to minimize various noises. Furthermore, our dataset's unique properties and metadata of health-related characteristics are more reliable against dataset shifts.

We build a model using FRILL embedding and XGBoost classifier for potential real-life context that necessitates rapid and accurate detection. It also helps the researchers easy to explore to improve the performance compared with the baselines. With the baseline system and dataset available, researchers have the advantage of rapid development of solutions in high demand. With the Sound-Dr dataset, we hope that researchers accelerate the building of Artificial Intelligence models to support doctors diagnose diseases faster and more accurately. The Sound-Dr dataset is collected from various mobile devices, with rigorous data collection methods, promising to apply widely in real-world situations.

With the increasing impact of respiratory illnesses, the Sound-Dr dataset is proposed in collaboration with medical experts to study respiratory anomalies, including pneumonia and COVID-19. As the baseline, this dataset can be useful to qualify respiratory disease screening/abnormal detection/symptom classification. In real-world scenarios, the dataset has been used in multiple medical apps for rapid screening due to its quality and robustness such as Respiratory diseases, COVID-19, and Respiratory anomalies.

In our pipeline, more data are needed in the field to enhance neural networks optimally. Although we provide an additional dataset on respiration to increase the distribution of data, more data are needed across many countries with a larger number of subjects. By collecting data from subjects from South East Asia, our research aims to provide the groundwork for future advancements in information processing and machine learning.

Acknowledgment

This work is supported by the FPT Software AI Center of FPT Software Company Limited (FPT, 1999). FPT Software is a global technology and IT services provider headquartered in Hanoi, Vietnam.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *The 25th acm sigkdd international conference on knowledge discovery and data mining* (p. 2623–2631). doi: 10.1145/3292500.3330701

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., ... Schuller, B. (2017, August). Snore sound classification using image-based deep spectrum features. In *International speech communication association (interspeech)* (pp. 3512–3516).

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., ... Mascolo, C. (2020). Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery and data mining* (p. 3474–3484). doi: 10.1145/3394486.3412865

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Acm sigkdd international conference on knowledge discovery and data mining* (p. 785–794). doi: 10.1145/2939672.2939785

Deb, S., Warule, P., Nair, A., Sultan, H., Dash, R., & Krajewski, J. (2022, 9). Detection of common cold from speech signals using deep neural network. *Circuits, Systems, and Signal Processing*. doi: 10.1007/s00034-022-02189-y

Eyben, F., Wöllmer, M., & Schuller, B. (2010, 01). opensmile – the munich versatile and fast open-source audio feature extractor. *ACM Multimedia 2010 International Conference*, 1459-1462. doi: 10.1145/1873951.1874246

FPT. (1999). *Fpt softwave company limited.* Retrieved from `https://www.fpt-software.com` ([Online; accessed 30-04-2023])

Friedman, J. (2000, 11). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*. doi: 10.1214/aos/1013203451

Hoang, T., Pham, L., Ngo, D., & Nguyen, H. D. (2022). A cough-based deep learning framework for detecting

covid-19. In *The 44th ieee engineering in medicine and biology society (embc)* (p. 3422-3425). doi: 10.1109/EMBC48229.2022.9871179

Islam, R., Abdel-Raheem, E., & Tarique, M. (2022). A novel pathological voice identification technique through simulated cochlear implant processing systems. *Applied Sciences*, *12*(5). doi: 10.3390/app12052398

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech, Language, and Hearing Research*, *33*(1), 103–115.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Eighth ieee international conference on data mining* (p. 413-422). doi: 10.1109/ICDM.2008.17

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18–25).

Mo, A., Gui, E., & Fletcher, R. R. (2022). Use of voluntary cough sounds and deep learning for pulmonary disease screening in low-resource areas. In *Ieee global humanitarian technology conference (ghtc)* (p. 242-249). doi: 10.1109/GHTC55712.2022.9911027

NYU Breathing Sounds for COVID-19. (2020). `https://breatheforscience.com/`. ([Online; accessed 30-March-2023])

Orlandic, L., Teijeiro, T., & Atienza, D. (2021). The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Springer Science and Business Media LLC*, *8*(1). doi: 10.1038/s41597-021-00937-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peplinski, J., Shor, J., Joglekar, S., Garrison, J., & Patel, S. (2021). FRILL: A Non-Semantic Speech Embedding for Mobile Devices. In *International speech communication association (interspeech)* (pp. 1204–1208). doi: 10.21437/Interspeech.2021-2070

Pham, L., Ngo, D., Tran, K., Hoang, T., Schindler, A., & McLoughlin, I. (2022). An Ensemble of Deep Learning Frameworks for Predicting Respiratory Anomalies. In *Ieee engineering in medicine & biology society (embc)* (p. 4595-4598). doi: 10.1109/EMBC48229.2022.9871440

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). *Failing loudly: An empirical study of methods for detecting dataset shift*. Curran Associates Inc.

Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., ... de Carvalho, P. (2019, mar). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, *40*(3), 035001. doi: 10.1088/1361-6579/ab03ea

Sakkatos, P., Barney, A., Bruton, A., Haitchi, H. M., Kurukulaaratchy, R. J., & Thackray, D. (2019). Quantified breathing patterns can be used as a physiological marker to monitor asthma. *European Respiratory Journal*, *54*(suppl 63). doi: 10.1183/13993003.congress-2019.PA5038

Sasaki, Y. (2007, 01). The truth of the f-measure. *Teach Tutor Mater*.

Schmitt, M., & Schuller, B. (2017). openxbow – introducing the passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*, *18*(96), 1–5.

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., R., N., ... Ganapathy, S. (2020). Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *International speech communication association (interspeech)* (pp. 4811–4815). doi: 10.21437/Interspeech.2020-2768

Sharma, N. K., Chetupalli, S. R., Bhattacharya, D., Dutta, D., Mote, P., & Ganapathy, S. (2022). The second dicova challenge: Dataset and performance analysis for diagnosis of covid-19 using acoustics. In *Ieee international conference on acoustics, speech and signal processing (icassp)* (p. 556-560). doi: 10.1109/ICASSP43922.2022.9747188

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chaumont Quitry, F., ... Haviv, Y. (2020). Towards Learning a Universal Non-Semantic Representation of Speech. In *International speech communication association (interspeech)* (pp. 140–144). doi: 10.21437/Interspeech.2020-1242

Song, I. (2015). Diagnosis of pneumonia from sounds collected using low cost cell phones. In *International joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN.2015.7280317

Woolcock institute of medical research vietnam. (1981). Retrieved from `https://www.woolcockvietnam.org` ([Online; accessed 30-04-2023])

Yang, Y., Yuan, Y., Zhang, G., Hao Wang, Y.-C. C., Liu, Y., Tarolli, C. G., ... Katabi, D. (2022, august). *Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals.* Nature Medicine. doi: 10.1038/s41591-022-01932-x

Zhao, Y., & Hryniewicki, M. K. (2018). Xgbod: Improving supervised outlier detection with unsupervised representation learning. In *International joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN.2018.8489605

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, *20*(96), 1-7.