# Pediatric Respiratory Sound Classification Using a Dual Input Deep Learning Architecture: The IEEE BioCAS 2023 Grand Challenge

Diogo Pessoa[1], Georgios Petmezas[2], Vasileios E. Papageorgiou[3], Bruno M. Rocha[1], Leandros Stefanopoulos[2], Vassilis Kilintzis[2], Nicos Maglaveras[2], Inéz Frerichs[4], Paulo de Carvalho[1], and Rui Pedro Paiva[1]
[1]University of Coimbra Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, 3030-290 Coimbra, Portugal
[2]2nd Department of Obstetrics and Gynaecology, The Medical School, 54124 Thessaloniki, Greece
[3]Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[4]Department of Anesthesiology, and Intensive Care Medicine, University Medical Center Schleswig-Holstein Campus Kiel, Kiel 24105, Schleswig-Holstein, Germany

***Abstract*—Patients with respiratory conditions typically exhibit adventitious respiratory sounds (ARS), such as wheezes and crackles. In recent years, computerized methods for analyzing respiratory function, namely ARS, have gained increased attention within the scientific community. Such methods primarily aim to facilitate diagnosing and monitoring patients suffering from respiratory diseases. In this work, we propose a deep learning model for the automatic classification of respiratory sounds within the proposed tasks of the "IEEE BioCAS 2023 Grand Challenge on Respiratory Sound Classification". The model was based on a dual input convolutional deep learning architecture, using the raw audio signal and the short-time Fourier transform (STFT) spectrogram as inputs. Our model obtained a challenge total score of 0.590 (Task 1-1: 0.756; Task 1-2: 0.467; Task 2-1: 0.658; Task 2-2: 0.458).**

***Index Terms*—Respiratory sound classification, Adventitious respiratory sounds, Respiratory diseases, Deep learning**

## I. Introduction

Respiratory diseases are among the leading causes of death worldwide, affecting people in multiple aspects of their lives [1]. Such diseases are responsible for a substantial strain on health systems and significantly impact the quality of life of people suffering from them [2]. Early diagnosis through lung auscultation can help limit the impact of respiratory diseases on patients, especially children and adolescents. To date, auscultation is one of the main tools clinicians use to analyze respiratory function. Respiratory sounds reveal significant information concerning the physiology of the lungs and any potential airway obstacles [3]. When performing auscultation, clinicians usually look up for the presence of adventitious respiratory sounds (ARS). These are additional respiratory sounds superimposed on normal respiratory sounds, and their presence is generally suggestive of a respiratory disorder [3]–[5]. ARS can be continuous (e.g., wheezes) or discontinuous (e.g., crackles), and their characteristics (such as timing in the respiratory cycle and frequency) are of great clinical relevance [4]–[6]. Despite auscultation's extensive adoption, its subjectivity is widely recognized, which has led to a new era of developments in computerized techniques for acquiring and analyzing respiratory sounds [4].

In recent decades, many machine learning algorithms have been developed for respiratory sound classification [7]. Despite the significant number of works published in the literature over the years, the automatic classification of ARS is still a challenging problem yet to be solved. Most previous works listed in [7] relied on small datasets, typically designed for teaching purposes. Thus, the models might not generalize in more challenging scenarios (i.e., with new external subjects or recordings with background noise). As a result, they might also be prone to overfitting, while their predictive efficiency is not extensively examined.

Bardou et al. [8] was one of the first works in which CNN models were proposed for the static classification of ARS (at the event level), using spectrogram, MFCCs, and LBPs as inputs. Li et al. [9] proposed ResNet-based architectures for deep feature extraction from spectrograms followed by a fully connected layer for ARS classification (at event and record level). Zhang et al. [10] proposed a feature polymerized-based two-level ensemble model (FP-TLEM) for respiratory sound classification (at event and record level). They have extracted several handcrafted features of different domains and employed the Synthetic Minority Oversampling Technique for data augmentation. Rocha et al. [11] proposed a dual-input deep learning architecture, with spectrograms and mel-spectrograms, for the static classification of ARS at the event level (wheezes, crackles, and normal). In another study by the same authors, several models based on the same architecture

This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020, by FCT Ph.D. scholarships (DFA/BD/4927/2020 and SFRH/BD/135686/2018), and by the Horizon 2020 Framework Programme of the European Union project WELMO (under grant agreement number 825572).

Corresponding author: D. Pessoa (email: dpessoa@dei.uc.pt).

were developed for the static classification of ARS at record level based on demographic characteristics (such as age, sex, BMI) [12].

In this work, we introduced a hybrid convolutional deep learning model consisting of two branches, each receiving a different input: the raw respiratory sound and spectrogram (time-frequency representation). The two branches are then fused to make the final model prediction. The models were trained to classify pediatric respiratory sounds at event and record levels (based on the IEEE BioCAS 2023 Grand Challenge tasks [1]).

The rest of the article is organized as follows: in section II, we describe the dataset and the proposed approaches; in section III we present and discuss the obtained results; lastly, in section IV we conclude and suggest possible directions for future work.

## II. Materials and Methods

This section describes the data used for this study and the proposed methodological framework. Figure 1 presents the main steps involved in the preprocessing and classification of ARS. To process the respiratory sounds and train all deep learning models, we used Python 3.8, Keras, and TensorFlow. The models were trained on an NVIDIA RTX 2060 Super with 8 GB of RAM. The computer was also equipped with an AMD Ryzen 9 3900X 3.8 GHz and 64 GB of RAM.

### A. Dataset

In the present study, we used the "SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database" [13], [14]. Respiratory sounds were collected from the pediatric respiratory department at Shanghai Children's Medical Center (SCMC). The sounds were recorded using a Yunting model II Stethoscope with a sampling frequency of 8000 Hz and involved a pediatric population aged between 1 month and 18 years old, with a mean age of 6.7 years. Also, each audio sample has been manually annotated by medical experts at event and record levels. At the event level, each recording has been segmented into multiple respiratory events annotated as Normal, Rhonchi, Wheeze, Stridor, Coarse Crackle, Fine Crackle, or Wheeze + Crackle. On the other hand, at the record level, following an initial separation based on the signal quality (poor and high), the high-quality recordings have been further annotated as Normal, CAS (continuous adventitious sound), DAS (discontinuous adventitious sound), or CAS & DAS according to the presence or absence of continuous or discontinuous ARS. Table I overviews the annotations at both event and record levels (training set).

The IEEE BioCAS 2023 Grand Challenge on Respiratory Sound Classification involves two main tasks: Task 1 (respiratory sound classification at event level) and Task 2 (respiratory sound classification at record level). Furthermore, each task is also divided into two sub-tasks. Task 1-1 includes binary classification into Normal and Adventitious events, and Task 1-2 refers to a multi-class classification challenge including, in total, seven classes. Regarding the record classification task, Task 2-1 is a ternary class classification challenge, including Normal, Adventitious, and Poor Quality records. Task 2-2 is a multi-class classification challenge comprising five classes. It is also worth noting that the challenge contained an external testing set used for the independent evaluation of models.

[1]https://2023.ieee-biocas.org/grand-challenge

TABLE I
The Grand Challenge Dataset (training set).

| | Class | Count |
|---|---|---|
| Events (Task 1) | Normal | 5159 |
| | Fine Crackle | 912 |
| | Wheeze | 452 |
| | Coarse Crackle | 49 |
| | Rhonchi | 39 |
| | Wheeze + Crackle | 30 |
| | Stridor | 15 |
| | **Total** | **6656** |
| Records (Task 2) | Normal | 1303 |
| | DAS | 248 |
| | Poor Quality | 177 |
| | CAS | 134 |
| | CAS & DAS | 87 |
| | **Total** | **1949** |

### B. Preprocessing and feature extraction

All the respiratory sounds were recorded with a sampling frequency of 8000 Hz. Nonetheless, as the typical frequency of interest of respiratory sound lies below 2000 Hz [3], we re-sampled all recordings at 4000 Hz.

Regarding Task 1-1 and Task 1-2, recordings have been segmented into distinct events based on the experts' annotations of each event. The duration of the events ranged from 0.126 s to 7.152 s, with an average duration of 1.278 s. Therefore, we established a maximum time of 7.152 seconds for the samples to be used as the inputs of our classification models and applied zero padding to the shorter samples (center zero padding). In Task 2, most of the respiratory recordings (1429) had a duration of 9.22 s, while about a quarter of the recordings (518) had a duration of 15.36 s. Two more recordings with a different duration have been detected, one lasting 0.3 s and one 8.51 s. Once again, a similar strategy has been used to deal with this issue; first, a fixed recording duration of 15.36 s has been selected. Then, all recordings with a shorter duration have been expanded using zero padding (center zero padding).

After the audio preprocessing, we computed the short-time Fourier transform (STFT) spectrogram of the respiratory sounds in all tasks and used it as input for the deep learning models, together with the raw audios. The spectrogram is one of the most commonly used tools in audio analysis and processing, since it describes the evolution of the frequency components over time. The STFT spectrogram ($X(n,\omega)$) of a given discrete signal ($x[n]$) is given by:

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega m} \quad (1)$$

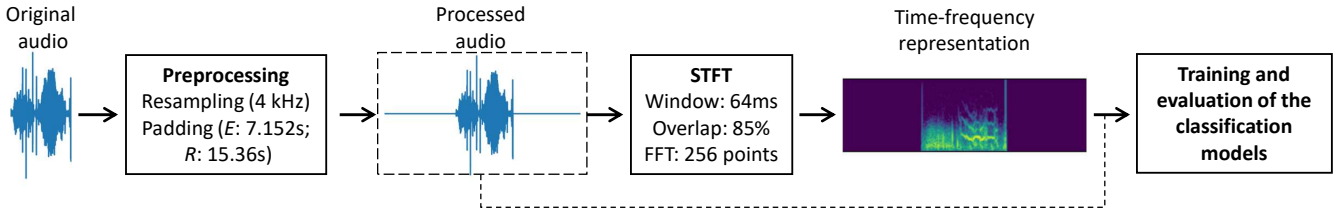


Fig. 1. Overall framework (*E*: maximum length of events, Tasks 1-1 and 1-2; *R*: maximum length of complete recordings, Tasks 2-1 and 2-2; the dashed line represents the combination of both inputs for the deep learning models - audio and spectrogram).

where $x[m]w[n-m]$ is a short-time section of $x[m]$ at time $n$, and $w[n]$ is a window function centered at instant $n$ [15]. To compute the STFT spectrogram we used a 64 ms Blackman–Harris window with 85% overlap. For the Fast Fourier Transform (FFT), 256 points were used [16]. Both inputs' samples, raw audio signals and spectrograms, were normalized between 0 and 1.

### C. Model architecture and training

In this work, we had as primary goal the development of models to automatically classify respiratory sounds at event (previously segmented) and recording levels. To do so, we developed deep learning models for both classification tasks that leveraged the use of multiple inputs, using the raw respiratory audio signal and the according STFT time-frequency representation. With the use of multiple inputs from different time domains, namely the time-frequency (STFT) and time (raw audio) domains, we intended to provide complementary information for the model to learn and be able to classify the different classes. Figure 2 presents a block diagram representation of the proposed model architecture.

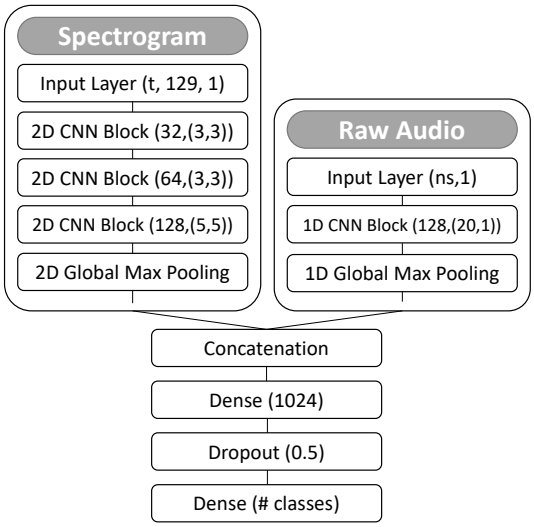


Fig. 2. Block diagram representation with the architecture of the deep learning model (t - number of time steps; ns - number of audio samples; CNN Block parameter 1 - number of filters; CNN Block parameter 2 - kernel size).

Our model was composed of two main ramifications: the spectrogram input's convolution module and the raw audio's convolution module (Figure 2). The convolution module of the spectrogram input consisted of three CNN blocks. Each block was composed of a 2D convolutional layer [17], a LeakyReLU activation layer (with $\alpha = 0.1$), a 2D max-pooling (pool size = 2), and, lastly, a dropout layer (dropout percentage = 0.5). The convolution module of the raw audio input consisted of only one CNN block. The block consisted of a 1D convolutional layer, a LeakyReLU activation layer (with $\alpha = 0.1$), a 1D max-pooling (pool size = 2), and, lastly, a dropout layer (dropout percentage = 0.5). After the convolutional modules of both inputs, 2D and 1D global maximum pooling layers [18] were applied with the subsequent concatenation of the extracted features from both sources. The concatenated features were then fed to a fully connected layer with 1024 neurons. Lastly, the model had a dense layer for the respective output according to each task (Task 1.1 - 2; Task 1.2 - 7; Task 2.1 - 3; Task 2.2 - 5).

We created five splits to validate our models in a subject hold-out fashion (with subject independence). In each split, we randomly selected approximately 80% of the subjects from each class to train the models, while the remaining 20% were kept for validation. The splits were performed independently for each of the four considered tasks (Task 1-1, Task 1-2, Task 2-1, and Task 2-2) and in a stratified manner, maintaining the same class ratio as the original training dataset. It is worth noting that by isolating the subjects, we ensure no data leakage between the training and validation sets.

The models were trained for 100 epochs for each task validation split. We used a learning rate of 3e-4 with the Adam optimizer and the categorical cross-entropy as the loss function. Simultaneously to the training process, the models were evaluated using the validation subset at every new epoch to save only the set of weights with the lowest validation loss. Then, we selected, for each task, the model with the highest score (see subsection II-D) based on the results obtained for the different splits. The best model out of the five for each task was then used in the external testing set.

### D. Evaluation metrics

To evaluate the performance of respiratory classification we used sensitivity, specificity, average score, harmonic score and task score. For Task 1, the equations for sensitivity (SE) and specificity (SP) are presented below:

$$Sensitivity(SE) = \frac{R_r + W_w + CC_{cc} + FC_{fc} + WC_{wc}}{R_t + W_t + CC_t + FC_t + WC_t}, \quad (2)$$

TABLE II
RESULTS SUMMARY TABLE (TRAINING AND TEST/CHALLENGE). SE - SENSITIVITY ; SP - SPECIFICITY ; AS - AVERAGE SCORE; HS - HARMONIC SCORE; ACC - ACCURACY; TS - CHALLENGE TASK SCORE; SD - STANDARD DEVIATION.

| | **SE** | **SP** | **AS** | **HS** | **ACC** | **TS** |
|---|---|---|---|---|---|---|
| Task 1-1 (training - mean ± sd) | 0.75 ± 0.02 | 0.95 ± 0.01 | 0.85 ± 0.01 | 0.84 ± 0.01 | 0.90 ± 0.01 | - |
| Task 1-1 (best run training) | 0.77 | 0.96 | 0.86 | 0.85 | 0.90 | - |
| Task 1-1 (challenge score) | 0.67 | 0.86 | 0.76 | 0.74 | - | 0.756 |
| Task 1-2 (training - mean ± sd) | 0.58 ± 0.05 | 0.93 ± 0.01 | 0.76 ± 0.03 | 0.71 ± 0.04 | 0.82 ± 0.02 | - |
| Task 1-2 (best run training) | 0.63 | 0.93 | 0.78 | 0.75 | 0.82 | - |
| Task 1-2 (challenge score) | 0.34 | 0.64 | 0.49 | 0.44 | - | 0.467 |
| Task 2-1 (training - mean ± sd) | 0.33 ± 0.09 | 0.89 ± 0.03 | 0.61 ± 0.04 | 0.47 ± 0.10 | 0.68 ± 0.04 | - |
| Task 2-1 (best run training) | 0.46 | 0.87 | 0.66 | 0.60 | 0.73 | - |
| Task 2-1 (challenge score) | 0.61 | 0.70 | 0.66 | 0.66 | - | 0.658 |
| Task 2-2 (training - mean ± sd) | 0.34 ± 0.04 | 0.78 ± 0.04 | 0.56 ± 0.03 | 0.48 ± 0.05 | 0.61 ± 0.03 | - |
| Task 2-2 (best run training) | 0.38 | 0.84 | 0.61 | 0.52 | 0.65 | - |
| Task 2-2 (challenge score) | 0.46 | 0.46 | 0.46 | 0.46 | - | 0.458 |

where $R_r$, $W_w$, $CC_{cc}$, $FC_{fc}$ and $WC_{wc}$ the number of correctly predicted events of each class, and $R_t$, $W_t$, $CC_t$, $FC_t$ and $WC_t$ the total number of Rhonchis, Wheezes, Coarse Crackles, Fine Crackles and Wheezes+Crackles, respectively, and

$$Specificity(SP) = \frac{N_n}{N}, \tag{3}$$

where $N_n$ denotes the number of correctly predicted normal events, and $N$ the total number of Normal events, respectively.

For Task 2, the equations for SE and SP are defined below:

$$Sensitivity(SE) = \frac{C_c + D_d + CD_{cd}}{C + D + CD}, \tag{4}$$

where $C_c$, $D_d$ and $CD_{cd}$ the number of correctly predicted recodrs, and $C$, $D$ and $CD$ the total number of CAS, DAS and CAS & DAS records, respectively, and

$$Specificity(SP) = \frac{N_n}{N}, \tag{5}$$

where $N_n$ the number of correctly predicted records, and $N$ the total number of Normal records, respectively.

For both tasks, the Average Score (AS), Harmonic Score (HS) and Task Score (TS) are defined as follows:

$$AverageScore(AS) = \frac{SE + SP}{2} \tag{6}$$

$$HarmonicScore(HS) = \frac{2 * SE * SP}{SE + SP} \tag{7}$$

$$TaskScore(TS) = \frac{AS + HS}{2} \tag{8}$$

## III. RESULTS

As discussed in subsection II-C, we have chosen the best model from the five validation splits and used it in the testing set (challenge result). Table II presents the obtained results regarding the five validation splits for all tasks (with the respective mean and standard deviation). Moreover, it also presents the results for the best model of the training runs in each task, as well as the respective task score obtained in the challenge testing set. From the analysis of Table II, we observe that the results obtained on the testing set by the best model are generally on pair with those obtained in validation. Our models have obtained a total score of 0.590 in the IEEE BioCAS 2023 Grand Challenge (Total Score = 0.2 * Score 1-1 + 0.3 * Score 1-2 + 0.2 * Score 2-1 + 0.3 * Score 2-2).

While our current approach demonstrates promising results, it is vital to acknowledge certain limitations that warrant consideration. Firstly, the reliance on a single best model may inadvertently introduce an element of model selection bias, potentially skewing the perception of overall performance. Exploring ensemble techniques or model averaging could potentially address this concern and provide a more comprehensive evaluation of the model's capabilities.

Our model provides encouraging results considering Task 1-1 and Task 2-2, which represent binary and ternary classification problems, respectively. Therefore, an attempt to improve the performance of our approach concerning the tasks corresponding to multi-class classification remains an open challenge for us. Considering this, data augmentation techniques must also be considered to increase the number of available samples and balance the different classes. Preprocessing steps, such as feature selection based on causality measures or the inclusion of additional inputs deriving for the time and/or frequency domain, can enhance predictive efficiency or even reduce the computational burden. The study of different inputs for the classification models must also be studied.

## IV. CONCLUSION

In this work, we tested several approaches on the tasks of the IEEE BioCAS Grand Challenge on Respiratory Sound Classification. The best results were achieved by the model with a dual input architecture, providing satisfactory diagnostic performance in all the investigated tasks. For future work, we intend to add a recurrent module to the current model to capture time dependencies. Additionally, we aim to implement other loss functions, such as focal loss, to deal with data imbalance [19]. Finally, we will study the use of different deep learning architectures and data augmentation methodologies to compensate further for the uneven class distribution.

## REFERENCES


[1] F. of International Respiratory Societies, *The Global Impact of Respiratory Disease – Second Edition*. 2017.

[2] G. J. Gibson, R. Loddenkemper, B. Lundbäck, and Y. Sibille, "Respiratory health and disease in Europe: The new European Lung White Book," *European Respiratory Journal*, vol. 42, no. 3, pp. 559–563, 2013.

[3] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of Respiratory Sounds: State of the Art," *Clin. Med. Circ. Respirat. Pulm. Med.*, vol. 2, p. CCRPM.S530, jan 2008.

[4] A. Marques and A. Oliveira, *Normal Versus Adventitious Respiratory Sounds*, ch. 10, pp. 181–206. Cham: Springer International Publishing, 2018.

[5] D. Pessoa, B. M. Rocha, P. de Carvalho, and R. P. Paiva, "Automated respiratory sound analysis," in *Wearable Sensing and Intelligent Data Analysis for Respiratory Management*, pp. 123–168, Elsevier, 2022.

[6] D. Pessoa, B. M. Rocha, C. Strodthoff, M. Gomes, G. Rodrigues, G. Petmezas, G.-A. Cheimariotis, V. Kilintzis, E. Kaimakamis, N. Maglaveras, A. Marques, I. Frerichs, P. de Carvalho, and R. P. Paiva, "Bracets: Bimodal repository of auscultation coupled with electrical impedance thoracic signals," *Computer Methods and Programs in Biomedicine*, p. 107720, 2023.

[7] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PLOS ONE*, vol. 12, pp. 1–43, 05 2017.

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

[9] J. Li, X. Wang, X. Wang, S. Qiao, and Y. Zhou, "Improving the resnet-based respiratory sound classification systems with focal loss," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 223–227, 2022.

[10] L. Zhang, Y. Zhu, S. Tu, and L. Xu, "A feature polymerized based two-level ensemble model for respiratory sound classification," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 238–242, 2022.

[11] B. M. Rocha, D. Pessoa, A. Marques, P. Carvalho, and R. P. Paiva, "Automatic classification of adventitious respiratory sounds: A (un)solved problem?," *Sensors*, vol. 21, no. 1, 2021.

[12] T. Fernandes, B. M. Rocha, D. Pessoa, P. de Carvalho, and R. P. Paiva, "Classification of adventitious respiratory sound events: A stratified analysis," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 01–05, 2022.

[13] Q. Zhang, J. Zhang, J. Yuan, H. Huang, Y. Zhang, B. Zhang, G. Lv, S. Lin, N. Wang, X. Liu, M. Tang, Y. Wang, H. Ma, L. Liu, S. Yuan, H. Zhou, J. Zhao, Y. Li, Y. Yin, L. Zhao, G. Wang, and Y. Lian, "Sprsound: Open-source sjtu paediatric respiratory sound database," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 867–881, 2022.

[14] Q. Zhang, J. Zhang, J. Yuan, H. Huang, Y. Zhang, B. Zhang, G. Lv, S. Lin, N. Wang, X. Liu, M. Tang, Y. Wang, H. Ma, L. Liu, S. Yuan, H. Zhou, J. Zhao, Y. Li, Y. Yin, L. Zhao, G. Wang, and Y. Lian, "Grand challenge on respiratory sound classification for sprsound dataset," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 213–217, 2022.

[15] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Pearson Education, 2002.

[16] V. E. Papageorgiou, T. Zegkos, G. Efthimiadis, and G. Tsaklidis, "Analysis of digitalized ecg signals based on artificial intelligence and spectral analysis methods specialized in arvc," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 38, no. 11, p. e3644, 2022.

[17] V. E. Papageorgiou, P. Dogoulis, and D. P. Papageorgiou, "A convolutional neural network of low complexity for tumor anomaly detection," *Proceedings of Eighth International Congress on Information and Communication Technology*, vol. 4, 2023.

[18] V. Papageorgiou, "Brain tumor detection based on features extracted and classified using a low-complexity neural network," *Traitement du signal*, vol. 38, no. 3, pp. 547–554, 2021.

[19] G. Petmezas, G.-A. Cheimariotis, L. Stefanopoulos, B. Rocha, R. P. Paiva, A. K. Katsaggelos, and N. Maglaveras, "Automated lung sound classification using a hybrid cnn-lstm network and focal loss function," *Sensors*, vol. 22, no. 3, 2022.