A benchmarking study of deep learning techniques applied for breath analysis

Vincenzo Dentamaro[1], Paolo Giglio[1], Donato Impedovo[1], Luigi A. Moretti[2], Giuseppe Pirlo[1], Elena Sblendorio[3]

[1] University of Bari Aldo Moro, Via Orabona 4, 70121, Bari Italy
[2] University of the West of England (UWE) - Coldharbour Ln, Stoke Gifford, Bristol BS16 1QY, UK3
[3] Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy

Abstract

In Machine Learning, new architectures are continually proposed, making difficult to evaluate which configurations better fit specific fields and tasks. The most reliable way to overcome this issue is to test them using the same data and parameters. In this work, five state of art deep neural network architectures has been performed in a promising field of health technology: the breath analysis. In particular it is reported that standard convolutional neural networks exploiting inductive bias, do not perform as well as the AUCO ResNet, an architecture designed for audio classification. In addition, the Vision Transformer model need lots of data to learn patterns showing the limitation of this technique even when transfer learning is performed.

Keywords
Breath analysis, transfer learning, AUCO ResNet, Mel Spectrogram, benchmark, ViT

1. Introduction

Three respiratory diseases were entrenched in the top 10 causes of death in the world. The chronic obstructive pulmonary disease (COPD) alone kills 3.2 million people every year [1] .COVID19 pandemic has remined us how vulnerable we are as a community. However, it has also been the opportunity to demonstrate the potential of digital solutions as a fundamental support for the healthcare system. Improving early detection is considered an essential step to reduce the burden of respiratory diseases. [1] However, accessible, affordable, and reliable tools must be designed for behavioral biometric analysis. [2] Machine Learning (ML) and Smart Sensors can be valuable resources in this matter, but the research must consider the practical implementations, to avoid wasting time and resources developing solutions which performance are not reproducible in the clinical environment. [3]
Benchmarks are of paramount importance when it comes to test different models under the same conditions. This allows a fair comparison of their accuracies and provides insights into the strengths and weaknesses of each model. This study aims to provide an overview of five architectures applied to a database of breath sounds. Mostly of them have been originally developed for computer vision tasks, thus a filter converting sounds in images (i.e spectrograms) has been applied. Two different tasks have been performed, to test the models both in a binary and multiclass classification.
The main research question is to understand which model is best suited to perform audio breath analysis when networks are trained in a transfer learning fashion.
The work is organized as follows: Section II introduces the state-of-the-art review. Section III Material describes the datasets used for training and transfer learning as well as the pre-processing applied. Section IV Methods presents the architectures used as well as the experimental setup. Section V

IEEESDS'23: Data Science Techniques for Datasets on Mental and Neurodegenerative Disorders, June 22, 2023, Zürich, Switzerland
vincenzo.dentamaro@uniba.it (V. Dentamaro); elena.sblendorio@students.uniroma2.eu (Elena Sblendorio);
0000-0003-1148-332X (V. Dentamaro);

sketches the results. In Section VI there is the discussion of results. While Section VII contains conclusions and future research directions.

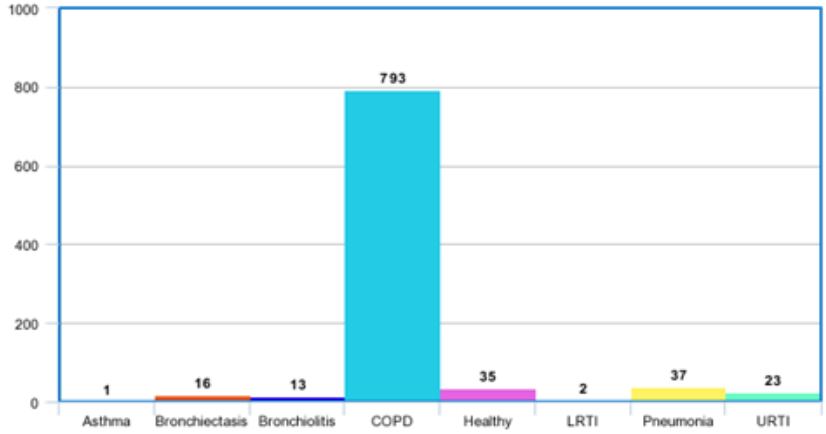

Figure 1 diagnosed diseases distribution in the Respiratory Sound Dataset

2. State of the art review

The database ICBHI 2017 has been used by various authors all in different conditions. Almost all authors since 2019 have been using deep neural network architectures, which have shown promising results. This brief literature review is focused on deep neural networks architectures already used on the ICBHI 2017 dataset. In particular in work [4] authors finetuned a pre-trained ResNet architecture on the ICBHI 2017 and multi-channel lung sound datasets for performing binary healthy/unhealthy classification reaching 87.59 of F1-Score. Authors in [5] addressed the limited size dataset issue using supervised contrastive learning. This technique relies on respiration cycle annotations as well as spectrogram frequency and temporal masking in order to generate augmented samples for representation learning with a contrastive loss. The reached accuracy is 0.759.

The Respiratory Sound Classification Network (ARSC-Net) [6] is a network designed for accurate respiratory sound classification. It combines residual blocks with channel-spatial attention to extract and classify two types of features from adventitious respiratory sounds: Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrogram. The two types of features are processed in parallel through encoder paths with residual attention to obtain a feature representation, which is then merged in a channel-spatial attention module to focus adaptively on important features in both the channel and spatial domains. The channel-spatial attention improves the feature representation by exploring inter-channel relationships in the spectrums using channel attention and generating inter-spatial correlation mapping through serial spatial attention. The reached accuracy is 80% in binary classification (healthy/unhealthy) but no inter-subjects separation scheme was used. In [7] authors come up with a new way to enhance the classification of respiratory sounds performing data augmentation. The approach involves changing and moving around the input data. The accuracy reached is about 0.704.

The LungBRN [8] and its evolution, the LungRN+NL [9] have been designed especially for the ICBH2017 challenge. The LungBRN architecture is an advanced bi-ResNet deep learning architecture, that utilizes STFT and wavelet feature extraction to enhance accuracy, while the LungRN+NL has incorporated the non-local block in the ResNet architecture. In addition, to address the imbalance problem, authors also added data augmentation to increase the accuracies. Respectively the LungBRN and LungRN+NL achieved 0.692 and 0.632 of accuracy.

As it is possible to observe from this small literature review, the majority of works are based on the ResNet deep learning architecture trained from scratch or in transfer learning fashion. In addition, it is possible to observe that some work used data augmentation. As pointed out in [10] data augmentation has caused some confusion in identifying correct patterns for certain classes when

initial data is not representative. Additionally, data augmentation decreases reproducibility and raises ethical concerns in the medical field [11].

For these reasons, it has been decided to perform a benchmark of the most famous deep neural networks architectures in computer vision as well as novel architectures such as the Vision Transformer and AUCO ResNet all trained on exactly same data and in exactly same conditions. Additionally, the methods used in this work do not make use of any data augmentation technique increasing reproducibility of the results.

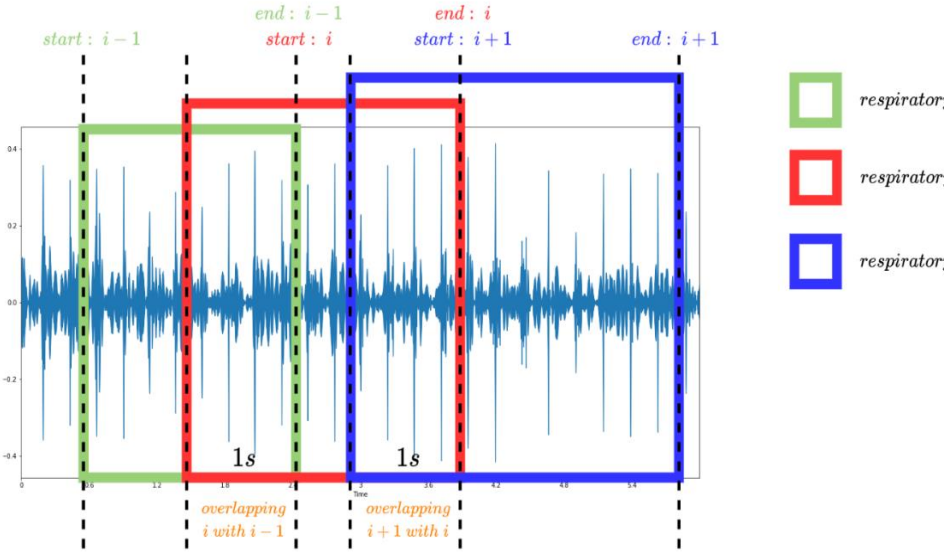

Figure 2 Overlapping between respiratory cycles

3. Material

3.1. Respiratory Sound Dataset

Compiled in 2017 at the International Conference on Biomedical Health Informatics (ICBHI 2017), the Respiratory Sound Dataset [1] was built to provide reliable data for comparing different automatic audio analysis. The database consists of a total of 5.5 hours of recordings collected in 920 annotated audio samples from 126 subjects ranging from young to old ages. 6898 respiratory cycles are included, of which 1864 contain crackles, 886 wheezes, and 506 both crackles and wheezes. The recordings were collected using heterogeneous equipment and their duration ranged from 7.85 s to 86.2 s. Data include clean breathing sounds and noisy recordings that simulate real-life conditions.

Unfortunately, this database is not free from limitations. Not only some subjects have been used to collect several audio samples (up to 60, making 6.5% of the entire sample size belonging to a single subject), but the file samples are also strongly unbalanced both in the distribution of the seven labelled diseases and between the number of healthy (no disease) and un-healthy subjects which respectively are 35 (3.8%) and 885 (96.2%) (separation used for the Binary Task). These issues, shown in Figure 1, have been taken in account and faced as later discuss in the present work. Given the lack of data, instead of considering each disease individually, it has been preferred to group all diseases into non-chronic diseases (i.e. Lower Respiratory Tract Infections (LRTI), Upper Respiratory Tract Infections (URTI), Pneumonia, e Bronchiolitis, for a total of 75 samples) and chronic diseases (i.e. COPD, Bronchiectasis e Asthma, including the remaining 810 samples), in addition of the third group of healthy subjects (used for the Multiclass Task). Acute (i.e. temporary) diseases sometimes (e.g. when under- or mis-treated) evolve in Chronic ones, thus any tool in support of medical decision-making is highly welcomed.

3.2. UrbanSound8K Dataset

UrbanSound8K Dataset [2] is one of the biggest and more used of its kind. It collects 8732 audio tracks, up to 4 seconds in length each, recorded from a urban environment and labelled in 10 quite well balanced classes. The files are stored in .wav format, as in Respiratory Sound Database, organized in 10 folds for an easier results comparison between different ML models. Given its remarkable dimension, this dataset was used in the transfer learning procedure, as later explained.

3.3. Preprocessing on Respiratory Sound Dataset

Since 94% of the Respiratory Sound Dataset was recorded at 44,100 Hz, it has been chosen to sample all audio tracks within that sampling rate. Moreover, since the audio tracks have different durations, as testify by the max and average length of 86.2 and 21.5 seconds respectively, each file audio has been divided into its respiratory cycles, avoiding cutting all audios to a prefixed length with a massive loss of data.

As it is known from signal processing theory, there are some chunks of audio yielding more information, resulting to be relevant for training the algorithms. Inspired by the conclusions of a study focused on a similar task, but with a different respiratory disease (COVID19)[4] it has been supposed that, also in this scenario, relevant information may be present between the transition from one respiratory cycle to another. For this reason, each respiratory cycle has been segmented to ensure a certain data overlapping (1s) with the next one (Figure 2). Some adjustments were needed when dealing with cycles shorter than 1s, in those cases the gap was filled adding enough zeros to compensate (padding technique).

Given the average length of 3s per respiratory cycle, the cycles longer than 4s (1s overlapping included) have been resized to be 4s long, while the shorter ones have been extended using padding. After this procedure, each respiratory cycle corresponds to a (4*4,4100 =) 176,400-dimensional vector of amplitudes.

To reduce the chance of biased results, it was performed the User-based Dataset splitting, which ensures that each subject, from which the data have been collected, belongs to the training or the testing set only [3] as it will be detailed later.

4. Methods

Two classification tasks have been performed in this work to compare different architectures. The first one aimed to recognised healthy vs unhealthy subjects, while the second one is a multi-class classification test to recognize the specific disease (if healthy or unhealthy with chronic disease or unhealthy with acute disease). The tested architectures are:

AUCO ResNet [4], the Auditory Cortex ResNet (AUCO ResNet) is a biologically inspired deep neural network especially designed for sound classification. It is built on the intuition that mammals have evolved the sound perception in order to focus on certain frequencies better than others not audible to the human ear even with a phonendoscope. This intuition is encoded by the presence of three attention mechanisms namely the squeeze and excitation mechanism [5], the convolutional block attention module [6], and the novel sinusoidal learnable attention [4]. This last attention mechanism acts by merging relevant information from activation maps at various levels of the network acting similarly to biological pyramidal-like neuronal cells that have been reported to code for high level concepts by neuroscientist. AUCO ResNet takes as input raw audio and outputs the respective class, without pre-processing, data augmentation or manual spectrogram generation. The model includes elements also present in the biological auditory cortex of mammals (rats), such as it is composed by six main blocks, it can evolve sound perception because also the mel spectrogram layer is trainable, it

has several attention levels and number of neurons within each stage has similar proportions of neurons found in rats and similar functionalities.

DenseNet 201 [17], with transfer learning pre-training and non-trainable Mel Spectrogram filter. In the DenseNet architecture, the input from one layer is later concatenated with the feature maps of all previous layers. This procedure allows for rich feature propagation and gradient flow, reducing the vanishing gradient problem faced in deep networks. This leads to a reduction in the number of parameters in the network and improved computation efficiency. The key features of the DenseNet architecture are the presence of Dense blocks: a group of layers where each layer is connected to every other layer in the same block. The Transition layers which are used to reduce the spatial size of the feature maps and prevent overfitting. The use of a Global average pooling as the final layer of the network, to generate the output predictions.

ResNet50 [18], with transfer learning pre-training and non-trainable Mel Spectrogram filter.

This architecture is designed to face the problem of training very deep networks, where the accuracy degrades with increasing depth due to the vanishing gradients problem. ResNet solves this problem by introducing residual connections. In residual connections, the input is directly added to the output of each layer. The residual connections allow for the effective propagation of gradients, even for very deep networks, and it is used as a technique to improve accuracy while reduced overfitting. ResNet, together with DenseNet and many others have been shown to generalize. [19] The architecture of ResNet consists of multiple residual blocks, where each block contains multiple convolutional layers and it could include additional operations such as batch normalization as well as attention mechanism [5]. The residual connections are implemented by summing the output of each block with its input, before passing the result to the next block.

InceptionResNet-V2 [20], with transfer learning pre-training and non-trainable Mel Spectrogram filter. This architecture is designed to perform multiple parallel convolutional filters of different sizes and pooling operations to capture information at multiple scales in the input image. It is similar to multi-scale learning process in computer vision allowing, thus, to learn features at different levels of detail, improving its overall representation and accuracy. The architecture is a sequence of multiple Inception blocks, each of which contains multiple parallel convolutional and pooling operations with in parallel filters, followed by a concatenation of the results. This allows the network to learn a wide range of features and is computationally efficient, as it reduces the number of parameters in the network.

Vision Transformer (ViT) [21].

The Vision Transformer architecture is based on the transformer architecture [22], which uses self-attention mechanisms (multiheaded attention) to learn the patterns between input elements in a sequence. In the case of Vision Transformer, the input elements are image patches (16x16 pixels), and the self-attention mechanism learns relationships between patches. This process is different from the one found in Convolutional Neural Network which seeks to exploit the inductive bias and produce the convolutional filters. Self-attention is a mechanism which allows to attend to different parts of its input and learn the relationships between the elements in the input. The ViT architecture consists of multiple stacked transformer blocks, each of which contains a self-attention mechanism and a fully connected layer used to generate the output predictions.

4.1. Mel Spectrogram filter

Some of the listed architectures (i.e. DenseNet201, ResNet50 and InceptionResnet-V2), already implemented in Keras API, are designed for image classification tasks. Therefore, it is necessary to implement the Mel Spectrogram filter, to convert audio files into images of their spectrograms. This filter firstly extracts the information about the audio frequencies, computing the short-time Fourier

transform (STFT) and its magnitude, and then the features from each audio signal. Thus, it computes and applies the matrix of the Mel Filter-Bank through a triangular Mel Filter-Bank, which imitates the perceptions of humans' ears. This setting choice has been led by the idea of making the results from the algorithms as explainable as possible for a hypothetical clinical implementation. In AUCO ResNet the Mel Spectrogram is inbuilt as a trainable layer of the network, and it learns the most discriminating frequencies for each class of the dataset during training [4].

4.2. Pre-training with UrbanSound4K

All the listed architectures have been pre-trained with the UrbanSound4K Dataset, using 9 of the 10 folds as a training-set and the remaining one as test-set. For AUCO ResNet the first 250 layers were frozen (apart the Mel Spectrogram layer), the later 150 layers were trainable, and a final dense layer with softmax activation function was added for the classification. For DenseNet, InceptionResNetV2 and ResNet50 and ViT, the last layer was substituted with a novel dense layer with softmax activation function for performing the final classification, while the entire network was allowed to be trained. The training hyperparameters are reported in Table I.

TABLE I. Hyperparameters used for traying

Parameters	*Values used with UrbanSound4K Dataset*	*Values used with Respiratory Sound Dataset*
fft 2048	fft 2048	fft 2048
n_mels 150	n_mels 150	n_mels 150
win_length 140	win_length 140	win_length 140
hop_length 344	hop_length 344	hop_length 344
optimizer RMSprop	optimizer RMSprop	optimizer RMSprop
epochs 100	epochs 100	epochs 50
batch size 16	batch size 16	batch size 16

4.3. Architectures training and finetuning

When training the architectures on the Respiratory Sound Dataset, the User-based Repeated random sub-sampling validation with under sampling was used to finetune the hyperparameters. The Repeated Random Sub-Sampling Validation, also called Monte Carlo Cross-Validation, has been preferred to K-fold Cross-Validation. It randomly splits the training set in 80% for the actual training and 20% for the validation at each iteration (10 in total) and finally provides the average of the metrics. Moreover, given the aforementioned database unbalances, at each iteration it was performed an under-sampling of the most numerous classes, to expose the models to the same number of patients for each class.

The labels of each target class, in both the tasks, was One-Hot encoded and the Softmax function was used as the Activation Function. The procedure was performed twice using two different loss functions: the Categorical Cross Entropy and the Balanced Categorical Cross-Entropy. The latter consists in multiplying the former by a weight computed by using the number of examples for each class. Therefore, if there is unbalancing between different classes, the loss function emphasizes the samples of the minority classes.

4.4. Architectures testing

The architectures have been tested classifying not only single respiratory cycles, but also complete audio files of a specific patient in the test-set, to simulate a real-world scenario. Testing the models on single patients was done by feeding the model with all the respiratory cycles of each patient. Thus, the associated class has been associated by computing the Mode over all the respiratory cycles of each patient.

The statistical Mode has been used considering the non-remarkable results collected by the tested architectures in this work. However, this approach may lead to a loss of data, especially in patients with light or not totally manifested conditions. Even more, certain conditions may be spotted in certain kind of breath cycles only (e.g. of specific duration range), and, again, using the Mode these data would be lost.

5. Results

The collected results have been computed using the categorical cross-entropy loss function. Results are computed per cycle in binary classification (Table I) and multi-class classification (Table III), as well as per patient, both in binary classification (Table II) and multi-task classification (Table IV). In bold there are the highest performances. In Tables V are reported the percentages of wrong predictions per respiratory cycle durations in the Binary classification task.

TABLE II. Binary Classification, Results on Respiratory Cycles

	Macro avg F1 score	***AUC***
AUCOResNet	**0.79**	**0.96**
DenseNet201	0.75	0.94
ResNet50	0.71	0.92
InceptionResNet-V2	0.53	0.76
ViT	0.12	0.52

TABLE III. Binary Classification, Results on Patients

	Macro avg F1 score
AUCOResNet	**0.81**
DenseNet201	0.70
ResNet50	0.74
InceptionResNet-V2	0.49
ViT	0.49

TABLE IV. Multiclass Classification, Results on Respiratory Cycles

	Macro avg F1 score	***AUC***
AUCOResNet	0.48	0.91
DenseNet201	0.47	**0.95**
ResNet50	**0.52**	0.94
InceptionResNet-V2	0.31	0.48
ViT	0.38	0.75

TABLE V. Multiclass Task, Results on Patients

	Macro avg F1 score
AUCOResNet	0.48
DenseNet201	0.46
ResNet50	**0.53**
InceptionResNet-V2	0.32
ViT	0.02

TABLE VI. Binary Task, Results on Respiratory Cycles

	≤ 1s	*]1s,1.5s]*	*]1.5s,2 s]*	*]2s,2.5 s]*	*]2. 5s,3s]*	*]3s ,3.5s]*	*]3. 5s, 4s]*	*> 4s*
AUCO ResNet	28%	7%	6%	8%	5%	4%	3%	1%
DenseNet 201	26%	11%	9%	8%	8%	2%	2%	0%
ResNet-50	29%	12%	8%	10%	9%	9%	3%	2%
Inception ResNet-V2	22%	9%	9%	10%	8%	3%	1%	0%
ViT	78%	91%	91%	89%	91 %	97 %	99 %	99 %

6. Discussion

In the Binary Task the AUCO ResNet performed better than the other architectures when pre-trained with the UrbanSound4K Database. It is not a surprising result considering that this architecture has been developed for audio analysis.

In the Multiclass Task the results are poor in all the architectures tested. Again, this was an expected result given the unbalanced database used as well as its limited size. In fact, especially the Vision Transformer architecture needs to be trained with a lot of data. In a similar fashion, also the AUCO ResNet needs more data as its several attention mechanisms are not capable of filtering the noise keeping the frequencies containing more information.

Analysing the wrong prediction percentages in different respiratory cycle lengths, as shown in Table V, it becomes clear that the architectures were better performing with longer cycles. This insight could lead to new experimentations, in which only cycles with a certain minimum length would be considered. Even if this suggested approach needs to be evaluated from a clinical perspective, because important information could be stored in the small respiratory cycles too.

Given the huge amount of COPD samples in the Respiratory Sound Database, it may be considered the idea of developing a COPD detector, training AUCO ResNet architecture on a binary task: COPD positive samples vs the sum of the other classes, for differential diagnosis purposes.

Moreover, to improve the performances of the AUCO ResNet architecture with transfer learning, it may be preferred a more contextualised dataset for the pre-training (e.g. the one used in the original paper, based on breath and cough audios of COVID19 patients).

Finally, the most obvious improvement would be achieved by feeding the models with a huge and well-balanced dataset.

7. Conclusions

Performing benchmarks is as important as building and releasing new architectures. It not only allows to compare different solutions, but also to better contextualise their limitations and their potential. This work goes even beyond, importing models from the computer vision field to concretely taste their applicability in a totally different context. This is the essence of innovation, and it should never be underestimated, even when the results are not that enthusiastic, as in this case. The hope is to inspire other researchers to explore and test new combinations of architectures and configurations, having in mind the real-world applicability of their work.

AUCO ResNet has provided remarkable results in audio analysis, but new studies, based on more balanced databases, are needed to deeply explore its potential.

Once a ML architecture would reach satisfying results in this context, validated by clinical trials, then endless opportunities, will be unlocked democratising access to care, bringing tangible benefits to people all around the globe. This is more than enough to keep investing in AI research applied to healthcare, despite the challenges it presents compared to other quicklier remunerative fields.

Acknowledgements

This work was partially supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU.

References

[1] S. M. Levine and D. D. Marciniuk, “Global Impact of Respiratory Disease: What Can We Do, Together, to Make a Difference?,” *Chest*, vol. 161, no. 5, p. 1153, May 2022, doi: 10.1016/J.CHEST.2022.01.014.

[2] M. Chimienti, I. Danzi, V. Gattulli, D. Impedovo, G. Pirlo, and D. Veneto, “Behavioral Analysis for User Satisfaction,” *Proceedings - 2022 IEEE 8th International Conference on Multimedia Big Data, BigMM 2022*, pp. 113–119, 2022, doi: 10.1109/BIGMM55396.2022.00027.

[3] V. Gattulli, D. Impedovo, G. Pirlo, and G. Semeraro, “Early Dementia Identification: On the Use of Random Handwriting Strokes,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13424 LNCS, pp. 285–300, 2022, doi: 10.1007/978-3-031-19745-1_21.

[4] T. Nguyen and F. Pernkopf, “Lung Sound Classification Using Co-tuning and Stochastic Normalization,” 2021, Accessed: Jan. 31, 2023. [Online]. Available: https://github.com/makcedward/nlpaug

[5] I. Moummad and N. Farrugia, “SUPERVISED CONTRASTIVE LEARNING FOR RESPIRATORY SOUND CLASSIFICATION”.

[6] L. Xu, J. Cheng, J. Liu, H. Kuang, F. Wu, and J. Wang, “ARSC-Net: Adventitious Respiratory Sound Classification Network Using Parallel Paths with Channel-Spatial Attention,” *Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021*, pp. 1125–1130, 2021, doi: 10.1109/BIBM52615.2021.9669787.

[7] Z. Wang and Z. Wang, “A DOMAIN TRANSFER BASED DATA AUGMENTATION METHOD FOR AUTOMATED RESPIRATORY CLASSIFICATION,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 9017–9021, 2022, doi: 10.1109/ICASSP43922.2022.9746941.

[8] Y. Ma *et al.*, “Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm,” *BioCAS 2019 - Biomedical Circuits and Systems Conference, Proceedings*, Oct. 2019, doi: 10.1109/BIOCAS.2019.8919021.

[9] Y. Ma, X. Xu, and Y. Li, “LungRN+NL: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2902–2906, 2020, doi: 10.21437/INTERSPEECH.2020-2487.

[10] V. Dentamaro, P. Giglio, D. Impedovo, L. Moretti, and G. Pirlo, “AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath,” *Pattern Recognit*, vol. 127, p. 108656, Jul. 2022, doi: 10.1016/J.PATCOG.2022.108656.

[11] F. Renard, S. Guedria, N. de Palma, and N. Vuillerme, “Variability and reproducibility in deep learning for medical image segmentation,” *Sci. Rep.*, vol. 10, no. 1, p. 13724, Aug. 2020.

[12] B. M. Rocha et al., “An open access database for the evaluation of respiratory sound classification algorithms,” Physiol Meas, vol. 40, no. 3, Mar. 2019, doi: 10.1088/1361-6579/AB03EA.

[13] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in Proceedings of the ACM International Conference on Multimedia - MM ’14, New York, New York, USA: ACM Press, 2014.

[14] G. Garcia, G. Moreira, D. Menotti, and E. Luz, “Inter-Patient ECG Heartbeat Classification with Temporal VCG Optimized by PSO,” Sci Rep, vol. 7, no. 1, Dec. 2017, doi: 10.1038/S41598-017-09837-3.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” IEEE Trans Pattern Anal Mach Intell, vol. 42, no. 8, pp. 2011–2023, Sep. 2017, doi: 10.1109/TPAMI.2019.2913372.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in Computer Vision – ECCV 2018, in Lecture notes in computer science. Cham: Springer International Publishing, 2018, pp. 3–19.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Dec. 2015.

[19] F. He, T. Liu, and D. Tao, “Why ResNet works? Residuals generalize,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 12, pp. 5349–5362, Dec. 2020.

[20] C. Szegedy, S. Ioffe, and A. Vanhoucke Vincent and Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” Feb. 2016.

[21] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020, doi: 10.48550/arxiv.2010.11929.

[22] A. Vaswani et al., “Attention Is All You Need,” Adv Neural Inf Process Syst, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Nov. 11, 2021. [Online]. Available: https://arxiv.org/abs/1706.03762v5