Interspeech 2022
18-22 September 2022, Incheon, Korea

Example-based Explanations with Adversarial Attacks for Respiratory Sound Analysis

Yi Chang[1,*], *Zhao Ren*[2,*], *Thanh Tam Nguyen*[3], *Wolfgang Nejdl*[2], *Björn W. Schuller*[1,4]

[1]GLAM – Group on Language, Audio, & Music, Imperial College London, United Kingdom
[2]L3S Research Center, Leibniz Universität Hannover, Germany
[3]Griffith University, Australia
[4]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

y.chang20@imperial.ac.uk,zren@l3s.de,t.nguyen19@griffith.edu.au,nejdl@l3s.de,schuller@ieee.org

Abstract

Respiratory sound classification is an important tool for remote screening of respiratory-related diseases such as pneumonia, asthma, and COVID-19. To facilitate the interpretability of classification results, especially ones based on deep learning, many explanation methods have been proposed using prototypes. However, existing explanation techniques often assume that the data is non-biased and the prediction results can be explained by a set of prototypical examples. In this work, we develop a unified example-based explanation method for selecting both representative data (prototypes) and outliers (criticisms). In particular, we propose a novel application of adversarial attacks to generate an explanation spectrum of data instances via an iterative fast gradient sign method. Such unified explanation can avoid over-generalisation and bias by allowing human experts to assess the model mistakes case by case. We performed a wide range of quantitative and qualitative evaluations to show that our approach generates effective and understandable explanation and is robust with many deep learning models.

Index Terms: Respiratory sound analysis, interpretable methods, explainable machine learning

1. Introduction

Respiratory sound classification plays an important role in today's diagnosis systems to assist physicians in identifying adventitious sounds [1]. While respiratory diseases such as COVID-19, bronchial asthma, and chronic obstructive pulmonary disease are affecting more and more the world population [2, 3], such computer-aided auscultation of respiratory sounds provides a remote and non-invasive instrument for early screening of the diseases. Owing to its promising prospect, respiratory classification has been studied intensively [4, 5, 6]. Especially, the success of deep neural networks (DNNs) in various application domains also boosts recent studies of respiratory sounds with better predictive accuracy [1, 2].

However, these advances have also introduced increasing complex and black-box models that are not explainable by nature, i. e., their decision boundaries are difficult to understand [7]. As a result, it is difficult for healthcare practitioners to fully trust the predictions if no explanation is available, especially when many respiratory sound classification results still have modest performance (e. g., the average score of around 50.16 % on the ICBHI 2017 dataset [8, 2]). Existing works tried to mitigate this problem with data augmentation [9] to feed more data to DNNs. Nevertheless, the interpretability of a model is crucial in high-stake domains such as healthcare [10, 11, 12].

Despite many recent advances in explainable artificial intelligence (AI) to mitigate mistakes [13, 14, 15], there are still enormous challenges for explaining respiratory sound classification. Most existing explainable methods focused on attention mechanisms (e. g., identifying parts of the input that most contributed to the final model decision) [16, 17]. Other works focused on example-based explanations using prototypes, which are data instances representative of a target class [18, 19]. Unlike attention mechanisms that do not provide actionable insights of the models, example-based explanations facilitate cognitive human understanding, in particular case-based reasoning [20], as well as vast potential to improve the classification quality via nearest neighbour classifiers [17, 21].

However, existing example-based explainable methods do not consider bias in data (as is often the case in real-world data). In fact, the distribution of a model decision cannot be represented by a set of prototypical examples, but also criticisms – data instances sampled from regions of the input space not well captured by the model. These criticism examples often lie close to the decision boundary of the same target class and often represent model mistakes (e. g., false positives) or out-of-distribution data. Indeed, including criticisms into explanations can avoid over-generalisation and bias by allowing human experts to assess the misclassified examples and outliers [21, 20].

An adversarial attack is a common tool to uncover model mistakes and biases by injecting adversarial perturbations into existing inputs. Such perturbed inputs (i. e., adversarial examples) are indistinguishable from original inputs by a human, yet, they are capable of fooling the model to change the target class [22, 23]. Motivated by adversarial attacks, we develop a unified solution for example-based explanations using prototypes and criticisms. Instead of using adversarial perturbations to change the target class, we consider a novel application of the adversarial perturbations to generate a spectrum of data instances that include both prototypes and criticisms simultaneously. Particularly, we propose an iterative fast gradient sign method (IFGSM) for generating perturbations, which offer a natural way of selecting prototypes and criticisms based on the number of steps of IFGSM.

Our work relates closely to existing works on prototype learning and criticism learning such as MMD-critic [21] and ProtoDash [18]. These works started selecting a set of prototypes first, then separate them into prototypes and criticisms by

* Y. Chang and Z. Ren contribute equally. T. T. Nguyen is the corresponding author. This research was partially funded by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003 and the research projects "IIP-Ecosphere", granted by the German Federal Ministry for Economics and Climate Action (BMWK) via funding code No. 01MK20006A. The code is released at https://github.com/glam-imperial/SoundPrototypeCriticism.

Copyright © 2022 ISCA

10.21437/Interspeech.2022-11355

solving a submodular optimisation problem on data distribution. However, these methods are model-agnostic (explanations are completely independent of the model) or only indirectly capture the discriminative nature of a model via a hidden kernel-based representation of the examples. Unlike these works, we argue that adversarial attacks can be used to unify example-based explanations, i. e., prototypes at one end and criticisms at the other end of the explanation spectrum generated by the IFGSM.

To the best of our knowledge, this is a novel application of adversarial attacks for explainable respiratory sound classification. We propose an interpretable and steerable explanation process for any type of DNNs. In doing so, we overcome the challenges of interpretability in audio data, which often exhibit high-order structures in temporal, spatial, and spectral dimensions. Especially, our approach in unifying prototypes and criticisms via adversarial attacks would benefit users in many ways: (i) the selected prototypes and criticisms can uncover new cases or outliers about the diseases, and (ii) they can be further analysed by post-hoc analysis such as attention tensor learning.

Related Work. Most existing approaches to respiratory sound classification neglect the question *why* certain patients have been classified as a target class. Although there exists many interpretable methods such as regression weights and attention maps [13, 24], they are difficult to validate in sound data, which often exhibit high-order structures in temporal, spatial, and spectral domains. We argue that explanations shall be based on a set of evidential examples, which enable human experts to compare real examples and generalise the problem properties. Some works tried to do so with prototype layers [25, 17, 26], which, however, are synthetic and biased, as the model is forced to focus more on typical examples and ignore extreme cases such as outliers and under-sampled data. Our work is a first attempt to unify example-based explanation for respiratory sound classification by creating an explanation spectrum of real examples to cover normal and abnormal characteristics of data.

2. Methodology

Let us define a dataset $\mathcal{D} = \{(\mathcal{X}, y)\}_{i=1}^{n}$, where $\mathcal{X}$ is the features, y denotes the labels, and n is the number of data samples. In the following, we will firstly give the definitions of prototypes and criticisms, and then explain the adversarial attacks that are used in our study. Finally, the whole process of our example-based explanation will be described.

2.1. Explanation Spectrum

Based on the dataset $\mathcal{D}$, a set of prototypes and criticisms will be searched out to represent the distribution of the model decision.

Prototypes. A strong DNN model often has small intra-class variations and relative large inter-class variations in a classification task [27]. Ideally, the high-level representations learnt by a DNN could be split into N groups according to N classes. The data sample at the centre of each group can be considered as the most representative example for the corresponding class. However, it is challenging to have such an ideal data distribution on real-world data due to a range of reasons, such as noise. Therefore, the high-level representations may be learnt into more than N groups. In this context, *prototypes* are the most representative examples of each group.

Criticisms. Despite multiple groups for each class, the high-level representations of real-world data often have outliers [20]. Although the outliers have only a few samples, they should be still correctly predicted by the model. Nevertheless, it is difficult to search or learn prototypes to represent only a small part of data in each class. In this regard, we call the data samples close to these outliers *criticisms*. Criticisms can be the outliers themselves or generated by the DNN model. In our study, both prototypes and criticisms are searched data examples for example-based explanation.

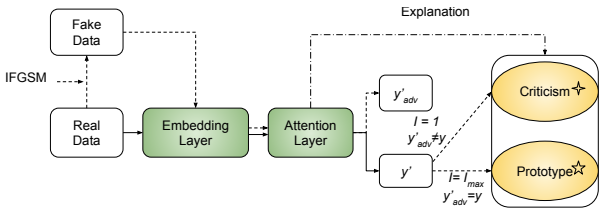


Figure 1: *The explanation pipeline with adversarial attacks. The solid lines are the data flow of real data, the dash lines are for adversarial (i. e., fake) data, and the dash line with dots is the explanation procedure with attention.*

2.2. Adversarial attacks

DNNs have shown their vulnerability to adversarial attacks on acoustic tasks in our prior studies [22, 23]. Due to the data distribution of prototypes and criticisms, we have an assumption that prototypes are the most difficult to be attacked, as they are close to the centre of the class groups; and the criticisms are very easy to be attacked, as they are the outliers. In this study, we search the prototypes and criticisms by the perturbations generated by white-box attack IFGSM [22] which is stronger than FGSM due to its I iterative steps. FGSM calculates the perturbations based on the gradient of loss function. For a data sample $\mathcal{X}_i$, the perturbation is generated by

$$\mathcal{X}_i^{per} = \text{clip}(\epsilon \ \text{sign}(\nabla_{\mathcal{X}_i} \mathcal{L}(\theta, \mathcal{X}_i, y_i'), -\psi, \psi) \qquad (1)$$

where ϵ is a coefficient that controls the difference between the perturbation and the original data, $\mathcal{L}$ is the loss function, θ stands for the model parameters, y_i' is the predicted label of $\mathcal{X}_i$, and the perturbation is clipped into an interval [-ψ,ψ], where ψ is a positive constant. Finally, the adversarial sample is calculated by $\mathcal{X}_i^{adv} = \mathcal{X}_i + \mathcal{X}_i^{per}$. Another benefit of using adversarial attacks is that the deeper the model is, the easier it is to attack [22].

2.3. Example-based explanation process

Inspired by the study of [20], adversarial attacks (i. e., IFGSM) can be an efficient alternative to MMD-critic [21] for prototype and criticism selection. Figure 1 shows our approach of selecting prototypes and criticisms. Specifically, since prototypes are the most representative examples, prototypes should be still correctly classified (i. e., $y'_{adv} = y$) after a certain number of maximum steps I_{max} of FGSM attack. Similarly, because the criticisms are those samples that are not well captured by the model, they should be very vulnerable to the IFGSM attack. Therefore, criticisms tend to be misclassified (i. e., $y'_{adv} \neq y$) after just one step or very few steps of IFGSM. Notably, the prototypes and criticisms are selected from the real data, since adversarial data may lie in a different data distribution, especially for criticisms.

To further verify and explain the selected prototypes and criticisms, it is essential to know which parts of these samples can effectively represent the corresponding distributions. We tackle this challenge by training DNNs with an embedding layer with dilation and an attention layer. Dilated kernels in the embedding layer focus on preserving the size of feature maps, and the

attention layer aims to learn the potential contribution of each unit in the prototypes and criticisms [28, 29, 16].

3. Experiments and Results

3.1. Experimental Settings

Data. To verify our proposed approach, our study is based on the Scientific Challenge database released at the International Conference on Biomedical and Health Informatics (ICBHI) 2017 [8], which is the largest publicly available acoustic database for respiratory sound classification. From seven chest locations of 126 participants, 920 audio waves were recorded with four devices, i. e., one microphone and three stethoscopes. From all recordings, totally 6 898 respiratory cycles were derived. Each respiratory cycle was annotated with one of the four classes, i. e., normal, crackle, wheeze, and both (crackle + wheeze). In the ICBHI challenge, the database was split into a training set (60 %) and a test set (40 %). Similar to our prior study [17], the training set is further split into a train set (70 %) and a validation set (30 %) for optimising the model hyperparameters. Notably, the split procedure is subject-independent to avoid the data from the same person appear in both train and validation set. The data distribution of the database on the four classes and the three datasets is described in Table 1.

Table 1: *The data distribution of the ICBHI database.*

#	Train	Devel	Test	$\sum$
Normal	1 513	550	1 579	3 642
Crackle	616	599	649	1 864
Wheeze	281	220	385	886
Both	131	232	143	506
$\sum$	2 541	1 601	2 756	6 898

Evaluation Metrics. We report the *unweighted average recall* (UAR) as the generic classification benchmark instead of *accuracy*, as UAR can provide fairer evaluation of the models over the four classes than *accuracy* [9, 22] in case of imbalance. It is also common to distinguish abnormal audio samples (i. e., crackles, wheezes, and both) from normal cases. Therefore, the following standard benchmarks are officially used in the ICBHI challenge [8]: *sensitivity* (SE) – the number of true abnormal cases over the total number of abnormal cases, *specificity* (SP) – the ratio of true normal cases over normal cases, and *average score* (AS) – the average of SE and SP.

Preprocessing. All audio recordings are re-sampled into 4 kHz. Since there are confounding noises in most files of the dataset (e. g., handling noise, speech) [8], we apply a fifth order butterworth bandpass filter as the denoising technique. Further, all respiratory cycles with various durations are unified into 4 s. Specifically, 4 s of audio signals are randomly chosen in the training procedure for better flexibility, whereas in the testing process, such length of audio signals are selected in the middle of each respiratory cycle for better performance. With the selected audios, we extract the log Mel spectrograms with a sliding window size of 256, a hop length of 128, and 128 Mel bins.

Model Architecture. In the CNN8 encoder, there are four convolutional blocks with output channel numbers of 64, 128, 256, and 512, respectively. Each of the convolutional blocks is composed of two convolutional layers with the kernel size 3×3, followed by a local max pooling layer with a kernel size of 2×2. For fair comparison, there are also four convolutional blocks in the ResNet encoder. Similarly, the output channel numbers of convolutional blocks are 64, 128, 256, and 512, each of which applies the 'shortcut connections' to add the identity mapping with the outputs of two stacked 3×3 convolutional layers [30]. For the classification, we either apply a global max pooling layer followed by an FC layer or a global attention pooling layer to learn the contribution of each time-frequency bin. For the dilated CNN8 and ResNet, the dilation rates applied for each convolutional block are 1, 2, 4, and 8, where each convolutional layer shares the same dilation rate, besides the one for the 'shortcut connections' [30].

Model training. During training, we utilise the 'Adam' optimiser with an initial learning rate of 0.001 and set the batch size as 16. Specifically, the learning rate is decayed by a factor of 0.9 at every 200-th iteration for stabilisation. All training processes are stopped at the 10 000-th iteration.

3.2. End-to-end Comparison with SOTA Systems

We compare our performance with those of all state-of-the-art (SOTA) approaches on the official test set. Please note that, the studies using a different test set are not comparable. In Table 2, our study is mainly compared with both hand-crafted features on classic machine learning classifiers [31, 4, 32] and time-frequency representations on deep neural networks[33, 34, 17].

Our approach outperforms all state-of-the-art methods on the test set when both AS and UAR are employed for evaluation. In particular, the CNN8 with dilation obtains 52.89 % AS, which is significantly ($p < 0.05$ in a one-tailed z-test) better than the 50.37 % in [17]. Moreover, the ResNet with dilation and attention achieves 46.82 % UAR, which is significantly ($p < 0.001$ in a one-tailed z-test) better than the 36.16 % UAR in [17]. Further, our best models have better performance on SE, which is quite important in clinical practice. Both of our best models have dilated convolutional kernels, indicating dilated kernels can improve performance of local max pooling.

Table 2: *Classification performance [%] comparison with the SOTA approaches on the test set.*

	SE	SP	AS	UAR
MFCC-HMM-GMM [31]	–	–	39.56	–
MFCC-Decision Tree [4]	20.81	78.05	49.43	–
STFT-Wavelet-SVM [32]	–	–	49.86	–
STFT-Wavelet-BiResNet [33]	31.12	69.20	50.16	–
STFT-ResNet-Attention [34]	17.84	81.25	49.55	–
LogMel-CNN8-Prototype [17]	27.78	72.96	50.37	36.16
Ours (CNN8-Dilation)	35.85	69.92	**52.89**	40.26
Ours (ResNet-Dilation-Att)	51.83	50.22	51.02	**46.82**

3.3. Ablation Study

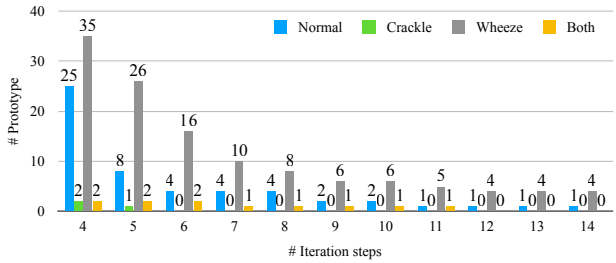
We evaluate the robustness of our prototype and criticism selection approach against various DNN models: CNN8 and ResNet. The performance of dilation and attention on the two CNN models are compared in Table 3. When comparing the UAR and AS values inside the two types of models, the performance of dilation and attention is better than that of the models with local max pooling layers in some cases. When we compare the two types of models, ResNet mostly outperforms CNN8 probably due to the residual blocks in ResNet. Interestingly, high UAR values do not always lead to high AS values. We think this is highly related to the class-imbalance nature.

3.4. Sensitivity Analysis

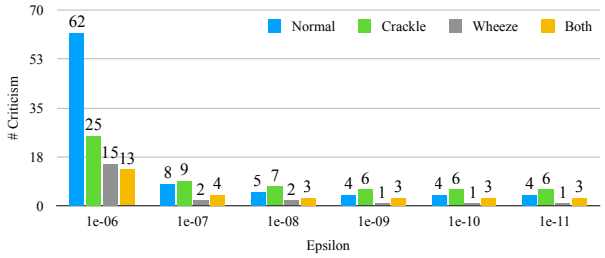
We analyse the number of prototypes under different iteration steps of IFGSM and number of criticisms under different ϵ values

Table 3: *The ablation study of the model performance [%] on residual block, dilation, and attention.*

	Four-class		Binary			
	Dev UAR	Test UAR	Dev AS	Test SE	Test SP	Test AS
CNN8[17]	–	40.36	52.99	39.42	59.72	49.57
CNN8-Att	38.51	42.75	49.56	43.76	49.65	46.70
CNN8-Dila	34.75	40.26	53.27	35.85	69.92	**52.89**
CNN8-Dila-Att	41.55	45.45	50.83	49.62	46.93	48.27
ResNet	41.69	45.33	54.48	43.67	58.01	50.84
ResNet-Att	37.59	43.62	47.66	39.51	62.76	51.13
ResNet-Dila	37.20	43.39	52.65	46.73	44.59	45.66
ResNet-Dila-Att	39.51	**46.82**	52.92	51.83	50.22	51.02



(a) *Number of prototypes with different I.*



(b) *Number of criticisms with different ϵ.*

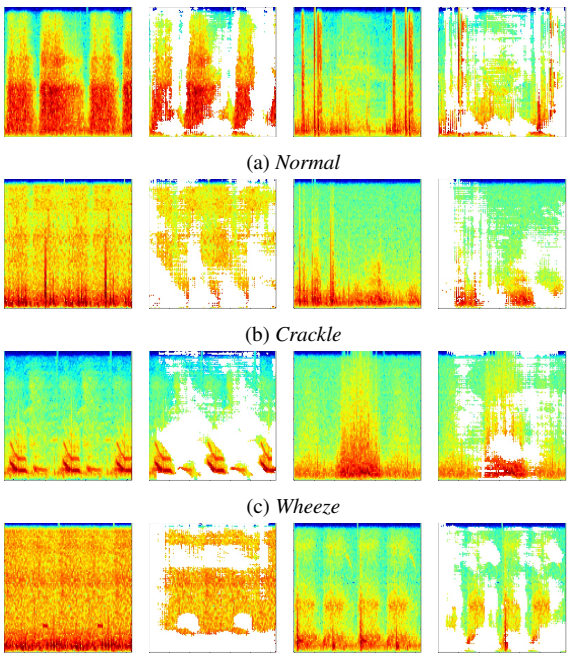
Figure 2: *Analysis of the number of prototypes and criticisms.*

when $I = 1$ in Figure 2 based on the developed dilated ResNet with the attention mechanism. In Figure 2 (a), the number of prototypes for each class is decreasing when IFGSM keeps iterating, indicating the generated adversarial data is stronger with larger I values. In each class, the number of criticisms is also decreasing when ϵ decreases (i. e., adversarial data is becoming more similar to real data). Setting appropriate I and ϵ is helpful for searching effective prototypes.

3.5. Visualisation of Prototypes and Criticisms

The log Mel spectrograms of the selected prototypes and criticisms are depicted in Figure 3. The four prototypes are representative sounds in the four classes. Particularly, the normal prototype sound is regular breathing in Figure 3 (a). As the crackle sounds are explosive, short-duration transient sounds, they can have a big range of magnitude and frequency content [35]. The selected prototype has the consistent characteristics on the duration and frequency Figure 3 (b). Compared to crackle sounds, wheezes have relatively long duration [35]. We can see the wheeze prototype has longer duration than the crackle one for each respiratory cycle Figure 3 (c). The "both" class (Figure 3 (d)) is a combination of crackle and wheeze, therefore, we can only see it is different from the normal one.

When comparing criticisms and prototypes, we can see that



(a) *Normal*

(b) *Crackle*

(c) *Wheeze*

(d) *Both*

Figure 3: *Visualisation of prototypes and criticisms, as well as their contribution parts for respiratory sound classification. X-axis: Time steps, Y-axis: Mel frequency bins. The first column contains the prototypes, the second column shows the contribution parts of prototypes, the third shows the criticisms, and the final one shows the criticisms' contribution parts.*

the typical characteristics of the sounds only appear in part of the whole waveform, especially normal, crackle, and wheeze. We think this is also the reason that the criticisms are easy to be misclassified in a single iteration step of IFGSM. We also project the attention heat maps to the prototypes and criticisms with a threshold at the attentions tensor' middle values. The time-frequency bins of a prototype/criticism are visualised when the corresponding bin in the attention heat map is larger than the threshold. We can see that, the respiratory cycles are preserved in the projections of normal and wheeze prototype sounds. In the projection of the crackle prototype, the non-respiratory part is reserved, probably because the respiratory duration is too short. For the projection of criticisms, the respiratory parts in all four classes are highlighted. Different from the crackle prototype, the non-respiratory part is learnt with low coefficients in the attention heat-map of criticism. We think this is caused by fewer high frequency sounds in the crackle criticism.

4. Conclusion

Existing explainable classification methods do not often consider bias in data. This paper developed a unified example-based explanation for respiratory sound classification by selecting prototypes and criticisms via an iterative fast gradient sign method. Not only applicable for any deep neural networks, our explanations can assist physicians in exploring extreme cases and making informed decisions. Experiments show that our approach can outperform the baselines, and achieve average score of 52.89 % and unweighted average recall of 46.82 %. In future work, we will explore the effect of adversarial attacks by analysing the attention map of adversarial data. We also plan to explore other types of explanations such as counterfactuals [36].

5. References

[1] R. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PLOS ONE*, vol. 12, no. 5, p. e0177926, 2017.

[2] Z. Yang, S. Liu, M. Song, E. Parada-Cabaleiro, and B. W. Schuller, "Adventitious respiratory classification using attentive residual neural networks," in *INTERSPEECH*, 2020, pp. 2912–2916.

[3] Z. Ren, Y. Chang, K. D. Bartl-Pokorny, F. B. Pokorny, and B. W. Schuller, "The acoustic dissection of cough: Diving into machine listening-based COVID-19 analysis and detection," *Journal of Voice*, 2022, 29 pages.

[4] G. Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," in *CBMI*, 2018, pp. 1–6.

[5] V. Ramanarayanan, O. Roesler, M. Neumann, D. Pautler, D. Habberstad, A. Cornish, H. Kothare, V. Murali, J. Liscombe, D. Schnelle-Walka *et al.*, "Toward remote patient monitoring of speech, video, cognitive and respiratory biomarkers using multimodal dialog technology." in *INTERSPEECH*, 2020, pp. 492–493.

[6] O. Rasskazova, C. Mooshammer, and S. Fuchs, "Temporal coordination of articulatory and respiratory events prior to speech initiation." in *INTERSPEECH*, 2019, pp. 884–888.

[7] J. Li, C. Wang, J. Chen, H. Zhang, Y. Dai, L. Wang, L. Wang, and A. K. Nandi, "Explainable cnn with fuzzy tree regularization for respiratory sound analysis," *IEEE Transactions on Fuzzy Systems*, pp. 1–13, 2022.

[8] B. Rocha *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiol. Meas.*, vol. 40, no. 3, 2019.

[9] W. Song, J. Han, and H. Song, "Contrastive embeddind learning method for respiratory sound classification," in *ICASSP*, 2021, pp. 1275–1279.

[10] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *BCB*, 2018, pp. 559–560.

[11] W. Du, L.-P. Morency, J. Cohn, and A. W. Black, "Bag-of-Acoustic-Words for mental health assessment: A deep autoencoding approach," in *INTERSPEECH*, 2019, pp. 1428–1432.

[12] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *MLHC*, 2019, pp. 359–380.

[13] S. Abderrazek, C. Fredouille, A. Ghio, M. Lalain, C. Meunier, and V. Woisard, "Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders — step 1: CNN model-based phone classification," in *INTERSPEECH*, 2020, pp. 2522–2526.

[14] D. Schiller, T. Huber, F. Lingenfelser, M. Dietz, A. Seiderer, and E. André, "Relevance-based feature masking: Improving neural network based whale classification through explainable artificial intelligence," in *INTERSPEECH*, 2019, pp. 2423–2427.

[15] M. Nauta, R. van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *CVPR*, 2021, pp. 14 933–14 943.

[16] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. W. Schuller, "CAA-Net: Conditional atrous CNNs with attention for explainable device-robust acoustic scene classification," *IEEE Trans Multimedia*, vol. 23, pp. 4131–4142, 2020.

[17] Z. Ren, T. T. Nguyen, and W. Nejdl, "Prototype learning for interpretable respiratory sound analysis," in *ICASSP*, 2022, to appear.

[18] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *ICDM*, 2019, pp. 260–269.

[19] Y. Ming, P. Xu, H. Qu, and L. Ren, "Interpretable and steerable sequence learning via prototypes," in *KDD*, 2019, pp. 903–913.

[20] P. Stock and M. Cisse, "ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases," in *ECCV*, 2018, pp. 498–512.

[21] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," *NIPS*, vol. 29, 2016.

[22] Z. Ren, A. Baird, J. Han, Z. Zhang, and B. W. Schuller, "Generating and protecting against adversarial attacks for deep speech-based emotion recognition models," in *ICASSP*, 2020, pp. 7184–7188.

[23] Z. Ren, J. Han, N. Cummins, and B. W. Schuller, "Enhancing transferability of black-box adversarial attacks via lifelong learning for speech emotion recognition models," in *INTERSPEECH*, 2020, pp. 496–500.

[24] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "SAN-M: Memory equipped self-attention for end-to-end speech recognition," in *INTERSPEECH*, 2020, pp. 6–10.

[25] J. Thienpondt, B. Desplanques, and K. Demuynck, "Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization," in *INTERSPEECH*, 2020, pp. 756–760.

[26] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in *CVPR*, 2018, pp. 3474–3482.

[27] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *CVPR*, 2018, pp. 1945–1954.

[28] Z. Ren, Q. Kong, K. Qian, M. Plumbley, and B. W. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in *DCASE*, 2018, pp. 39–43.

[29] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *ICASSP*, 2019, pp. 56–60.

[30] Y. Chang, X. Jing, Z. Ren, and B. W. Schuller, "CovNet: A transfer learning framework for automatic COVID-19 detection from crowd-sourced cough sounds," *Frontiers in Digital Health*, vol. 3, no. 799067, pp. 1–11, 2022.

[31] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *ICBHI*, 2017, pp. 39–43.

[32] G. Serbes, S. Ulukaya, and Y. Kahya, "An automated lung sound preprocessing and classification system based onspectral analysis methods," in *ICBHI*, 2017, pp. 45–49.

[33] Y. Ma *et al.*, "LungBRN: A smart digital stethoscope for detecting respiratory disease using bi-ResNet deep learning algorithm," in *BioCAS*, 2019, pp. 1–4.

[34] Z. Yang *et al.*, "Adventitious respiratory classification using attentive residual neural networks," in *INTERSPEECH*, 2020, pp. 2912–2916.

[35] J. Quintas, G. Campos, and A. Marques, "Multi-algorithm respiratory crackle detection." in *HEALTHINF*, 2013, pp. 239–244.

[36] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *ICML*, 2019, pp. 2376–2384.