Convolution-Vision Transformer for Automatic Lung Sound Classification

José Neto
Center of Informatics
Federal University of Pernambuco
Recife, Brazil
jdcn@cin.ufpe.br

Nicksson Arrais
SiDi
Recife, Brazil
nicksson.a@sidi.org.br

Tiago Vinuto
SiDi
Recife, Brazil
t.vinuto@sidi.org.br

João Lucena
SiDi
Recife, Brazil
j.lucena@sidi.org.br

***Abstract*—Auscultation is an essential part of clinical examination since it is an inexpensive, noninvasive, safe, and one of the oldest diagnostic techniques used to diagnose various pulmonary diseases. In literature, machine learning models were proposed in various studies for lung sound classification to overcome the ear acuity and the inherent inter-listener variability. In this work, we propose a hybrid Convolution-Vision Transformer architecture that explores the usage of Convolutional with Vision Transformers in a single system. We evaluate our proposed method on ICBHI 2017 database for the four-class sound classification of lung sounds to demonstrate the effectiveness of our method which has achieved a score of 57.36% surpassing many state-of-art models.**

***Index Terms*—Auscultation, Lung Sound Classification, Vision Transformer, ICBHI dataset**

I. INTRODUCTION

Auscultation is one of the main methods used in the diagnosis of respiratory systems due to its simplicity, practicality and low cost. This process consists of listening to the internal sounds of the human body through a stethoscope in order to verify the integrity of lung function [1].

These internal sounds are produced by the flow of air along the respiratory tract [2], during the process of expiration and inspiration. The pulmonary auscultation method requires a well-trained professional capable of interpreting lung sounds and providing the diagnosis [3].

Therefore, even if it is an effective form of diagnosis, auscultation is a subjective method that is subject to false diagnosis [4]. Palaniappan *et al.* [5] surveyed a range of studies about the process of monitoring lung health using techniques of machine learning on handcrafted features while in [6] and [7] showed that Convolutional Neural Networks can works as well or outperform classical models demonstrating that machine learning techniques can assist health professionals, such as nurses and doctors.

In this work, we aim to successfully detect and classify the adventitious sounds, a sound whose presence usually indicates a pulmonary disorder [8], and the normal breath sounds through a combination of machine learning models and signal processing techniques. We also present works that propose the use of machine learning for the classification of respiratory anomalies through auscultation, exploring the individual approach to the problem, as well as its methodologies.

For short, the main contributions of this work are described as follow:

1) Proposed hybrid model, which combines Convolutional Neural Networks (CNNs) and Vision Transformer (ViT). To the best of our knowledge, this is the first to explore these combinations in the ICBHI challenge [9].
2) We demonstrate it's performance in the ICBHI 2017 dataset achieving excellent results at specificity and score metrics outperforming other state-of-the-art approaches.

II. RELATED WORKS

In literature, Deep Learning mainly based on convolutional neural networks is widely used for sound classification and recognition. Ma *et al.* [10] proposed a bilinear bi-ResNet neural network in the task of respiratory sounds classification, which was training on the features extracted via Short-time Fourier transform (STFT) and wavelet analysis. Their experimental results achieved a score of 50.16% on the official ICBHI 2017 60-40 split.

Ma *et al.* [11] introduce LungRBN+NL, a model in the ResNet backbone with a non-local block [12] to calculate the relationship across time and frequency domain. They also use short-time Fourier transform (STFT) and wavelet feature extraction. In the official split, their LungRN+NL architecture has reported a performance score of 52.26%.

On the other hand, Gairola *et al.* [13] proposed a Deep Neural Network (DNN) called RespireNet, It is formed by ResNet34 and fully connected layers, additionally, They introduced data augmentation based on concatenation, device-specific fine-tuning, blank region clipping and smart padding to improve the accuracy. Their best score for the 4-class task was 56.2% in the 60-40 split.

Meanwhile, Zhao *et al.* in [14] proposed to explore the effectiveness of a multi-branch Temporal Convolutional Network (TCN) architecture integrated with Squeeze-and-Excitation Network. They denoted their system as MBTCNSE making use of spatial and temporal information from the log mel spectrogram features for respiratory sound classification. The authors reached a score of 75.7% in the 80-20 train/test split.

Using the ResNeSt [15] as a backbone, which is a model based on ResNe, ResNeXt, SK-Net, and SE-Net, Wang *et al.* [16] tested data augmentation methods like circular padding, splice and also mixup in the spectrogram features. Their score on the ICBHI dataset using the official 60-40 train/test split was 55.7.%

The ARSC-Net proposed by Xu *et al.* [17] make use of two types of features from adventitious respiratory sound, the Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrogram. The two types of features are entered into the parallel encoder paths with residual attention for extracting feature representation and then fused into a channel-spatial attention mechanism. They achieved a great score of 56.76% for the four-class sound classification in the 60-40 official split.

Finally, Using CRNN (Convolutional Recurrent Neural Network) by inputting multiple respiratory sound image features such as spectrogram, scalogram and Constant-Q Transform, Asatani *et al.* [18] obtained a score of 72% using 5-fold validation.

III. PROPOSED SCHEME

Our proposed framework illustrated in the Fig. 1 is composed of two parts: Feature Extraction, a Convolutional block formed by three independents kernels and an attention module to enhance channel and spatial information and last, a Vision Transformer Network, the last two parts compose the model. In this section each one will be described.

A. Feature Extraction

For the classification of respiratory sounds, we extract three types of features such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrogram and Constant-Q Transform (CQT) as frequency-time representations, each one will be discussed hereafter.

1) Mel Spectrogram: Mel Spectrogram is a combination of the spectrogram and the Mel scale, where the first one is the magnitude squared of the short-time Fourier transform (STFT) a tool designed to analyze the way the frequency of non-stationary signal changes over time. The STFT can be described mathematically as [19]:

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-j\omega n} \tag{1}$$

Where:

X_m - Is the Discrete Fourier Transform of windowed data centered about time mR;
$w(n)$ - Is Window function of length M (e.g., Hamming);
R - The difference between the window length M and the overlap length L, known as Hop size.

A Hamming window is applied to each frame to greatly reduce spectral leakage before conducting DFT. The Hamming window has the form [20]:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M}\right), \quad 0 \leq n \leq M \tag{2}$$

The frequency bands are extracted by applying the Mel filter bank on the power spectrum of each frame to obtain the Mel Spectrogram. The Mel scale simulates the way how the human ear reacts to a sound, being more sensitive at the lower frequency and less so at the higher frequency. To compute the Mel Scale, a non linear transformation is applied in the original frequency using the formula as below::

$$f_{mel} = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

where f and f_{mel} denotes respectively, the physical and perceived frequency in hertz.

2) MFCC: Mel Frequency Cepstral Coefficients (MFCC) is a representation of a sound's short-term power spectrum. It refers to the inverse Fourier transform of the logarithm of the estimated signal spectrum. The magnitude frequency response of each frame is obtained by computing the Discrete Fourier Transform (DFT) of each frame using pre-processed sound data. DFT computation can be expressed as;

$$X(k) = \sum_{n=0}^{n-1} x(n)e^{\frac{-j2\pi kn}{N}}, \quad 0 < k < N-1 \tag{4}$$

here N is the number of points to compute de DFT. Like the Mel Spectrogram, Mel spectrum is computed by passing the Fourier transformed signal through a set of Mel-filter banks. The Mel spectrum of the magnitude spectrum X(k) is computed by multiplying the magnitude spectrum by each Mel weighting filters.

$$s(k) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k), \quad 0 < m < M-1 \tag{5}$$

where M is total number of Mel weighting filters. $H_m(k)$ is the weight given to the $k-th$ energy spectrum bin contributing to the $m-th$ output band and is expressed as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m)-f(m-1)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{6}$$

To conclude, the Discrete Cosine Transform (DCT) is performed on the Logarithm compressed Mel spectrum. The DCT can be expressed as:

$$C_n = \sum_{m=0}^{M-1} log_{10} s_m \frac{\cos(\pi n(m-0.5)}{M} \tag{7}$$

3) Constant-Q Transform: when the window function is set larger In the Short-Time Fourier Transform we obtain higher frequency resolution but with low time resolution which concentrates much more information in the high-frequency region. Constant-Q Transform, CQT [21], allows changing the length of the window function to allow harmonic frequencies to be represented in equal intervals in the transform domain. The Constant-Q Transform equation is shown below

$$X(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} W(n,k)x(n)e^{-j\frac{2\pi Q}{N_k}n} \tag{8}$$

Where:

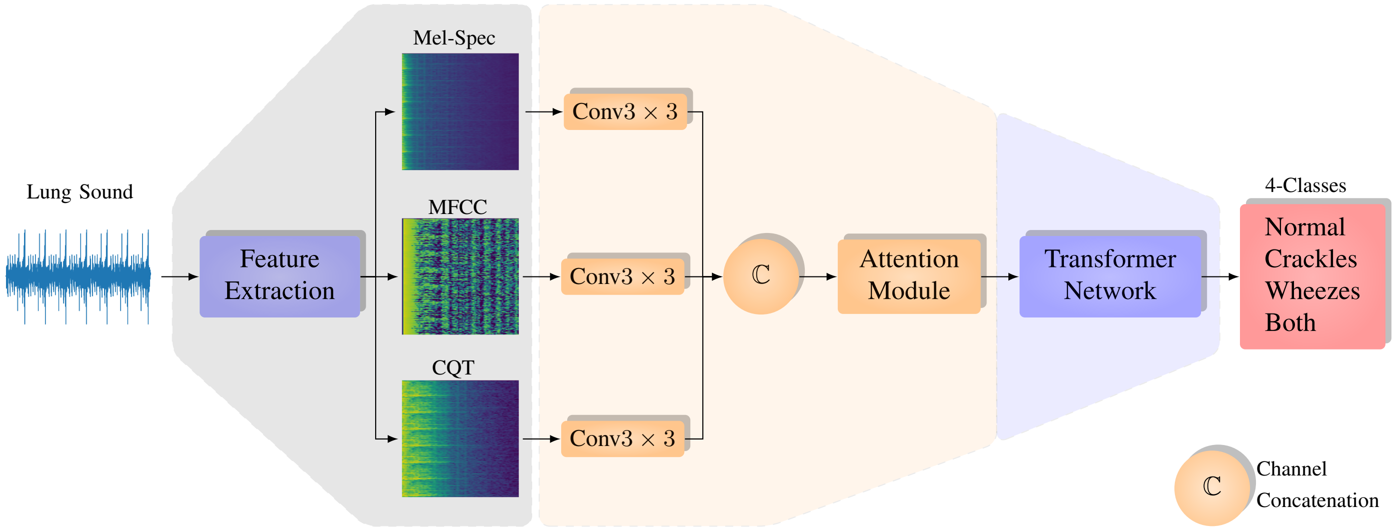

Fig. 1. Proposed Framework.

$W(n)$ - is the window function;
x(n) - is the original signal and
Q - is a constant.

Furthermore, $N(k)$ is represented by:

$$N(k) = \frac{f_s}{f_k} Q \tag{9}$$

Here f_s is the sampling frequency and f_k is the center frequency. The range of k is can be found like the equation 10.

$$k = \left\lceil b \log_2 \left(\frac{f_{\max}}{f_{\min}} \right) \right\rceil \tag{10}$$

b is the number of octave divisions, and fmax and fmin are the maximum and minimum frequencies, respectively. The CQT representation increase the low-frequency resolution [22] which can be useful for classifying adventitious sounds.

B. Convolutional Block

Convolutional Neural Networks - CNN is a class of artificial neural networks more suited for image-focused tasks [23], [24]. It takes this name from the linear operation called convolution. CNNs have including convolutional layers, non-linearity such as ReLu, pooling, and fully-connected layers [25]. In this work the Convolutional module is first designed with three branches convolution blocks using a 3×3 kernel size to extract each channel's feature separately while maintaining the size of the original spectrogram input. We aim to increase and learn more generalizable features representations from fewer data while maintain a low cost model.

Second, according to Woo *et al.* in [26], We include the Convolutional Block Attention Module (CBAM) which contains two sequential sub-modules called the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The CAM module is defined as the expression in Eq. 11 where it is used multi-layer perceptron (*MLP*) with one hidden layer, Global Average Pooling (*GAP*) and Global Max Pooling (*GMP*) to leverage global information of the input feature $x \in \mathbb{R}^{C \times H \times W}$ [27].

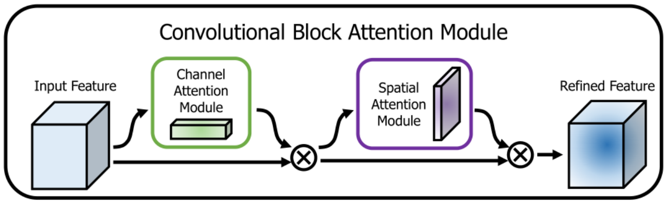

Fig. 2. CBAM Attention Module [26].

$$CAM(x) = \sigma(MLP(GAP(x)) + MLP(GMP(x))) \tag{11}$$

here σ is the sigmoid activation.

In the SAM module, the goal is to generate a spatial attention map by utilizing the inter-spatial relationship of features. It works in the principle of 'where' is an informative part. To compute the spatial attention first is apply average-pooling (AVG) and max-pooling (MP) operations along the channel axis and concatenate them to generate an efficient feature descriptor. The spatial attention module is computed as:

$$SAM(x) = \sigma\left(f^{7x7}\left([\text{AVG}(x); \text{MP}(x)]\right)\right) \tag{12}$$

where $f^{7\times7}$ is the 7×7 kernel size of the convolution operation. CBAM is suitable to explore relationships of features to tell the network what and where to pay attention enhancing informative local regions [27].

C. Vision Transformer

The Vision Transformer (ViT) in Fig 3, is a model similar to initial transformer [28] applied in natural language processing

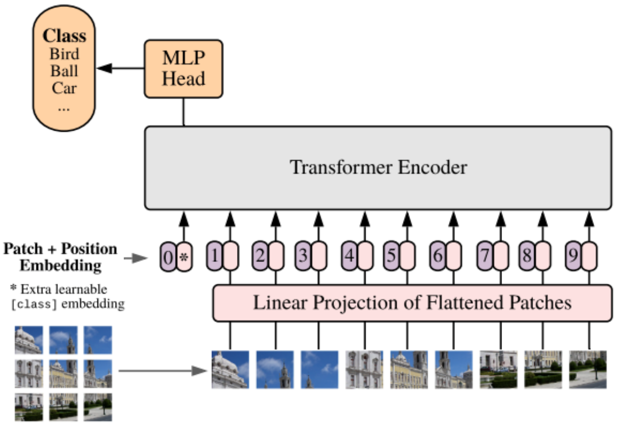
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Fig. 3. Vision Transformer architecture [29].

tasks but created to deal with images. To avoid the quadratic computational cost, the ViT computes relationships among pixels in small fixed-sized patches of the image, each of them are then linearly embedded, position embeddings are added and the resulting sequence of vectors is fed to a standard transformer encoder.

To perform classification, the standard approach of adding an extra learnable "classification token" to the sequence is used. In the original Vision Transformers (ViT) [29], the authors concluded that to perform on par with Convolutional Neural Networks (CNNs), ViTs need to be pre-trained on larger datasets. This is mainly due to the lack of inductive biases in the ViT architecture unlike CNNs, which don't have layers that exploit locality. To deal with this problem, Touvron *et al* [30] proposed a novel model called Data-efficient image Transformers, (DeiT). The authors introduced a new distillation procedure based on a distillation token, which plays the same role as the class token, except that it aims at reproducing the label estimated by the teacher.

IV. Experiments

A. Dataset

Dataset ICBHI 2017 challenge is a scientific challenge organized by International Conference on Biomedical and Health Informatics in 2017, which provides a respiratory sound database and an official scoring method [31]. This database consists of a total of 5.5-hours recordings containing annotated respiratory cycles from 126 subjects. For simplicity, a record is defined as the lung sounds collected from one patient and a cycle is defined as a respiratory cycle from a patient. Hence, the total recording contains 6,898 cycles which comprise 3,642 "normal", 1864 "crackles", 886 "wheezes", and 506 "crackle plus wheeze" cycles.

B. Loss Function

TABLE I
Performance comparison of our proposed model with SOTA in four class classification and official 60-40 split.

Method	Se (%)	Sp (%)	Sc(%)
LungBRN [10]	31.12	69.20	50.16
LungRN+NL [11]	41.32	63.20	52.26
ResNeSt+augmentation [16]	40.20	70.40	55.30
CNN+CBA+BRC [13]	39.60	71.80	55.70
CNN+CBA+BRC+FT [13]	40.10	72.30	56.20
ARSC-Net [17]	**46.38**	67.13	56.76
Ours	36.41	**78.31**	**57.36**

Due to the high imbalance of the dataset we used the class balance loss proposed by Cui *et al.* [32]. This loss uses a re-weighting that uses the effective number of samples, the expected volume of samples, for each class. This method is agnostic and can be applied in any loss, in the cross entropy loss it can be written as:

$$CB(p,y) = \frac{1-\beta}{1-\beta^{n_y}}\mathcal{L}(p,y) \tag{13}$$

Where:

n_y - Is the number of samples of the class y;
β - Is a hyper parameter and
p - Is the model output.

C. Settings

The proposed model was all built in an open source framework. We use Pytorch for machine learning and librosa to extract sound features, the Adam optimizer, and a fixed learning rate of 0.0001 in the Four-Class Task study.

We use the pre-trained weights provided by DeiT to initialize the transformer model. All others layers without pre-training are randomly initialized. In the four-class respiratory sound classification, we adopt the official 60-40 data split and the score as evaluate metric according to the original challenger. The score, Eq. 16, is the mean of sensitivity (Se) Eq. 14 and specificity (Sp) Eq. 15. We extract 96-dimensional MFCCs, Mel-spectrogram and Constante-Q transform, all features were processed with a window length of 2048 and hop length of 512.

$$Se = \frac{TP}{TP+FN} \tag{14}$$

$$Sp = \frac{TN}{TN+FP} \tag{15}$$

$$Sc = \frac{Se+Sp}{2} \tag{16}$$

D. Performance Comparison

Table I shows the SOTA for only published four-class classification task on official 60-40 split. Compared to the current state-of-the-art works, our model outperforms all compared models achieving the highest score which was improved from 56.76% to 57.36%. However, it is observed that our model does not perform well while trying to differentiate the adventitious lung sounds, resulting in a lower sensitivity score which can be seen in the Fig 4 which shows the confusion matrix where the "Both" class contains both Wheezes and Crackles adventitious sounds.

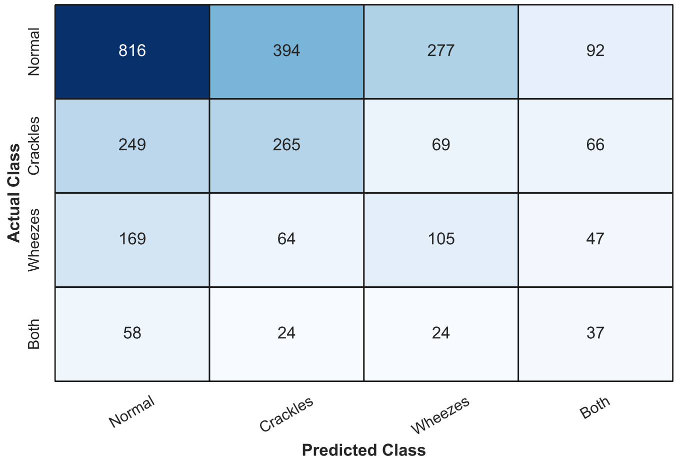

Fig. 4. Confusion matrix. where W_C contains both Wheezes and Crackles adventitious sounds.

E. Ablation Studies

Table II shows the results of different analyses while using distinct DeiT models. The two first models were tested without CNN and attention mechanism while the last two were done with aggregation of these modules. We can see that, combining CNN and attention module performs better than the DeiT model alone demonstrating that a hybrid approach is effective to identify adventitious sounds.

TABLE II
THE ABLATION STUDY TABLE

Model	Params	Se (%)	Sp (%)	Sc(%)
DeiT Small	21M	30.18	76.72	53.45
DeiT Base	86M	32.22	77.37	54.80
Deit Base + CNN	86M	33.78	77.70	55.74
Deit Base + CNN + Att	86M	**36.41**	**78.31**	**57.36**

V. Conclusion

In this paper we demonstrated how pre-trained new models like Vision Transformers can be a useful tool and jointly with CNN, learn on small datasets. Experimental results show our best model, which uses an attention mechanism block outperforms state-of-the-art score metric, leading by it's specificity, systems when evaluate on the ICBHI 2017 database proving the effectiveness of our proposed method. In the future, we want to evaluate two-class and lung sound disease classification tasks and improve the sensitivity of our model using data argumentation techniques.

Acknowledgement

The results presented in this paper have been developed as part of a project between SiDi, financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/91.

References

[1] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka, "Respiratory sounds: advances beyond the stethoscope," *American journal of respiratory and critical care medicine*, vol. 156, no. 3, pp. 974–987, 1997.

[2] A. Sovijarvi, "Characteristics of breath sounds and adventitious respiratory sounds," *Eur Respir Rev*, vol. 10, pp. 591–596, 2000.

[3] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Annals of thoracic medicine*, vol. 10, no. 3, p. 158, 2015.

[4] S. Mangione and L. Z. Nieman, "Pulmonary auscultatory skills during training in internal medicine and family practice," *American journal of respiratory and critical care medicine*, vol. 159, no. 4, pp. 1119–1124, 1999.

[5] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: a systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013.

[6] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–9, 2017.

[7] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial intelligence in medicine*, vol. 88, pp. 58–69, 2018.

[8] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman, "Definition of terms for applications of respiratory sounds," *European Respiratory Review*, vol. 10, no. 77, pp. 597–610, 2000.

[9] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[10] 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 biresnet deep learning algorithm," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[11] 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." in *Interspeech*, 2020, pp. 2902–2906.

[12] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[13] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 527–530.

[14] Z. Zhao, Z. Gong, M. Niu, J. Ma, H. Wang, Z. Zhang, and Y. Li, "Automatic respiratory sound classification via multi-branch temporal convolutional network," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9102–9106.

[15] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[16] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9017–9021.

[17] 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," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1125–1130.

[18] N. Asatani, T. Kamiya, S. Mabu, and S. Kido, "Classification of respiratory sounds by generated image and improved crnn," in *2021 21st International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2021, pp. 1804–1808.

[19] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[20] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[21] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[22] P. Singh, G. Saha, and M. Sahidullah, "Non-linear frequency warping using constant-q transformation for speech emotion recognition," in *2021*

International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1–6.

[23] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[24] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*. Ieee, 2017, pp. 1–6.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[27] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *arXiv preprint arXiv:2111.07624*, 2021.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[31] B. Rocha, D. Filos, L. Mendes, I. 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 *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[32] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.