INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Personalization for Robust Voice Pathology Detection in Sound Waves

Khanh-Tung Tran[1], *Truong Hoang*[1], *Duy Khuong Nguyen*[1], *Hoang D. Nguyen*[2], *Xuan-Son Vu*[3#]

[1]AI Center, FPT Software Company Limited, Vietnam
[2]School of Computer Science and Information Technology, University College Cork, Ireland
[3]Department of Computing Science, Umeå University, Sweden

`tungtk2@fpt.com, truonghv1@fpt.com, khuongnd6@fpt.com, hn@cs.ucc.ie, sonvx@cs.umu.se`

Abstract

Automatic voice pathology detection is promising for non-invasive screening and early intervention using sound signals. Nevertheless, existing methods are susceptible to covariate shifts due to background noises, human voice variations, and data selection biases leading to severe performance degradation in real-world scenarios. Hence, we propose a non-invasive framework that contrastively learns personalization from sound waves as a pre-train and predicts latent-spaced profile features through semi-supervised learning. It allows all subjects from various distributions (e.g., regionality, gender, age) to benefit from personalized predictions for robust voice pathology in a privacy-fulfilled manner. We extensively evaluate the framework on four real-world respiratory illnesses datasets, including Coswara, COUGHVID, ICBHI, and our private dataset - ASound under multiple covariate shift settings (i.e., cross-dataset), improving up to 4.12% in overall performance.

Index Terms: covariate shift, robust voice pathology detection

1. Introduction

The advent of sound and speech technology has opened many new possibilities in voice pathology detection, such as chronic diseases and infectious respiratory conditions (e.g., COVID-19) [1, 2, 3]. Conventional methods for respiration monitoring, such as thoracic impedance pneumography [4], or capnography [5], are either too invasive or inconvenient. Recent studies show the use of artificial intelligence (AI) based pathology detection for cost-effective and non-invasive screening and monitoring of a wide spectrum of diseases. Nevertheless, such models are susceptible to reliability issues in real-world scenarios as trained models tend to learn mixed characteristics linked to personal information under various categories rather than downstream features, thereby leading to poor performance in action. Figure 1 illustrates covariate shift in a cross-dataset validation scenario, where a classification model is trained to achieve acceptable performances on one dataset (source) and then fails when being tested on similar (target) datasets. Indeed, human sounds have distinctive characteristics associated with disease-related information which can be exploited as latent profile features for better learning and reliability.

In this work, we propose an end-to-end AI-based Voice Pathology Detection framework with two aims: (1) learning latent profile characteristics in sound waves, and (2) maintaining the robustness of AI systems due to the distribution shift issue. A transformer-based model is pre-trained to learn personalized features through masked contrastive learning with negatives sampled from other users. Then, this model is used to extract latent profile embedding for each subject based on all samples they provided. Additionally, the model can be self-trained to enhance the embedding and support users with insufficient samples, i.e., new patients. The profile embedding is used in finetuning with a traditional classification model in an end-to-end manner. Our proposed approach fulfills privacy compliance by learning latent space features through anonymized sound data without utilizing additional personal information.

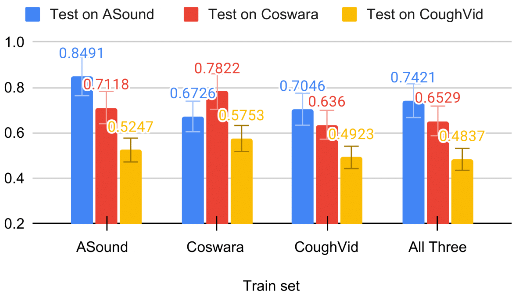

Figure 1: *Performance degradation due to covariate shift. A binary classification model is trained on one of the three datasets (ASound, Coswara, or COUGHVID) or all three at once. We observed that there is a* huge performance gap *(up to 32.44%) when trained and tested in cross-dataset scenarios.*

Our main contributions are listed as follows:

- To our best knowledge, it is the first work to propose a personalization strategy within a unified deep-learning framework to mitigate covariate shifts in the domain of sound data without touching sensitive meta information.
- Introduce a novel personalized encoding method using each data subject's sound samples to build their profiles based on pre-training with contrastive learning and can be applied to all users through semi-supervised learning.
- Evaluate our methods with comprehensive experiments on multiple real-world datasets, both publicly available, such as Coswara, COUGHVID, ICBHI, and our private datasets, namely ASound. The results demonstrate the efficacy of our proposed framework.
- Source codes, reproducible baselines, and our new dataset are made publicly available at `https://github.com/ReML-AI/RoPADet` for future research and benchmarking purposes.

2. Related Work

Covariate shift refers to possible changes in the distribution of the input variables present in the training and the testing data

[#] Corresponding author.

10.21437/Interspeech.2023-1332

- i.e., $P_{train}(X, y) \neq P_{test}(X, y)$. The continuous nature of sound data brings a variety of noisy aspects unique to each user, such as intrinsic medical conditions, gender, and age, leading to covariate shifts, especially when deployed to real-world settings. Covariate shift is a reoccurring problem and it has been studied in multiple works [6, 7, 8, 9] regarding the applicability and efficacy of such models in real-world scenarios. However, it is still an under-research problem in the sound domain.

Personalization is a well-established and complex topic, with multiple applications, such as in personalized speech recognition [10] or hate speech detection [11]. It also has a multitude of medical applications [12]. A typical personalized machine learning method leverages additional cues, such as sensitive metadata related to each user. Another direction utilize additional samples of the same user as the user embedding. For instance, in [13], the authors leveraged other samples of the present user as pseudo-sources for emotion recognition or speech enhancement. Perhaps the most related work to this research is the work done by [14] in which they explore longitudinal audio samples over time of each patient as time series for COVID-19 progression prediction and recovery trend prediction. To our best knowledge, we are the first to introduce a personalization technique that enables learning latent characteristics as a pre-train and predicting profile embeddings for unseen data towards robust sound classification.

3. Methodology

In this section, we describe our novel personalization framework for sounds, named RoPADet, that leverages pre-training to learn latent features for each individual directly from connected sound signals and infer profile embeddings for unseen human sounds through semi-supervised self-training. These embeddings capture individual respiratory-related characteristics and disentangle them from illness-related features. With the profiling mechanisms, covariate issues can be addressed, for example, when audio samples are collected from different areas (e.g., in a different country), but can be useful for early detection in other areas with personalization. In COVID-19, the intuition is that $y_{EU}^{Omicron} = \{y_{Africa}^{Omicron} + pred_{Africa}\} - pred_{EU}$, where disentangled information learnt from one region can aid in early prevention of the pandemic in another region. Therefore, our framework is robust to covariate shift and privacy-preserving without using sensitive data such as gender or maturity.

3.1. Latent profiling for personalization

We investigate profile information that captures each subject's unique characteristics and base signals, regardless of the downstream task. In order to train an effective profile extractor, masked contrastive learning is employed to predict masked time step with distractors sampled from audios of other users. This helps learn a discriminative model [15] that distinguishes between users by their uniqueness.

Formally, define $D_P = (X_i, y_i, u_i)$ as the dataset of samples with profile labels u_i (user identifier) and $D_{NP} = (X_j, y_j)$ as the dataset of samples with no profile labels. X_i is the sound samples with label y_i for downstream task T.

After the profile pre-training phase on $D = D_P \cup D_{NP}$, we obtained a profile extractor $f_{\theta_{profile}}$. We use it to generate initial profiles for each user u in D_P:

$$pred_u = \frac{1}{N_u} \Sigma_i^{u_i=u} f_{\theta_{profile}}(X_i) \tag{1}$$

where N_u is the total amount of samples belong to user u. Profile $pred_u$ will be used for all input X_i to the downstream task where $u_i = u$. Grouping and extracting profile features from multiple samples of the same subject, invariant to the target task labels are necessary to obtain features unique to each user.

Algorithm 1 RoPADet

```
Input:
$\mathcal{D}$ is dataset and $\mathcal{W}$ is the linear classifier head of RoPADet;
$f$: 1D-CNN+Transformer architecture of RoPADet
Output:
$(\theta^*_{feat}, \theta^*_{profile}, \theta^*_{\mathcal{W}})$: final RoPADet
// Initialization
$\mathcal{W} \leftarrow$ random initialization
$\theta_{feat}, \theta_{profile} \leftarrow$ random initialization
$\mathcal{D}_P = (X_i, y_i, u_i), \mathcal{D}_{NP} = (X_j, y_j) \leftarrow$ select samples
    with and without profile from $\mathcal{D}$
// Pre-train with masked contrastive learning (MCL)
$f_{\theta_{feat}} \leftarrow$ pre-train $f$ on $\mathcal{D}$ with MCL
$f_{\theta_{profile}} \leftarrow$ pre-train $f$ on $\mathcal{D}$ with MCL where negatives
    are sampled from everywhere
// Self-train
$f_{\theta_T} \leftarrow f_{\theta_{profile}}$
for all user $u$ in $\mathcal{D}_{\mathcal{P}}$ do
    $N_u \leftarrow$ number of samples belong to user $u$
    $pred_u \leftarrow \frac{1}{N_u} \Sigma_i^{u_i=u} f_{\theta_T}(X_i)$
end for
$pred^T_{\mathcal{D}_P} \leftarrow pred_{u_i}, \forall X_i \in \mathcal{D}_P$
for $iter = 1$ to $MaxIter$ do
    $pred^T_{\mathcal{D}_{NP}} \leftarrow f_{\theta_T}(X_j), \forall X_j \in \mathcal{D}_{NP}$
    // Train student
    $pred^S_{\mathcal{D}_P} \leftarrow f_{\theta_S}(X_i), \forall X_i \in \mathcal{D}_P$
    $pred^S_{\mathcal{D}_{NP}} \leftarrow f_{\theta_S}(X_j), \forall X_j \in \mathcal{D}_{NP}$
    $Loss_S \leftarrow$ mse-loss$(pred^T_{\mathcal{D}_P}, pred^S_{\mathcal{D}_P})$ + mse-
    loss$(pred^T_{\mathcal{D}_{NP}}, pred^S_{\mathcal{D}_{NP}})$
    $\theta_S \leftarrow \theta_S - \eta_S * \nabla_{\theta_S} Loss_S$
    $\theta_T \leftarrow \theta_S$ // Update teacher
end for
// Update profile extractor
$\theta^*_{profile} \leftarrow \theta^*_S$
$pred_D \leftarrow \{pred^S_{\mathcal{D}_P}, pred^S_{\mathcal{D}_{NP}}\}$
// Fine-tune
$label_D \leftarrow \mathcal{W}(f_{\theta_{feat}}(D) \oplus pred_D)$
$Loss \leftarrow$ cross-entropy-loss$(label_D, y_D)$
$\theta_{feat} \leftarrow \theta_{feat} - \eta * \nabla_{\theta_{feat}} Loss$
$\theta_{\mathcal{W}} \leftarrow \theta_{\mathcal{W}} - \eta * \nabla_{\theta_{\mathcal{W}}} Loss$
return $(\theta^*_{feat}, \theta^*_{profile}, \theta^*_{\mathcal{W}})$
```

However, naturally, most data samples fall into D_{NP}, and the above extracting scheme can not be applied effectively for D_{NP}, since each user provides only 1 sample, so the profile extractor can not extract useful distilled profile information.

To enable learning with profile on D_{NP}, we view this setting as semi-supervised learning. Our main task for this self-training pipeline is profile inference, where student learn to extract profile features provided by teacher model through a mean-squared error loss. Motivated by STraTA [16], we first learn a strong base model f_{θ_T} through an auxiliary task (discriminative masked contrastive learning) and then use it as the initial teacher for self-training. This self-training process converges quickly, in which the first or second iteration's results typically yield the best performances on downstream tasks for RoPADet.

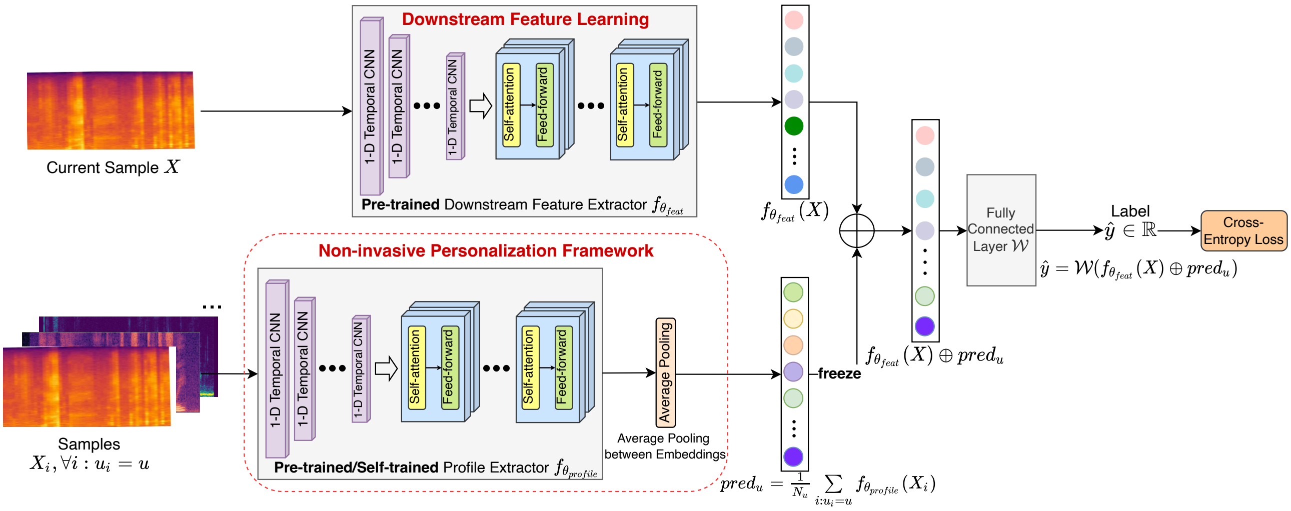
Downstream Feature Learning
1-D Temporal CNN
1-D Temporal CNN
1-D Temporal CNN
Self-attention
Feed-forward
Self-attention
Feed-forward
Current Sample X
Pre-trained Downstream Feature Extractor $f_{\theta_{feat}}$
$f_{\theta_{feat}}(X)$
Fully Connected Layer $\mathcal{W}$
Label
$\hat{y} \in \mathbb{R}$
Cross-Entropy Loss
$\hat{y} = \mathcal{W}(f_{\theta_{feat}}(X) \oplus pred_u)$
Non-invasive Personalization Framework
Samples
$X_i, \forall i : u_i = u$
1-D Temporal CNN
1-D Temporal CNN
1-D Temporal CNN
Self-attention
Feed-forward
Self-attention
Feed-forward
Average Pooling
Average Pooling between Embeddings
Pre-trained/Self-trained Profile Extractor $f_{\theta_{profile}}$
freeze
$f_{\theta_{feat}}(X) \oplus pred_u$
$pred_u = \frac{1}{N_u} \sum_{i:u_i=u} f_{\theta_{profile}}(X_i)$

Figure 2: *Model architectures. Our solution, RoPADet consists of a large pre-training and self-training with personalization, achieving robust performances on respiratory sound illnesses detection tasks.*

3.2. Downstream task training

To learn the downstream task T with the advantage of the personalized profile information, a downstream feature extractor $f_{\theta_{feat}}$ is leveraged with profile features:

$$\hat{y_i} = \mathcal{W}(f_{\theta_{feat}}(X_i) \oplus pred_u), u_i = u \tag{2}$$

where $\mathcal{W}$ denotes a linear classifier head and $\oplus$ denotes the concatenate operation. $f_{\theta_{profile}}$ used to compute $pred_u$ is obtained from the previous pre-training and self-training stages and is frozen in this stage. This is correspondence with RoPADet.

Finally, the model is trained in an end-to-end manner through back-propagation via a cross-entropy loss function:

$$Loss = \sum_i y_i \log(\sigma(\hat{y_i})) + (1 - y_i) \log(1 - \sigma(\hat{y_i})) \tag{3}$$

Algorithm 1 describes the overall training pipeline. As listed in Algorithm 1, from line 5 to 7, we apply a pre-training method that learns discriminative features between samples through contrastive learning. The self-train stage is carried out for learning pseudo profiles, as in line 8 to line 23. The pre-trained or self-trained model, act as the profile extractor, has the same architecture as the conventional model for downstream task, excluding the downstream classification layers. We perform an average pooling operation on the features extracted from samples provided for personalization and used this as the profile feature. The profile feature is concatenated with the classification features before input to a downstream classifier, to perform end-to-end training, demonstrated in line 28-31.

3.3. Model architecture

Following recent advancements, we adapt a transformer-based model for downstream feature learning branch and neural profile extractor of RoPADet, as depicted in Figure 2. Unlike prior works that directly input raw waveform, a spectrum representation of waveform is extracted and fed into front-end 1-D temporal CNN layers before input through the transformer encoder blocks. Our method views each timestep that constitutes frequency signals and nearby information as tokens for the self-attention mechanism. While previous works input waveform, or patches of spectrogram, ours is an efficient way of combining these signals w.r.t the nature of sound data. Moreover, we can leverage the masked pre-training task of transformers as auxiliary task for self-training.

Table 1: *Performances when train on ASound-P (with profile information) and test on ASound-P and ASound-NP (without profile information).*

Exp	Method	Profile-learning	Profile-inference	AUC on ASound-P	AUC on ASound-NP
#1	CNN-based	✗	✗	0.6783	0.3714
#2	Transformer -based	✗	✗	0.8561	0.3822
#3	RoPADet	✓	✗	*0.8693*	*0.3858*
#4	RoPADet	✓	✓	**0.8699**	**0.3964**

Table 2: *Performances of proposed methods on ICBHI dataset (official 60-40 split). ↑ means higher number is better.*

Exp	Method	Additional training data	Personalization	ICBHI↑	Sens.↑	Spec.↑
#1	SoTA#1 [17]	✓	✗	57.3	30.0	**85.6**
#2	SoTA#2 [18]	✓	✗	57.55	*39.15*	75.95
#3	SoTA#3 [19]	✓	✗	*58.29*	37.24	*79.34*
#4	SoTA#4 [20]	✗	✗	53.90	36.36	71.44
#5	SoTA#5 [18]	✗	✗	54.74	33.84	75.35
#6	RoPADet	✗	✓	**58.86**	**40.79**	76.93

4. Results and Discussions

4.1. Datasets

We extensively conduct experiments on 3 real-world respiratory sound datasets: **COUGHVID** (n=7379) [21], **Coswara** (n=4306) [22], and **ICBHI** (n=6898) [23]. Moreover, we collected a crowdsourced dataset, named **ASound** (n=4495), for respiratory illness detection recorded using mobile phones without profile information. Motivated by reliability issues, we conducted a second collection phase, in which if consent is given, we assigned each participant a unique anonymized identifier for personalization. Two additional variants are composed for pre-training and self-training investigations: **ASound-P** includes samples with *profile information* (n=570 by 117 users), and **ASound-NP** *without profile information* (n=1651). We further

Table 3: *Performances when train on ASound, Coswara, and COUGHVID. Trained models are then evaluated on each in-domain or out-domain test sets. We focus on a comparison between our proposed model with or without personalization. Here we denote results with subscript L, M, and H as in the scenario where the effect of covariate shift is low, medium or high, consecutively. The notation ↑ indicates that a higher number is better, whereas ↓ indicates that a lower number is better.*

Exp	**Train Dataset**	**Method**	**Test on ASound**		**Test on Coswara**		**Test on COUGHVID**	
			AUC↑	Brier score↓	AUC↑	Brier score↓	AUC↑	Brier score↓
#1	ASound	Transformer-based	0.9673	**0.0420**	0.7258	0.1722	0.5156	0.1215
#2		RoPADet	**0.9687**	0.0450	**0.7386**$_M$	**0.1631**	**0.5309**$_H$	**0.1210**
#3	Coswara	Transformer-based	0.7619	0.1720	0.8081	0.1912	**0.5031**	0.2481
#4		RoPADet	**0.7691**$_L$	**0.1390**	**0.8102**	**0.1342**	0.5015$_M$	**0.1726**
#5	COUGHVID	Transformer-based	0.4987	0.2819	0.6078	**0.2414**	**0.6878**	**0.2049**
#6		RoPADet	**0.5157**$_H$	**0.2804**	**0.6173**$_H$	0.2487	0.6835	0.2159

Table 4: *Performances when train on ASound-P (obtained from experiments in Table 1), and then evaluate (AUC scores) on each sub-group of users that contributed to ASound-NP. ♢ denotes profiles were only extracted by pre-trained based profile extractor.*

Exp	**Method**	**Age Group**					**Variance between**	**Gender Group**		**Variance between**
		$0-18\uparrow$	$18-23\uparrow$	$23-28\uparrow$	$28-33\uparrow$	$33-100\uparrow$	**Age groups↓**	Male↑	Female↑	**Gender groups↓**
#1	Transformer-based	0.1724	0.3114	0.3537	0.3751	0.535	0.1165	0.4429	0.3736	0.0347
#2	RoPADet♢	0.1776	0.3219	0.3982	**0.3901**	0.5217	0.1124	0.4295	**03915**	**0.0190**
#3	RoPADet	**0.2086**	**0.364**	**0.4324**	0.3641	**0.5486**	**0.1104**	**0.4594**	0.3767	0.0414

validate our personalization strategy on ICBHI as profile information is provided (multiple sound samples per patient).

4.2. Setup

Evaluation metrics. Standard evaluation metrics such as precision, recall, F1, AUC are all reported for future benchmarking purposes. Among these metrics, **AUC score** is our main focused metric, to be consistent with previous works and for evaluation on imbalanced datasets; **ICBHI metric** is used in the ICBHI challenge (i.e., the average of the sensitivity and the specificity); **Brier score** [24] is used to evaluate covariate shift in cross-dataset scenarios.

Experimental settings. Our proposed framework is implemented using fairseq [25]. All experiments are carried out on an Ubuntu Server (20.04 LTS) with 2 RTX 3090 GPUs. For all settings, we pre-trained our model on combined training sets. For finetuning, models are trained for 100 epochs with AdamW optimizer, a learning rate of 1e-4, and a batch size of 64. Unless otherwise specified, all scores are reported based on 5-fold cross-validation (on ASound, Coswara, and COUGHVID) or the average of 5 random seeds (on ICBHI).

4.3. Results

4.3.1. Performances with personalization

In this experiment, we evaluate the effectiveness of personalization strategy through pre-training for profile learning and self-training for profile inference on ASound-P (with profile information) and ASound-NP (with no profile information). Results in Table 1 indicate that while pre-trained profile extractor improves performance on data with, but maintains same performance on data without sufficient profile labels over Transformer-based model that share same architecture but without personalization. When we apply semi-supervised learning through self-training, we observe a large improvement for ASound-NP, up to 2.5% over the baseline CNN. This proves the usefulness of our proposed personalization approach through pre-training and self-training. Moreover, results on ICBHI dataset demonstrated in Table 2 where RoPADet with personalization advance the SoTA for solutions without additional training data by 4.12% in ICBHI score. It also achieved a gap of 0.5% compared to solutions using extra training data, without additional training data and fewer parameters. This also shows the advantages of our framework.

4.3.2. Performances under covariate shift scenarios

Next, we compare models with or without personalization in real-world scenarios where the profile information may not be available. We first train a separate model for each dataset: ASound, Coswara, and COUGHVID. Then, models are evaluated on each test set of the 3 datasets. Results shown in Table 3 demonstrate consistent improvements in cross-dataset evaluations, up to 1.7% in case of train on COUGHVID and then test on ASound, as shown in Exp#5 and Exp#6.

4.3.3. Fair performance evaluation across sub-groups

To ensure fair classification for each sub-group, we evaluate models on each sub-group of users that contributed samples to ASound-NP. Sub-groups are constructed based on the user's age or gender. The results illustrated in Table 4 show that our proposed personalization framework improves performances across all sub-groups and allows for fair classification results between groups with low variances.

5. Conclusion

In this study, we proposed RoPADet, a neural network personalization approach with good improvements through pre-training and self-training. We focused on the problem of covariate shift, which is a significant problem leading to a detrimental performance in AI systems. These consequences might decrease the system's trustworthiness and safety, especially in the healthcare domain. Our personalization technique can be utilized for a variety of patient-related healthcare tasks. It can be applied even if the user is new to the system (i.e., no profile is provided), and as the user continues to use the system, it becomes more adaptable and reliable. Our approach not only helps improve the system's performance but also takes into account users' safety and privacy and aids in gaining their belief. Extensive experiments on various benchmark datasets and tasks from real-world settings confirmed the effectiveness and generalizability of the proposed approach. Future works may be interested in more advanced methods for fusing user profiles and classification features or in weighting user samples, such as exploiting temporal orders.

6. References

[1] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[2] G. T. Frost, G. Theron, and T. Niesler, "TB or not TB? Acoustic cough analysis for tuberculosis classification," in *Proc. Interspeech 2022*, 2022, pp. 2448–2452.

[3] P. A. Pérez-Toro, P. Klumpp, A. Hernandez, T. Arias, P. Lillo, A. Slachevsky, A. M. García, M. Schuster, A. K. Maier, E. Noeth, and J. R. Orozco-Arroyave, "Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings," in *Proc. Interspeech 2022*, 2022, pp. 2483–2487.

[4] J. H. Yoo, H. Jeong, J. Lee, and T.-M. Chung, "Federated learning: Issues in medical application," in *Future Data and Security Engineering*. Springer International Publishing, 2021, pp. 3–22.

[5] A. F. Pacela, "Impedance pneumography—a survey of instrumentation techniques," *Medical and biological engineering*, vol. 4, no. 1, pp. 1–15, 1966.

[6] H. Coppock, L. Jones, I. Kiskin, and B. Schuller, "COVID-19 detection from audio: seven grains of salt," *The Lancet Digital Health*, vol. 3, no. 9, pp. e537–e538, Sep. 2021.

[7] J. Han, T. Xia, D. Spathis *et al.*, "Sounds of COVID-19: exploring realistic performance of audio-based digital testing," *NPJ digital medicine*, vol. 5, no. 1, pp. 1–9, 2022.

[8] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, A. Ruggiero, A. Korhonen, E. Jefferson, E. Ako, G. Langs, G. Gozaliasl, G. Yang, H. Prosch, J. Preller, J. Stanczuk, J. Tang, J. Hofmanninger, J. Babar, L. E. Sánchez, M. Thillai, P. M. Gonzalez, P. Teare, X. Zhu, M. Patel, C. Cafolla, H. Azadbakht, J. Jacob, J. Lowe, K. Zhang, K. Bradley, M. Wassin, M. Holzer, K. Ji, M. D. Ortet, T. Ai, N. Walton, P. Lio, S. Stranks, T. Shadbahr, W. Lin, Y. Zha, Z. Niu, J. H. F. Rudd, E. Sala, and C.-B. S. and, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, Mar. 2021.

[9] M. Mohammadamini, D. Matrouf, J.-F. Bonastre, S. Dowerah, R. Serizel, and D. Jouvet, "Barlow Twins self-supervised learning for robust speaker recognition," in *Proc. Interspeech 2022*, 2022, pp. 4033–4037.

[10] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays *et al.*, "Personalized speech recognition on mobile devices," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5955–5959.

[11] M. R. Awal, R. Cao, R. K.-W. Lee, and S. Mitrović, "Angrybert: Joint learning target and emotion for hate speech detection," in *PAKDD*. Springer, 2021, pp. 701–713.

[12] T. Golany and K. Radinsky, "PGANs: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification," *AAAI*, vol. 33, no. 01, pp. 557–564, Jul. 2019.

[13] A. Sivaraman, S. Kim, and M. Kim, "Personalized Speech Enhancement Through Self-Supervised Data Augmentation and Purification," in *Proc. Interspeech 2021*, 2021, pp. 2676–2680.

[14] T. Dang, J. Han, T. Xia *et al.*, "Exploring Longitudinal Cough, Breath, and Voice Data for COVID-19 Disease Progression Prediction via Sequential Deep Learning: Model Development and Validation," *Journal of Medical Internet Research*, vol. 24, 02 2022.

[15] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *Patterns*, vol. 3, no. 12, p. 100616, 2022.

[16] T. Vu, M.-T. Luong, Q. Le, G. Simon, and M. Iyyer, "STraTA: Self-training with task augmentation for better few-shot learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5715–5731.

[17] L. Pham, D. Ngo, K. Tran, T. Hoang, A. Schindler, and I. McLoughlin, "An Ensemble of Deep Learning Frameworks for Predicting Respiratory Anomalies," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 4595–4598.

[18] I. Moummad and N. Farrugia, "Supervised Contrastive Learning for Respiratory Sound Classification," *arXiv preprint arXiv:2210.16192*, 2022.

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

[20] J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, Y. Li, L. Zhao, and G. Wang, "LungAttn: advanced lung sound classification using attention mechanism with dual TQWT and triple STFT spectrogram," *Physiological Measurement*, vol. 42, no. 10, p. 105006, oct 2021.

[21] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, Jun. 2021.

[22] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, N. R., P. K. Ghosh, and S. Ganapathy, "Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proc. Interspeech 2020*, 2020, pp. 4811–4815.

[23] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, N. Maglaveras, R. P. Paiva, I. Chouvarda, and P. de Carvalho, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological Measurement*, vol. 40, no. 3, p. 035001, Mar. 2019.

[24] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[25] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53.