INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adventitious Respiratory Classification using Attentive Residual Neural Networks

Zijiang Yang[1], *Shuo Liu*[1], *Meishu Song*[1], *Emilia Parada-Cabaleiro*[1], *Björn W. Schuller*[12]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Group on Language, Audio, & Music, Imperial College London, UK

zijiang.yang@ieee.org

Abstract

Every year, respiratory diseases affect millions of people worldwide, becoming one of the main causes of death in nowadays society. Currently, the COVID-19—known as a novel respiratory illness—has triggered a global health crisis, which has been identified as the greatest challenge of our time since the Second World War. COVID-19 and many other respiratory diseases present often common symptoms, which impairs their early diagnosis; thus, restricting their prevention and treatment. In this regard, in order to encourage a faster and more accurate detection of these kinds of diseases, the automatic identification of respiratory illness through the application of machine learning methods is a very promising area of research aimed to support clinicians. With this in mind, we apply attention-based Convolutional Neural Networks for the recognition of adventitious respiratory cycles on the International Conference on Biomedical Health Informatics 2017 challenge database. Experimental results indicate that the architecture of residual networks with attention mechanism achieves a significant improvement w. r. t. the baseline models.

Index Terms: deep learning, adventitious respiratory classification, residual neural network, attention mechanism

1. Introduction

Respiratory diseases affect more and more people all around the world. For instance, the COVID-19, a new disease identified firstly in Wuhan (China) in 2019 [1], has caused, according to the World Health Organisation (WHO) [1], more than 4 000 000 infections and 290 000 deaths. Respiratory diseases, such as COVID-19, Bronchial Asthma, or Chronic Obstructive Pulmonary Disease, are characterised by very similar symptoms, e. g. , the adventitious breathing, which could be a confounding factor during diagnosis [2]. Due to their serious consequences—particularly in the case of COVID-19, which according to the WHO is a global pandemic—an early and accurate diagnosis of this type of diseases has become crucial.

In this regard, the automatic identification of respiratory diseases, for instance through the classification of adventitious breathing via machine learning, has been successfully performed [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Nevertheless, considering that solutions to prevent and treat diseases as COVID-19 are more urgent than ever, quicker, more accurate, and more convenient solutions for adventitious breathing recognition are still needed. To this end, we present a novel machine learning-based model for respiratory recognition, performed on the International Conference on Biomedical Health Informatics (ICBHI) 2017 challenge database. Our model, focused on a Residual Network (ResNet), i. e. an advanced Convolutional Neural Network (CNN) model optimised with four attention mechanisms, achieves significant improvements w. r. t. the baseline results.

The rest of the manuscript is laid out as follows: in Section 2, the related work is introduced; in Section 3, the methodology is outlined; in Section 4, the experimental setup is described; finally, in Section 5 and Section 6, the results, discussion, and conclusions are given.

2. Related Work

2.1. Respiratory Sound Analysis: ICBHI 2017

Although machine learning methods have been recently applied to automatically identify respiratory diseases, this area of investigation is still impaired by the limited publicly available databases [4]. In this regard, the ICBHI 2017 Database [3], i. e. , a respiratory sound database specially tailored for the automatic identification of respiratory diseases, was presented. The objective of the ICBHI 2017 challenge was the classification of adventitious respiratory cycles in four classes: *Normal*, *Crackle*, *Wheeze*, and *Both* (*Crackle* and *Wheeze*) [3]. The best results were achieved by the SUK and JL teams. The SUK team [4, 6] applied non-dynamic Tunable Q-Factor Wavelet Transform to decompose the input signal into three channels: low resonance, high resonance, and the residue part; then, the fusion of the features extracted through Short Time Fourier Transform (STFT) and through Tunable Wavelet Transform from the three channels, was fed into a Support Vector Machine (SVM) classifier, achieving 49.86 % of the official score (the average of sensitivity and specificity). The JL team [4, 5] utilised two methods of spectral subtraction in order to suppress the stationary noise in the background—thus, decreasing the influence of noise in the signal. After signal preprocessing, Mel-frequency Cepstral Coefficients (MFCCs) were extracted as features and the combination of a Hidden Markov Model (HMM) and a Gaussian Mixture Model (GMM) was considered for classification, achieving 39.56 % of the score.

After the challenge, further work on the same task was performed, showing that MFCCs and spectral low-level features with a boosted decision tree with one leave per class were particularly suitable, achieving 49.43 % of the score [8]. However, the best performance so far was achieved through a model based on ResNet [15], which has considered, instead of the classic ResNet structure, a bi-ResNet model, i. e. two individual ResNet models were trained by STFT and wavelet features, respectively, before being concatenated into one dense layer. The performance of this approach achieved 50.16 % of the score [12].

Additionally, although originally a four-class classification task was presented by the challenge, other researches on the

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019

http://dx.doi.org/10.21437/Interspeech.2020-2790

ICBHI 2017 Database have also been proposed. For instance, a three-class classification problem, considering adventitious respiratory cycles: *Normal*, *Crackle*, and *Wheeze* [7]; or the classification of respiratory diseases: *Healthy*, *Chronic*, and *Non-chronic* [10, 13].

In addition, other works exist in a more broader context of breathing analysis such as featured in this year's Interspeech 2020 Computational Paralinguistic Challenge's Breathing Sub-Challenge [16].

2.2. Attention Mechanisms

The efficiency of CNNs has been proven during the last years [17, 18, 19, 20, 21], showing a particularly accurate performance in image recognition [22]. In order to improve CNN performance, a novel mechanism called 'attention', which emulates the way humans understand images, was proposed. At first sight, humans tend to focus on areas with salient features e. g. , a house, rather than processing the whole scene, e. g. , the whole painting [22]. Similarly, the attention mechanism leads CNN models to focus on the most relevant areas of an image—by this increasing their understanding—which usually optimises CNN performance [23]. Several attention-based modules have been proposed, showing that this mechanism can optimise the abilities of neural networks on feature recognition [24, 25, 26, 27].

The Squeeze-and-Excitation (SE) [28, 29] block is a classic channel-wised attention mechanism. The SE block 'squeezes' feature maps to acquire the channel descriptor, then applies 'excitation' on the descriptor to learn the relation among all channels and obtains the weights of every channel, where the more important channel gets a larger weight. The SE block can enhance the sensitivity of the model on features by suppressing the contribution of less useful channels [28].

Another attention mechanism is the spatial attention block [22, 30], which in order to reduce the number of channels to 1, applies max pooling or average pooling on the original feature map, and generates the spatial weights. As its name has it, the spatial attention module makes efforts on the spatial relation to localise the most important area of the image.

Finally, another attention mechanism called component attention has also been proposed [31], showing an excellent performance, e. g. , on affective computing tasks [32]. Differently to other attention mechanisms, the component attention module extracts component-wise attention vectors from the original feature map instead of channel-wise or spatial attention vectors. Therefore, the component attention mechanism focuses on observing the relation among all components of the input vector.

3. Methodology

3.1. Feature Sets

After evaluating a variety of feature extraction methods from the literature, such as MFCCs, STFT, or wavelet features [5, 6, 8, 12], the STFT and wavelet features showed to be those with the best performance. Furthermore, since one of the advantages of a CNN is its excellent ability to focus on the adjacent area of every unit in the time-frequency domain [33], CNN based models especially benefit from the spectrogram. With this in mind, we consider that the STFT spectrogram is the most appropriate feature to be taken into account. Concerning the parameters of the STFT, n_fft was set to 100, and the window length was determined to be 100 ms with a 40 ms hop length.

3.2. Residual Network (ResNet-18)

The proposed model is based on ResNet-18 [15, 34], since the result of model selection shows that a deeper ResNet (such as ResNet-50) caused severe overfitting. ResNet-18 contains 4 residual blocks, and every residual block has two residual modules. A residual module includes three main components: two convolutional layers and an identity **x** one. Before the input is fed into the residual module, the input is saved and marked as the identity **x**. At the end of the module, the sum of the original output and **x** is regarded as the output and propagated in the network [15]. ResNet-18 was applied in experiments with and without different attention mechanisms.

3.3. Attention Blocks

As pointed out in Section 2.2, both the SE block and the spatial attention block, were adopted separately in every residual block. In addition, the SE block and the spatial attention block considered together, as well as the component attention block, were also taken into account. The structure of the four considered attention blocks is shown in Fig. 1.

3.3.1. Squeeze-and-Excitation Block

The Squeeze-and-Excitation (SE) block includes two parts of processing: *Squeeze* and *Excitation*. In the *Squeeze* procedure, spatial information of the output vector is squeezed into a channel descriptor with the size of $1 \times 1 \times C$, where C stands for the number of channels [28]. Afterwards, the channel descriptor is sent to the *Excitation* module, including one dense layer with C/r hidden units and another dense layer with C hidden units, in which r represents reduction ratio. After applying a sigmoid activation function, the channel-wise attention vector is obtained. An SE block is defined as follows:

$$\mathbf{F_{SE}} = \mathbf{A_{SE}}(\mathbf{F}) \bigotimes \mathbf{F}, \tag{1}$$

where **F** represents its input feature maps. The attention $\mathbf{A_{SE}}$ is learnt via an average pooling layer across all locations (time and frequency axes) and two dense layers:

$$\mathbf{A_{SE}}(\mathbf{F}) = \sigma(f_{C \to \frac{C}{r} \to C}(\text{avg\_pool}(\mathbf{F}))), \tag{2}$$

where $f_{C \to \frac{C}{r} \to C}$ indicates the process of converting the average-pooled output from the length of C to C/r first, and back to C again. σ stands for the sigmoid function, which limits the learnt attention values to the range of 0 and 1.

3.3.2. Spatial Attention Block

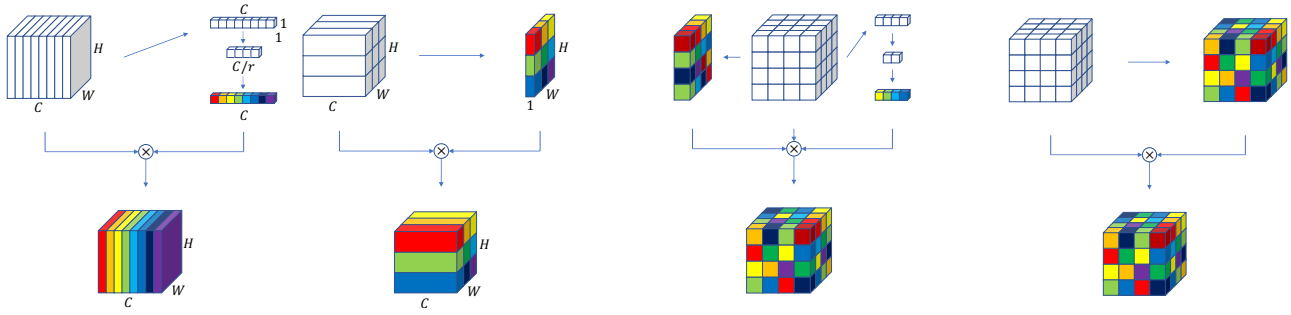
Unlike the SE block, the spatial attention block generates a spatial descriptor rather than a channel descriptor. The size of the output vector is compressed from $H \times W \times C$ into the spatial-wise attention vector, with the size of $H \times W \times 1$ [22]. In the end, the output vector is multiplied by the attention vector, and the final output is sent to the following network components. The spatial attention is computed as:

$$\mathbf{F_{SP}} = \mathbf{A_{SP}}(\mathbf{F}) \bigotimes \mathbf{F}, \tag{3}$$

where $\mathbf{A_{SP}}$ stands for the spatial attention, and

$$\mathbf{A_{SP}}(\mathbf{F}) = \sigma(\text{Conv}_{1\times1\times1}(\mathbf{F})), \tag{4}$$

in which $\text{Conv}_{1\times1\times1}$ reveals the 1×1 convolutional layer with a single output channel.



(a) *Squeeze-and-Excitation Block* (b) *Spatial Attention Block* (c) *SE + Spatial Attention Block* (d) *Component Attention Block*

Figure 1: *Structure of the considered attention mechanisms: the Squeeze-and-Excitation (SE) block, the spatial attention block, the SE + spatial attention block, and the component attention block.*

3.3.3. SE + Spatial Attention Block

Besides, in order to explore the performance of the attention on both channel and spatial domain at the same time, the ResNet model with the SE and the spatial attention blocks together, was also tested. Attentions on channel and spatial were computed and multiplied by the output vector together to generate the final output. The computation is:

$$\mathbf{F_{SP}} = \mathbf{A_{SE}}(\mathbf{F}) \bigotimes \mathbf{A_{SP}}(\mathbf{F}) \bigotimes \mathbf{F}, \tag{5}$$

where $\mathbf{A_{SE}}(\mathbf{F})$ and $\mathbf{A_{SP}}(\mathbf{F})$ were defined in Eq. 2 and 4.

3.3.4. Component Attention Block

Additionally, the component attention block, which is given by a slight modification of the spatial attention block, i. e. , the output of the convolutional layer was modified from 1 to C, was adopted in ResNet-18. In other words, an attention with a size of $H \times W \times C$ was generated instead of $H \times W \times 1$. The computation of the component attention is defined as:

$$\mathbf{F_{CO}} = \mathbf{A_{CO}}(\mathbf{F}) \bigotimes \mathbf{F}, \tag{6}$$

where $\mathbf{A_{CO}}$ represnts the component attention,

$$\mathbf{A_{CO}}(\mathbf{F}) = \sigma(\mathrm{Conv}_{1\times1\times C}(\mathbf{F})), \tag{7}$$

where $\mathrm{Conv}_{1\times1\times C}$ means the 1×1 convolutional layer with C output channels.

4. Experimental Setup

4.1. Database

To investigate the performance of attention mechanisms on respiratory sound classification, we considered the ICBHI 2017 Database [3]. The ICBHI 2017 Database contains 920 recordings obtained from seven chest locations. In total, 6 898 respiratory cycles were recorded: 3 642 for *Normal*, 1 864 for *Crackle*, 886 for *Wheeze*, and 506 for *Both*, i. e. *Crackle* plus *Wheeze*. Four different devices, i. e. three stethoscopes and one microphone, were used to gather the data—a process, in which two medical centres were involved. A subject-independent partitioning into a training set (60 %) and a test set (40 %), performed according to a fixed distribution given by the organiser of the challenge, was considered (cf. Table 1). From a statistical perspective, the duration of every respiratory cycle varies from a minimum of 0.20 *sec.* to a maximum of 16.16 *sec.* ($mean = 2.70$ *sec.*, $median = 2.54$ *sec.*, $std = 1.17$ *sec.*).

Table 1: *Details of the distribution of ICBHI 2017 Database*

	# Training Set	# Test Set	# Σ
participants	79	49	128
recordings	539	381	920
normal	2 063	1 579	3 642
crackle	1 215	649	1 864
wheeze	501	385	886
both	363	143	506
Σ	4 142	2 756	6 898

4.2. Preprocessing

Before to train the model, several preprocessing procedures were carried out. Due to the variety of devices taken into account for the recording process [3], the sample rate for some recordings was 4 kHz while for others it was 44.1 kHz; thus, the sample rate of all recordings was down-sampled to 4 kHz. And a 5-th Butterworth bandpass filter (100 - 2000 Hz) was applied to prevent meaningless information [5, 12]. Additionally, due to the unbalanced length of the instances, to keep all spectrograms at the same size, a duration of 2.5 sec. was set for every instance. For this, the spectrograms of instances longer than 2.5 sec. were trimmed into the same duration, while for short instances, the missing part was padded with 0. In the training set, instances were randomly trimmed, while in the test set—to guarantee comparability across the considered approaches—instances were trimmed according to a fixed start position.

Moreover, in order to achieve more convincing results, another subject-independently partitioning, considering 70 % of the samples for the training set and 30 % for the validation set, was taken into account from the original training set, while the original test set was only used for evaluation.

4.3. Evaluation Metric

In this paper, the evaluation metric of the ICBHI 2017 challenge was applied for a better comparison with other works. According to the challenge [3], the metric Average Score (AS) was defined as the average of Sensitivity (SE) and Specificity (SP). SE, SP, and AS are computed as:

$$SE = \frac{C_T + W_T + B_T}{C + W + B}, \tag{8}$$

where C_T, W_T, B_T stand for the number of correctly classified *Crackle*, *Wheeze*, *Both* instances, and C, W, B stand for the

Table 2: *Experimental results. At the top, outcomes from previous work; at the bottom, outcomes from the baseline and attentive ResNet-18 models as proposed here. The evaluation metrics average score (AS), sensitivity (SE), and specificity (SP), are considered.*

	Validation Set (%)	**Test Set (%)**		
	AS	SE	SP	**AS**
MFCCs + HMM + GMM [5]	—	—	—	39.56
STFT + wavelet features + SVM [6]	—	—	—	49.86
MFCCs + low-level features + Boosted Decision Tree [8]	—	20.81	78.05	49.43
STFT + wavelet features + bi-ResNet [12]	—	31.12	69.20	**50.16**
ResNet-18	51.19	19.46	77.96	48.71
ResNet-18 + Squeeze-and-Excitation (SE) Block	53.28	15.04	78.85	46.94
ResNet-18 + Spatial Attention Block	**54.99**	11.38	82.23	46.86
ResNet-18 + SE Block + Spatial Attention Block	54.14	17.84	81.25	**49.55**
ResNet-18 + Component Attention Block	54.24	20.65	77.83	49.24

total number of *Crackle*, *Wheeze*, *Both* instances, respectively,

$$SP = \frac{N_T}{N}, \tag{9}$$

where N_T and N stand for the number of correctly classified and all *Normal* instances, respectively, and

$$AS = \frac{SE + SP}{2}. \tag{10}$$

5. Results and Discussion

The experimental results, including those methods presented in previous works (cf. Section 2.1), and those proposed by us, are given in Table 2. For the considered models, i. e. , ResNet-18 without attention block (our baseline), ResNet-18 with the Squeeze-and-Excitation (SE) block, ResNet-18 with the spatial attention block, ResNet-18 with the SE and the spatial attention blocks, and ResNet-18 with the component attention Block (cf. Section 3.3). Results on the validation and the test sets are presented; for the previous work, since results on the validation set are not available, only those for the test set are given.

For all the considered approaches, our results for the test set (AS $\leq 49.55\,\%$) display a detriment with respect to the validation set (AS $\geq 51.19\,\%$); cf. Table 2. Although this cannot be observed in the Table 2 for the previous research—no concrete results for the validation set were reported in the evaluated works—it has been indicated in those works, in any case, that the scores achieved on the validation set (although not reported), were much lower than those from the test set [4]. This tendency might be explained on the one side by the fact that four devices, i. e. three stethoscopes and one microphone, as well as seven chest locations, were utilised in data recording, which introduced a variety of noises of diverse intensity all over the database. On the other side, this tendency might be also due to the unbalanced original distribution of the training and the test sets. Indeed, although the organisers of the ICBHI 2017 challenge performed a subject-independent partitioning, for the training set, instances collected by all the devices were considered, while for the test set, only those collected by three devices were taken into account [3].

Our baseline, i. e. , the results from the ResNet-18 without any attention mechanism, was outweighed by all the considered attentive models on both the validation and test sets, which indicates that attention mechanisms truly optimise ResNet-18 on adventitious respiratory classification. On the validation set, the best score was achieved by the spatial attentive model, which achieved AS $= 54.99\,\%$ w. r. t. AS $= 51.19\,\%$ given by the baseline. Differently, on the test set, the baseline (AS $= 48.71\,\%$), was only outperformed by the ResNet-18 with SE and Spatial Attention blocks (AS $= 49.55\,\%$), which reached the level of the state-of-art. This might be due to the fact that using SE and the spatial attention in a unified CNN enables its retrieval of salient acoustic characteristics across channels and locations, concurrently. The reason why attentive models perform better than the baseline is probably that each evaluated class presents specific acoustic characteristics on the frequency and time domains, i. e. , *Wheeze* is always longer than $100\,ms$, while *Crackle* is always shorter than $20\,ms$. In this regard, attention blocks give larger weights to these particular characteristics; thus, encouraging the classification performance of the model.

Although our best result was slightly outperformed by the approaches based on a bi-Resnet with STFT and wavelet features [12], our models replace a complex and deep network architecture, i. e. two 34-layers ResNet [12], by utilising attention mechanisms with an 18-layers ResNet, which alleviates the request on computational complexity, reducing also time and storage resources. Finally, all models, i. e. those proposed by us as well as those considered in previous research, achieved excellent results on SP but very low on SE, which might be due to a not sufficiently clear discrimination of the spectrograms among the three adventitious respiratory classes.

6. Conclusion

The application of attention mechanisms on adventitious respiratory classification was investigated, showing that these can contribute to improve the performance of residual neural networks. Performance was elevated from $51.19\,\%$ to $54.14\,\%$ average score on the validation set, and from $48.71\,\%$ to $49.55\,\%$ on the official test set. A Resnet-18 with Squeeze-and-Excitation and spatial attention blocks achieved the best results, being equivalent to other state-of-art works. In the future, we plan to utilise different feature sets, such as wavelet or Mel-spectrograms, and data augmentation as a strategy to deal with the unbalanced database. Our work indicates that attention mechanisms are very promising for adventitious respiratory classification.

7. Acknowledgements

We offer our deepest condolences to people who lost their family and friends because of COVID-19 and the highest respect to all clinicians, nurses as well as researchers who are fighting against COVID-19 for all human beings. We further acknowledge funding from the KIrun project (grant no. FKZ-16KN069402) supported by a German BMWi ZIM grant.

8. References

[1] N. Chen, M. Zhou, X. Dong, J. Qu, G. Fengyun, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: A descriptive study," *The Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[2] V. Cukic, V. Lovre, D. Dragisic, and A. Ustamujic, "Asthma and chronic obstructive pulmonary disease (COPD)–differences and similarities," *Materia Socio-Medica*, vol. 24, no. 2, p. 100, 2012.

[3] B. M. Rocha, D. Filos, L. Mendes, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *Proc. ICBHI: Precision Medicine Powered by pHealth and Connected Health*, Thesaloniki, Greece, 2017, pp. 33–37.

[4] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological Measurement*, vol. 40, no. 3, 2019, 28 pages.

[5] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *Proc. ICBHI: Precision Medicine Powered by pHealth and Connected Health*, Thesaloniki, Greece, 2017, pp. 39–43.

[6] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based onspectral analysis methods," in *Proc. ICBHI: Precision Medicine Powered by pHealth and Connected Health*, Thesaloniki, Greece, 2017, pp. 45–49.

[7] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32 845–32 852, 2019.

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

[9] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. ICANN*, Rhodes, Greece, 2018, pp. 208–217.

[10] D. Perna, "Convolutional neural networks learning from respiratory data," in *Proc. BIBM*, Madrid, Spain, 2018, pp. 2109–2113.

[11] S. Ntalampiras and I. Potamitis, "Classification of sounds indicative of respiratory diseases," in *Proc. EANN*, Crete, Greece, 2019, pp. 93–103.

[12] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "LungBRN: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *Proc. BioCAS*, Nara, Japan, 2019, pp. 1–4.

[13] M. T. García-Ordás, J. A. Benítez-Andrades, I. García-Rodríguez, C. Benavides, and H. Alaiz-Moretón, "Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data," *Sensors*, vol. 20, no. 4, 2020, 16 pages.

[14] Z. Neili, M. Fezari, and R. Abdeghani, "Analysis of acoustic parameters from respiratory signal in copd and pneumonia patients," in *Proc. SIVA*, Guelma, Algeria, 2018, pp. 1–4.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[16] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Proceedings INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association*, ISCA. Shanghai, China: ISCA, September 2020, 5 pages, to appear.

[17] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.

[18] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 113–117.

[19] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 5861–5865.

[20] Z. Ren, K. Qian, Y. Wang, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018.

[21] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 849–853.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, 2018, pp. 3–19.

[23] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 2956–2964.

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

[25] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 221–225.

[26] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019.

[27] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2020.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[29] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2019.

[30] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 5659–5667.

[31] A. Das, J. Li, R. Zhao, and Y. Gong, "Advancing connectionist temporal classification with attention modeling," in *Proc. ICASSP*, Calgary, AB, Canada, 2018, pp. 4769–4773.

[32] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 206–210.

[33] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 3768–3772.

[34] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. NIPS*, Montreal, QC, Canada, 2018, pp. 6389–6399.