IberSPEECH 2022
14-16 November 2022, Granada, Spain

Respiratory Sound Classification Using an Attention LSTM Model with Mixup Data Augmentation

Noelia Salor-Búrdalo[1], *Ascensión Gallardo-Antolín*[1]

[1]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain

100346893@alumnos.uc3m.es, gallardo@ing.uc3m.es

Abstract

Auscultation is the most common method for the diagnosis of respiratory diseases, although it depends largely on the physician's ability. In order to alleviate this drawback, in this paper, we present an automatic system capable of distinguishing between different types of lung sounds (neutral, wheeze, crackle) in patient's respiratory recordings. In particular, the proposed system is based on Long Short Term-Memory (LSTM) networks fed with log-mel spectrograms, on which several improvements have been developed. Firstly, the frequency bands that contain more useful information have been experimentally determined in order to enhance the input acoustic features. Secondly, an Attention Mechanism has been incorporated into the LSTM model in order to emphasize the more relevant audio frames to the task under consideration. Finally, a Mixup data augmentation technique has been adopted in order to mitigate the problem of data imbalance and improve the sensitivity of the system. The proposed methods have been evaluated over the publicly available ICBHI 2017 dataset, achieving good results in comparison to the baseline.

Index Terms: respiratory sound classification, wheeze, crackle, frequency analysis, LSTM, attention mechanism, mixup data augmentation

1. Introduction

Respiratory diseases are one of the most common causes of mortality in the world according to the World Health Organization (WHO) [1] causing millions deaths worldwide [2]. A relevant example of this kind of diseases is COVID-19, new pneumonia emerged in Wuhan, China, in December 2019 [3]. Since its appearance, more than 550 million people have been infected in the world [4] and 6.3 million have died [5].

The auscultation is the most common way to check the respiratory conditions of the patient. Lung sounds can be classified into two groups: normal and abnormal. The two most important abnormal lung sounds are wheezes and crackles. The first ones are continuous high-pitched adventitious sounds that result from obstruction of breathing airway [6]. The second ones are explosive and discontinuous sounds present during inspiratory and expiratory parts of the breathing cycle with a significantly smaller duration compared to the total breathing cycle [7].

However, the auscultation is dependent on the knowledge and skills of the healthcare staff. A study concludes that only 80% of wheezes can be detected correctly [8]. For this reason, it is necessary to research into the development of automatic respiratory sound classification systems to help medical diagnosis.

The architecture of this kind of systems typically consists of a pre-processing and feature extraction stage followed by a classification step that provides the type of breath sound to which the input audio signal belongs. Several acoustic features have been proposed for this task, as, for example, spectrograms [9], wavelet coefficients [10], and Mel-Frequency Cepstral Coefficients (MFCC) [11, 12, 13]. Regarding the classification stage, initial works proposed the use of traditional machine-learning algorithms such as K-Nearest Neighbours [14], Support Vector Machines [14, 15, 16] or Hidden Markov Models [13]. More recently, deep-learning models have been successfully developed for this task. Among the different architectures evaluated, it is worth mentioning ResNet networks with spectrograms and wavelet features as input [17], convolutional neural networks [18], different recurrent neural networks with MFCC features [19] and the combination of recurrent layers with Gammatone spectrograms as input features [20].

One of the main problems with working with medical data is the difficulty of collecting databases with a large number of samples. To address this problem, the ICBHI 2017 database was released [21, 22] at the International Conference on Biomedical Health Informatics (ICBHI) in 2017. Although the database developers established a train-test division to facilitate the comparison between different systems, studies using other 80/20 train-test distributions began to be published. This has led to the loss of its comparative power, what, however, has not prevented the use of this database in subsequent studies.

In this context, in this paper we propose an automatic system based on Long Short-Term Memory (LSTM) networks fed with log-mel spectrograms for classifying lung sounds. Specifically, we have focused on several techniques that improve the baseline system, such as the analysis of the frequency components of the audio signals in order to enhance the input features, the incorporation of an Attention Mechanism into the LSTM network in order to better characterize the temporal evolution of the input sequences, and the use of Mixup data augmentation methods in order to balance the training data and therefore, improve the sensitivity of the system. The proposed techniques have been evaluated over the ICBHI 2017 dataset [21], achieving good results in comparison to the baseline.

This paper is organized as follows. Section 2 gives a description of the database used. Section 3 explains the different steps of the proposed methodology: database preprocessing and feature extraction, LSTM-based classifier, enhancement of the input features, incorporation of the Attention Mechanism and use of Mixup data augmentation techniques. Section 4 gives an account of the experiments carried out together with the main results obtained. Finally, Section 5 describes the final conclusions and some possible lines of research.

2. ICBHI 2017 DATABASE

The experiments have been carried out over the ICBHI 2017 database [21, 22]. It contains 920 audio recordings of lung sounds from 126 patients of different ages and with different respiratory conditions (lower and/or upper respiratory tract infections, pneumonia, asthma, etc.) with a total duration of 5.5

10.21437/IberSPEECH.2022-13

hours. The samples have been taken with four different stethoscopes and in seven different chest locations. It is important to note that ambient noise and conversations are present in the audios to simulate the normal conditions in which an auscultation is performed.

The database has been manually segmented into respiratory cycles yielding $6,898$ audio signals labelled into 4 different acoustic classes (normal, wheeze, crackle, and crackle and wheeze). Table 1 contains the number of samples per class. As can be seen, the database is highly imbalanced as class 0 (normal sounds) accounts for more than 50 % of the total samples.

Table 1: *Class distribution of the ICBHI* 2017 *database.*

Class	Label	No. samples	Ratio
0	Normal (N)	3,642	52.8 %
1	Wheeze (W)	886	12.8 %
2	Crackle (C)	1,864	27.0 %
3	Crackle and Wheeze (C + W)	506	7.4 %
Total		6,898	100.0 %

The database has been divided into train ($4,223$ samples), validation ($1,386$ samples) and test ($1,289$ samples) sets, which correspond approximately to, respectively, 60%, 20% and 20% of the total samples, in a patient-independent fashion. That is, all the respiratory cycles of a certain patient are included into the same set.

3. Respiratory sound classification system

In this Section, the main components of the proposed lung sound classification system are described.

3.1. Data preprocessing and feature extraction

This stage is composed of the following processes:

3.1.1. Resampling

As the original audios are recorded at different sampling frequencies ranging from 4 KHz to 44.1 KHz, it is necessary to resample them to a common frequency (16 KHz, in this case).

3.1.2. Truncation

The audio files have a variable duration between 0 and 10 s seconds, although only a small number of them have a length greater than 6.25 s. Therefore, in order to reduce the temporal dimension of the input features of the LSTM network, longer audios than 6.25 s are truncated to this length.

3.1.3. Feature extraction

The chosen acoustic features are the log-mel spectrograms that are computed at a frame period of 10 ms using Hamming windows of 20 ms length and a mel-scale filterbank composed of $n_B = 40$ triangular filters. Note that, due to the previous truncation process, the maximum number of frames of the resulting log-mel sequences is $T = 625$.

3.1.4. Normalization and padding

Log-mel spectrograms are normalized to mean zero and variance one. As the length of the LSTM input sequences is set to the fixed value $L = 625$, shorter log-mel sequences than this quantity are padded with dummy values.

As original audios are contaminated with background noise, some preliminary experiments were performed by applying several denoising techniques. However, as results did not improve, this denoising process was finally discarded.

3.2. LSTM-based classifier

The proposed system is based on LSTM networks [23], a kind of Recurrent Neural Network (RNN) capable of modeling temporal sequences, as is our case.

After a preliminary experimentation, the chosen baseline architecture is shown in Figure 1. The input is made up of a 2-dimensional matrix where the first dimension is the number of frames of each sequence ($T = 625$) and the second one is the number of features that corresponds to the number of Mel filters ($n_B = 40$). Next, there is a Masking layer whose function is that the dummy values of the padded sequences not to be used in further computations (see Subsection 3.1). Then, the output of this layer passes through a LSTM layer with 128 cells and a hyperbolic tangent as activation function. Next layer is an Average Pooling (green box) whose function is to compact the LSTM output sequence into an utterance-level representation, by computing its mean along the time dimension. Then, this output is normalized by means of a Batch Normalization layer and passes through a Dense layer with 128 neurons and a dropout of 40 % rate that helps prevent overfitting. The process finishes with a Batch Normalization and a Dense layer whose activation function is a softmax and has 4 neurons that corresponds to the number of acoustic categories to classify.

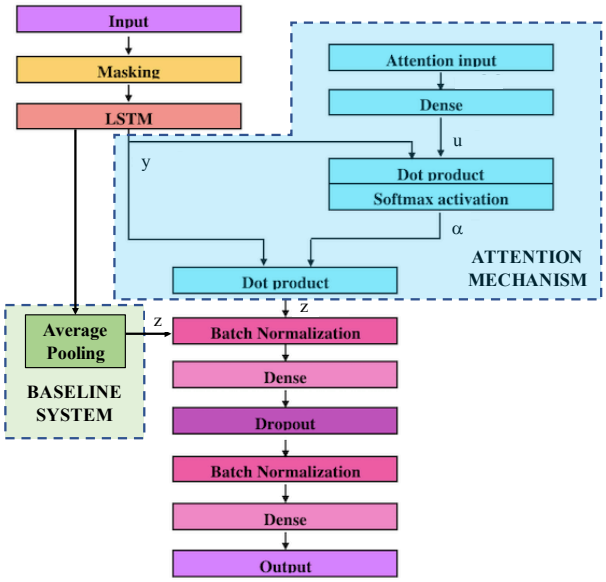

Figure 1: *Architecture of the respiratory sound LSTM-based classification system. The difference between the baseline and the attentional model is that the Average Pooling step in the first one (green box) is replaced by the Attention Mechanism in the second one (blue boxes).*

3.3. Enhancement of the input features

All components of the log-mel spectrograms are not of equal importance as respiratory sounds present more energy content in certain frequency bands. In addition, we observed that, usually, low frequencies contain more useful information whereas high frequencies are more prone to noise. Figure 2 represents the normalized average log-energy of each of the 40 mel filters for the training audio files. It can be seen that the first 10

filters (corresponding to the band from 0 to 800 Hz) and 20 filters (band from 0 to $1,800$ Hz) account for approximately the 60 % and the 85 % of the total energy, respectively. The peak of filter 32 (frequency band around $4,000$ Hz) corresponds to noise likely due to the recording equipment or environment.

From these observations, we propose to enhance the input features by manually selecting the range of low frequency filters that concentrates most of the energy, while discarding higher frequency bands.

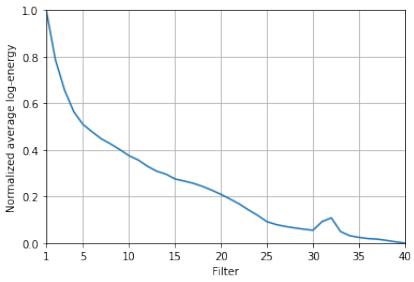

Figure 2: *Normalized average log-energy of each mel filter.*

3.4. Attention Mechanism

Attention is a basic psychological process, whose main function is to select the most relevant stimuli from the environment. Following this idea, the attentional LSTM networks try to emphasize the most relevant LSTM frames for the classification task in hand, whereas diminishing the contribution of the less informative ones. This approach has been successfully used in other automatic learning problems that deal with temporal sequences, as is the case of speech and audio-related tasks [24, 25, 26, 27, 28, 29].

Among the different alternatives for including the attention mechanism into a LSTM framework, in this paper we have adopted the one proposed, among others, by [25] for speech emotion recognition, that is suitable for training in data scarcity conditions. In summary, in contrast to the baseline system where the LSTM output sequence is compacted into an utterance-level representation z by applying an average pooling operation, here z is computed following this equation,

$$z = \sum_{t=1}^{T} \alpha_t y_t \tag{1}$$

where y_t represents the output of the LSTM sequence $y = [y_1, ..., y_T]$ at the time instant t, T is the length of the sequence and α_t is the weight corresponding to the tth LSTM frame. These weights are computed through the following expression,

$$\alpha_t = \frac{exp(u' y_t)}{\sum_T^{t=1} exp(u' y_t)} \tag{2}$$

where $'$ is the transpose operation and u is the attention parameter which is learnt during the training process. The product between u and y_t evaluates the importance of each temporal frame. Subsequently, the weights are normalized by applying a softmax transformation to guarantee that their sum is one.

To implement this method, it is necessary to modify the LSTM network defined in Section 3.2 as shown in Figure 1 where the Attention Mechanism is represented in blue.

3.5. Mixup data augmentation

In order to alleviate the imbalance of the database, we have adopted the Mixup data augmentation technique [30] for increasing the number of samples of the classes other than the majority class (which is class 0, i.e., normal sounds).

In this method, the augmented samples are built through the linear combination of the original examplers of the database. Given two input vectors (x_i, x_j) and their corresponding one-hot encoding labels (w_i, w_j), the new data is built as follows,

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{3}$$

$$\tilde{w} = \lambda w_i + (1 - \lambda) w_j \tag{4}$$

where $\lambda \in [0, 1]$. In the original paper, the mix parameter λ is variable as it is sampled from a beta distribution. Nevertheless, it can be kept constant (see Subsection 4.5).

We have applied the Mixup technique in two different ways:

- **Class-Dependent Mixup (CD Mixup)**. This option takes into account the characteristics of each type of respiratory sound as proposed in [17]. This way, the normal class is considered as a kind of background sound on which adventitious lung sounds (wheeze and/or crackles) are superimposed. Therefore, to obtain new samples of classes 1 (wheeze), 2 (crackle) or 3 (crackle and wheeze), the first input vector in Eq. (3) belongs to class 0 (normal), and the second one belongs to the other class.
- **Class-Independent Mixup (CI Mixup)**. In this option, the two input vectors in Eq. (3) are randomly chosen regardless of the class to which they belong.

4. Experiments

In this Section, we describe the results achieved by applying the techniques described in the previous one.

4.1. Evaluation metrics

To evaluate the different methods, we have used the metrics proposed in [21]: sensitivity (Se), specificity (Sp) and score ($Score$) that are defined in, respectively, Eqs. (5), (6) and (7).

Sensitivity is the ratio between the number of correctly classified samples of crackle ($C_{correct}$), wheeze ($W_{correct}$) and crackle and wheeze ($B_{correct}$) and the total number of samples of these classes (C_{total}, W_{total} and B_{total}).

$$Se = \frac{C_{correct} + W_{correct} + B_{correct}}{C_{total} + W_{total} + B_{total}} \tag{5}$$

Specificity is the ratio between the correctly classified normal samples ($N_{correct}$) and the total number of samples of this class (N_{total}).

$$Sp = \frac{N_{correct}}{N_{total}} \tag{6}$$

Score is the mean of sensitivity and specificity.

$$Score = \frac{Se + Sp}{2} \tag{7}$$

All the experiments have been repeated 10 times. Therefore, results shown in next Subsections correspond to the average of the 10 repetitions performed.

4.2. Baseline results

The baseline LSTM model defined in Subsection 3.2 was trained using the categorical crossentropy as loss function, the Adadelta optimization algorithm with an initial learning rate of 0.1, a batch size of 128 and a maximum number of epochs of 100. The baseline results achieved with this configuration are the following: $Se = 34.36\ \%$, $Sp = 76.12\ \%$ and $Score = 55.24\ \%$. The poor performance in terms of sensitivity can be explained for the imbalance of the database.

4.3. Results with the enhancement of the input features

Table 2 shows the average results achieved by modifying the number of selected filters in the input log-mel spectrograms. In the case of the Score, the standard deviation is also indicated.

Table 2: *Results obtained after selecting the first n filters.*

First n filters	Se	Sp	Score
10	36.77 %	78.30 %	57.53 % ± 1.12 %
20	34.42 %	79.40 %	56.91 % ± 0.61 %
30	36.41 %	75.85 %	56.13 % ± 0.65 %
40	34.36 %	76.12 %	55.24 % ± 0.56 %

As can be observed, the score increases as the number of selected filters decreases, being the baseline case (40 filters) the worst scenario. This result suggests that low frequencies are more relevant than high ones for the discrimination between different breath sounds.

4.4. Results with the Attention Mechanism

Table 3 includes the results obtained when incorporating the Attention Mechanism to the baseline LSTM system. As can be seen, regardless of the number of filters of the input features, the use of attentional networks improves the final score as the specificity increases drastically. However, the sensitivity decreases, which is a negative effect. Again, best results are obtained when considering the 10 first filters of the input vectors.

Table 3: *Results obtained with the Attention Mechanism after selecting the first n filters.*

First n filters	Se	Sp	Score
10	30.40 %	88.07 %	59.24 % ± 0.84 %
20	31.00 %	85.55 %	58.27 % ± 0.37 %
30	29.66 %	85.17 %	57.42 % ± 0.30 %
40	30.05 %	83.94 %	56.99 % ± 0.49 %

4.5. Results with the Mixup data augmentation

In the *CD Mixup* method, we tried to equalize the number of samples of all classes, while keeping unmodified the samples of class 0, obtaining an augmented training set of $8,577$ exemplars. For a fair comparison, the training size set in the *CI Mixup* method was set to the same value.

Table 4 shows the results achieved by the two mixup methods when the mix parameter is either variable ($\lambda =$ Var.) or set to a fixed value ($\lambda = 0.25$). These experiments correspond to the best previous system (10 filters and Attention Mechanism) whose metrics are also included for a better comparison.

As can be seen, both mixup methods improve the results of the system without data augmentation, although the differences are more noticeable for *CI Mixup*. A possible reason is that in the *CD Mixup* method, all the augmented data use samples of the majority class, increasing indirectly the presence of class 0 examples in the training set.

Table 4: *Results obtained with the Attention Mechanism after selecting the first 10 filters and the different variants of Mixup data augmentation methods.*

Method	λ	Se	Sp	Score
No augm.	-	30.40 %	88.07 %	59.24 % ± 0.84 %
CD Mixup	Var.	38.38 %	87.99 %	63.19 % ± 0.57 %
CD Mixup	0.25	42.26 %	80.35 %	61.30 % ± 0.54 %
CI Mixup	Var.	42.47 %	86.15 %	64.31 % ± 0.34 %
CI Mixup	0.25	44.71 %	85.44 %	65.08 % ± 0.35 %

The system that achieves best results is *CI Mixup* with fixed λ. In particular, with respect to the system without data augmentation, it produces an increment in sensitivity and score of 14.31 % and 5.84 % absolute respectively, whereas the specificity suffers a slight decrease of 2.63 % absolute. It is worth noting that the data augmentation technique has partially overcome the problem of data imbalance obtaining an important improvement in terms of sensitivity.

Finally, when comparing to the baseline system (see Subsection 4.2), this configuration improves the sensitivity by 10.35 %, the specificity by 9.32 % and the score by 9.84 % absolute.

5. Conclusions and future lines

In this paper, we have proposed an automatic system capable of classifying respiratory sounds into four different classes (normal, wheeze, crackle, and crackle and wheeze) that is based on Long Short Term-Memory (LSTM) networks with log-mel spectrograms as input features. We have focused our research on three different issues: the enhancement of the input features by means of the empirical selection of the frequency bands containing more useful information for the task at hand; the incorporation of an Attention Mechanism into the LSTM model to modulate the contribution of each audio frame; and the use of a Mixup data augmentation technique to alleviate the problem of data imbalance and improve the sensitivity of the system. The different methods have been evaluated over the publicly available ICBHI 2017 dataset. Experiments have shown that with the use of the 10 first filters of the log-mel spectrograms, the attentional LSTM networks and the *CI Mixup* method with a fixed mix parameter, best results are obtained, improving the sensitivity by 10.35 %, the specificity by 9.32 % and the score by 9.84 % absolute with respect to the baseline system.

For future work, we plan to extend our research in three directions: to analyze automatic methods for the selection of the frequency bands more suitable for the discrimination of lung sounds, to analyze the relationship between the emphasized frames by the Attention Mechanism and some acoustic properties of the respiratory sounds, and to study the use of alternative data augmentation techniques.

6. Acknowledgements

The authors acknowledge support from the Spanish State Research Agency (MCIN/AEI/10.13039/5011000110) through project PID2020-115363RB-I00.

7. References

[1] A. Brunier and A. Muchnik, "WHO reveals leading causes of death and disability worldwide: 2000-2019", 2020. [Online]. Available: https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 (accessed: July 24, 2022).

[2] World Health Organization, "The top 10 causes of death", Dec. 09, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed: July 24, 2022).

[3] N. Chen 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, Feb. 2020, doi: 10.1016/S0140-6736(20)30211-7.

[4] A. Orús, "Number of cumulative cases of coronavirus (COVID-19) worldwide from January 22, 2020 to July 13, 2022", Statista. [Online]. Available: https://www.statista.com/statistics/1103040/cumulative-coronavirus-covid19-cases-number-worldwide-by-day/ (accessed: July 24, 2022).

[5] A. Orús. "Number of novel coronavirus (COVID-19) deaths worldwide as of July 15, 2022", Statista. [Online]. Available: https://www.statista.com/statistics/1093256/novel-coronavirus-2019ncov-deaths-worldwide-by-country/ (accessed: July 24, 2022).

[6] N. Meslier, G. Charbonneau, and J. L. Racineux, "Wheezes", vol. 8, no. 11, pp. 1942–1948, Nov. 1995, doi: 10.1183/09031936.95.08111942.

[7] P. Piirila and A. R. Sovijarvi, "Crackles: recording, analysis and clinical significance", vol. 8, no. 12, pp. 2139–2148, Dec. 1995, doi: 10.1183/09031936.95.08122139.

[8] S. Mangione and L. Z. Nieman, "Pulmonary auscultatory skills during training in internal medicine and family practice", vol. 159, no. 4, pp. 1119–1124, 1999, doi: 10.1164/ajrccm.159.4.9806083.

[9] M.-L. Hsueh et al, "Respiratory wheeze detection system", IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005, doi: 10.1109/iembs.2005.1616260.

[10] S. Ulukaya, G. Serbes, I. Sen and Y. P. Kahya, "A lung sound classification system based on the rational dilation wavelet transform", 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, doi: 10.1109/embc.2016.7591542.

[11] H. Yamamoto, S. Matsunaga, M. Yamashita, K. Yamauchi and S. Miyahara, "Classification between normal and abnormal respiratory sounds based on stochastic approach", International Congress on Acoustics (ICA), 2010.

[12] N. Sengupta, M. Sahidullah and G. Saha, "Lung sound classification using cepstral-based statistical features", Computers in Biology and Medicine, vol. 75, pp. 118-129, 2016, doi: 10.1016/j.compbiomed.2016.05.013.

[13] N. Jakovljevic and T. Loncar-Turukalo, "Hidden Markov Model based respiratory sound classification", ICBHI: Precision Medicine Powered by pHealth and Connected Health, Thesaloniki, Greece, p. 39–43, 2017.

[14] R. Palaniappan, K. Sundaraj and S. Sundaraj, "A comparative study of the SVM and KNN machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals", BMC Bioinformatics, vol. 15, no. 1, pp. 223, Jun. 2014, doi: 10.1186/1471-2105-15-223.

[15] M. Lozano, J. A. Fiz and R. Jané, "Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode eecomposition and instantaneous frequency", IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 2, pp. 486-497, 2016.

[16] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based on spectral analysis methods", Precision Medicine Powered by pHealth and Connected Health, Springer Singapore, p. 45-49, Nov. 2017, doi: 10.1007/978-981-10-7419-6_8.

[17] Y. Ma, 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", Proc. of Interspeech 2020, 2020.

[18] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning", IEEE Transactions on Biomedical Circuits and Systems, 2020, doi: 10.1109/tbcas.2020.2981172.

[19] D. Perna and A. Tagarelli, "Deep auscultation: predicting respiratory anomalies and diseases via recurrent neural networks", IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 50-55, 2019.

[20] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen and R. Palaniappan, "Robust deep learning framework for predicting respiratory anomalies and diseases", arXiv 2002.03894, 2020.

[21] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms", vol. 40, no. 3, p. 035001, Mar. 2019, doi: 10.1088/1361-6579/ab03ea.

[22] B. Rocha, D. Pessoa, A. Marques, P. Carvalho, and R. P. Paiva, "Automatic classification of adventitious respiratory sounds: a (un)solved problem?", Dec. 2020, doi: 10.3390/s21010057.

[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[24] J. Chorowski, D. Bahdanau, D. Dzmitry, D.Serdyuk, K. Cho and Y. Bengio, "Attention-based models for speech recognition", Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, pp. 577–585, 2015.

[25] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention", ICASSP 2017, pp. 2227–2231, 2017.

[26] M. Fernández-Díaz and A. Gallardo-Antolín, "An attention Long Short-Term Memory based system for automatic classification of speech intelligibility", Engineering Applications of Artificial Intelligence, vol. 96, Nov. 2020, doi: 10.1016/j.engappai.2020.103976.

[27] N. Zacarias-Morales, P. Pancardo, J. A. Hernández-Nolasco and M. Garcia-Constantino, "Attention-inspired artificial neural networks for speech processing: a systematic review", Symmetry, 13(2):214, 2021.

[28] A. Gallardo-Antolín and J. M. Montero, "On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification", Neurocomputing, vol. 456, pp. 49-60, 2021, doi: 10.1016/j.neucom.2021.05.065.

[29] A. Gallardo-Antolín and J. M. Montero, "An auditory saliency pooling-based LSTM model for speech intelligibility classification", Symmetry, 13, no. 9: 1728, 2021, doi: 10.3390/sym13091728.

[30] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "*mixup*: beyond empirical risk minimization", International Conference on Learning Representations (ICLR), 2018.