Exploring Model Architectures for Real-Time Lung Sound Event Detection

Michiel Jacobs[1,2,3,4] , Lode Vuegen[1,2,3] , Tom Verresen[5], Marie Schouterden[5], David Ruttens[4,5] and Peter Karsmakers[1,2,3] *

1 - KU Leuven, Dept. of Computer Science, Kleinhoefstraat 4, 2440 Geel, Belgium
2 - Flanders Make @ KU Leuven
3 - Leuven.AI - KU Leuven Institute for AI
4 - Hasselt University, Fac. of Medicine & Life Sciences, 3500 Hasselt, Belgium
5 - Ziekenhuis Oost-Limburg, Dept. of Pulmonary Medicine, 3600 Genk, Belgium
Corresponding authors: {michiel.jacobs, peter.karsmakers}@kuleuven.be

Abstract. Computerized detection of relevant lung sound events has the potential to assist physicians during auscultation and to monitor the severity of pulmonary diseases in ambulatory settings. In some cases, real-time detection of adventitious lung sounds is required to provide instant feedback to physicians, e.g. during autogenic drainage therapy. State-of-the-art solutions for this task leverage deep learning models, which vary significantly in complexity. For real-time applications on resource-constrained devices, such as stethoscope-integrated hardware, both detection accuracy and model complexity are important to consider. While most existing research focusses primarily on accuracy, this work evaluates both accuracy and computational complexity. The contributions of this work are threefold. First, the effect of using a full breathing cycle as input is studied to assess its impact on event detection performance. This approach introduces a computational cost due to the required segmentation process. Second, a transformer-based architecture is compared with two relatively simple convolutional models, each utilizing different input horizons. Evaluations are conducted on both public and in-house lung sound datasets. Third, recognizing that the event detection task aligns better with a multi-label setting than the commonly used multi-class setup, this study compares both approaches. We conclude that a multi-label output outperforms a multi-class approach, that inputs segmented per breathing cycle are preferred, and that the high complexity models have similar performance to the models with low complexity on unseen data.

The source code is available through this GitHub repository.

1 Introduction

In recent years, there has been a growing trend in using digital and wearable stethoscopes for auscultation. These devices allow for the convenient storage of recorded lung sounds from both hospital and ambulatory environments directly into electronic health records [1]. Furthermore, when integrated with a real-

*The authors would like to thank Flanders Innovation & Entrepreneurship Agency (VLAIO) for providing funding for PlugNPatch (3E230186) and KU Leuven IOF for funding COMPASS (3H230337)

time computerized lung Sound Event Detection (SED) system, they have the potential to instantly assist pulmonologists in making their diagnoses [2].

Multiple types of adventitious lung sound events exist, which can be divided into two groups: continuous and discontinuous events [3]. An example of a continuous event is wheezing, which has frequencies in the range 100 to 5000 Hz and typically lasts more than 100 ms [3, 4]. Crackles are discontinuous events characterized as impulsive sounds and typically last between 5 to 15 ms [4].

The authors of [5] evaluated the use of the recently proposed Audio Spectrogram Transformer (AST) [6] to determine whether a manually segmented breathing cycle contains a crackle, a wheeze, both or if it can be considered as normal. The AST was compared with traditional convolutional networks such as CNN6 [7, 8], EfficientNet, and ResNet50. The obtained results indicate that AST with pre-training on ImageNet and AudioSet obtains the highest ICBHI score (balanced accuracy), which was equal to 62.37%. The results for EfficientNet, ResNet50, and CNN6 were 56.56%, 56.85% and 55.24% resp.

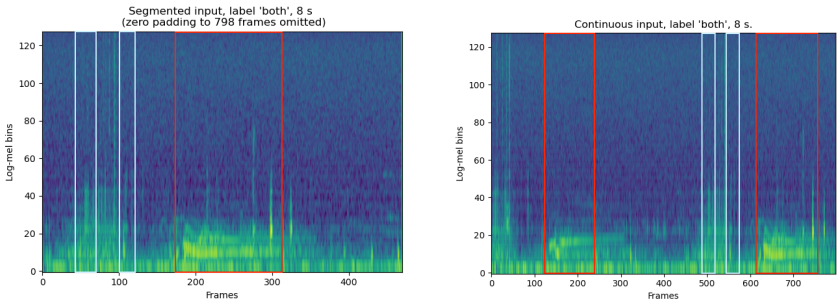
Although the results from [5] are promising, further research is necessary to develop a viable solution for clinical settings. One challenge lies in the requirement for segmented breathing cycles, which adds an additional layer of processing to the lung sound data. Accurately estimating breathing cycles is not only challenging, but also adds computational complexity and is prone to errors. In addition, the proposed AST model architecture is computationally demanding, making it unsuitable for resource-constrained processing devices that may be integrated into the stethoscope. Moreover, while lung sound classification is often treated as a multi-class problem in the literature, it is inherently a multi-label problem, as lung sound events are not mutually exclusive [4].

The main contributions of this work are: (1) the study of the effect of using a full breathing cycle as input to assess its impact on event detection performance, (2) the comparison of a state-of-the-art transformer-based architecture with two relatively simple convolutional models in terms of detection performance and computational complexity. In case of continuous audio input, models are also evaluated for their generalizability on a second, independent in-house lung sound dataset, (3) recognizing that the event detection task aligns better with a multi-label setup.

2 Methodology

2.1 Data

The public ICBHI dataset [3] contains respiratory sounds and was first presented at the International Conference on Biomedical Health Informatics (ICBHI) in 2017. In total, 126 participants were recorded to create 920 recordings and two sets of annotations. The sounds were recorded with four unique stethoscopes, each with its own sampling rate. The dataset contains annotated breathing cycles, crackle and wheeze events, and information about participants' respiratory disease. The ICBHI dataset has approximately 5.5 h of audio, and 6,898 annotated breathing cycles. The data are divided into official training (60%) and



(a) Segmented (b) Continuous

Fig. 1: Breathing cycle segmented and continuous model input for AST and CNN6 models. Wheezes are annotated in red and crackles in white boxes.

test (40%) sets based on the patient IDs.

The in-house ZOL dataset was collected at the Ziekenhuis Oost-Limburg (ZOL) hospital and contains 46 1-minute auscultation recordings originating from three Chronic Obstructive Pulmonary Disease (COPD) patients. Data were collected using a Littmann 3200 digital stethoscope, which is one of the stethoscopes also used in the ICBHI dataset. Auscultation positions were the following: anterior left & right, posterior left & right, and right-hand side of the neck. Annotations were obtained by a majority vote over 3 trained annotators. As this dataset does not have breathing cycle annotations, it can only be used to evaluate models trained on continuous audio.

2.2 Pre-Processing

In order to ease the comparison, our work used the same pre-processing as in [5]. First, the audio is resampled to 16 kHz. Second, spectrograms are created using a Short-Time Fourier Transform (STFT) with 25 ms windows and 10 ms steps. The Hann window function is applied. Third, the spectrogram is converted to the log-mel domain by applying a mel-filterbank of length 128. The spectrograms always have a time dimension of 8 seconds, as this is the duration of the longest breathing cycle segment. For shorter segments, the spectrogram is zero padded until 798 frames (8 s). For continuous audio, the entire 8 second window is filled with audio using a 50% overlap between windows. In contrast to [5], the training samples were not augmented, and breathing cycles shorter than 8 s were zero padded instead of repeated. Figures 1a and 1b show the segmented and continuous input resp.

2.3 Evaluation

The ICBHI Challenge [3] proposed balanced accuracy as evaluation metric (referred to as "ICBHI score"), which is the average of true positive and true negative rates. Evaluation metrics are averaged over three runs which use the same validation set and which have random seeds for reproducible model initialisation. In order to make a fair comparison with a multi-class setup, the

multi-label predictions are scored by aggregating metrics from 4 binary comparisons, i.e. normal-vs-rest, crackle-vs-rest, wheeze-vs-rest and both-vs-rest.

Both [5] and [7] used the official test set for tuning hyperparameters, as well as evaluating the final model. To keep the ICBHI official test set independent from the data used during training, in this work a validation set was sampled from the official training set without replacement (80% train, 20% val.). Data from a single patient is either fully in the training or validation set.

Since patients 156 and 218 occur both in the official training set and in the official test set, these recordings are added to the validation set. In this way, no weights are fitted to these test-set patients.

2.4 Model Architectures and Training

For model training, Adam optimiser and a cosine learning rate (LR) schedule were used. For multi-class output, the cross-entropy (CE) loss was used with 4 output neurons ("normal", "crackle", "wheeze" & "both"). For multi-label output, binary CE loss was used with 2 output neurons ("crackle" & "wheeze"). Class weights were applied to accomodate for the class imbalance. For AST and CNN6, training was done in two stages. In the first stage, pre-trained weights were loaded and only the classifier was trained without cosine LR schedule. In the second stage, the model with the lowest validation loss was entirely trained and cosine LR schedule was applied. A grid search was performed to find the optimal combination of learning rate and weight decay parameter. The weights were saved when the model reached the lowest validation loss. Details can be found in Table 1.

The baseline model is composed of four convolutional blocks having 32, 32, 64 and 128 filters resp., all of size 5×5. Max pooling with dimension 2×2 is always performed in between convolution layers. Next, three fully-connected layers of 64, 32 and 32 neurons are applied. ReLU activation is always used, as well as batch normalisation. Dropout (50%) is only applied on the convolutional layers. For this baseline model, the multi-label decision thresholds were tuned using the point closest in terms of Euclidean distance to coordinates (0, 1) on the ROC curve. The input to the baseline spans 0.5 s instead of 8 s, as the goal is to deploy this model on an edge device and in real-time.

Model	Initialisation	Classifier only		Full model		Dropout rate	Batch size	Weight decay
		Learning rate	Epochs	Learning rate	Epochs			
AST MC-S	IN + AS	1e-3	30	5e-7	50	n/a	16	1e-2
AST ML-S	IN + AS	1e-3	30	1e-6	50	n/a	16	1e-4
CNN6 MC-S	AS	1e-2	100	1e-3	200	50%	64	1e-5
CNN6 ML-S	AS	1e-2	100	5e-3	200	50%	64	1e-5
AST ML-C	IN + AS	5e-3	30	1e-6	50	n/a	16	1e-3
CNN6 ML-C	AS	1e-2	100	5e-3	200	50%	64	1e-5
Baseline	Random	n/a	n/a	1e-3	200	50%	128	1e-3

Table 1: Settings of the various models (MC: multi-class, ML: multi-label, S: segmented input, C: continuous input). ImageNet initialisation is abbreviated as "IN", and AudioSet initialisation is abbreviated as "AS".

Model	Nr. of param. Nr. of FLOPS	ICBHI Validation Set Score	ICBHI Official Test Set Score	ZOL Dataset Score
AST MC-S	87,531,736 97.6 billion	60.78 ± 0.17	47.40 ± 0.40	n/a
AST ML-S	87,528,660 97.6 billion	63.64 ± 0.51	53.66 ± 0.50	n/a
CNN6 MC-S	4,306,372 15.8 billion	59.42 ± 1.84	43.09 ± 1.09	n/a
CNN6 ML-S	4,305,346 15.8 billion	62.87 ± 4.41	51.04 ± 2.85	n/a
AST ML-C	87,528,660 97.6 billion	59.18 ± 1.93	48.85 ± 0.56	57.00 ± 3.46
CNN6 ML-C	4,305,346 15.8 billion	58.96 ± 1.70	46.24 ± 0.80	56.19 ± 19.67
Baseline ML-C	482,914 84.3 million	53.78 ± 1.25	45.40 ± 0.87	58.72 ± 8.56

Table 2: Obtained ICBHI scores for the different settings (MC: multi-class, ML: multi-label, S: segmented input, C: continuous input). All numbers represent mean ± sample standard deviation (3 runs). The number of floating-point operations (FLOPS) was calculated using `fvcore` package [9]. One combined multiplication & addition is counted as one FLOP by `fvcore`.

The CNN6 architecture [7, 8] was used with AudioSet pre-trained weights. These pre-trained weights came from [8] and were used to initialise the model. For CNN6, the multi-label decision thresholds were set at 0.50.

The AST architecture was initialised with ImageNet and AudioSet pre-trained weights, as was done in [5] and [6]. The AST's patch size was set to 16×16, with overlap 10×10. The multi-label decision thresholds were set at 0.50.

3 Results & Discussion

The first experiment was to compare a multi-label output to existing multi-class models, using breathing cycle segmented audio input. The corresponding results are shown in the upper block of Table 2. When looking at the ICBHI score, AST performs best on the official ICBHI test set in both multi-class and multi-label setups. It can be seen that a multi-label setup results in improved ICBHI score. The results on the validation set are comparable to the test set results of [5].

In the second experiment, the model input was switched from breathing cycle segmented spectrograms to continuous audio spectrograms. The results are presented in the middle block of Table 2. With continuous audio input, ICBHI scores declined around 5% for both AST and CNN6. It is likely that this is caused by the start of a breathing cycle to not always be at the beginning of the segment that is offered as input to the model.

When evaluating the models on the in-house ZOL data the simple baseline model has the highest ICBHI score, indicating that AST and CNN6 might still overfit the ICBHI data. Furthermore, it is observed that for the ZOL data, the standard deviations increased compared to those on the ICBHI data. Potentially this is caused by the limited number of Littmann 3200 samples in the ICBHI training set (1.1% for continuous audio).

When the baseline model is compared to the state-of-the-art models with continuous input, it can be seen that the baseline model has almost comparable ICBHI score as the CNN6 model, with 8 times less parameters. Compared to AST, the baseline model performs worse, but also has 180 times less parameters. This poor performance could be caused by the short time span of the input.

4 Conclusion & Future Work

In this work, three contributions were made. First, a multi-class output was compared to a multi-label output using existing state-of-the-art lung event detection models. It was found that the multi-label setup improves the classification performance. Second, continuous audio input was compared to breathing cycle segmented audio input. A decline in model performance was seen when switching from breathing cycle segmented input to continuous input. Third, it was observed that a CNN model with low computational complexity has similar performance on the unseen in-house dataset compared to the AST model which has considerably more parameters and floating-point operations.

Future work could study the impact of data augmentation and channel characteristics on the accuracy of models with varying computational complexity.

References

[1] Sung Hoon Lee, Yun-Soung Kim, Min-Kyung Yeo, Musa Mahmood, Nathan Zavanelli, Chaeuk Chung, Jun Young Heo, Yoonjoo Kim, Sung-Soo Jung, and Woon-Hong Yeo. Fully portable continuous real-time auscultation with a soft wearable stethoscope designed for automated disease diagnosis. *Science Advances*, 8(21):eabo5867, 2022.

[2] Luca Arts, Endry Hartono Taslim Lim, Peter Marinus van de Ven, Leo Heunks, and Pieter R Tuinman. The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. *Scientific reports*, 10(1):7347, 2020.

[3] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, Pantelis Natsiavas, Ana Oliveira, Cristina Jácome, Alda Marques, Nicos Maglaveras, Rui Pedro Paiva, Ioanna Chouvarda, and Paulo de Carvalho. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3):035001, mar 2019. Accessed 30 July 2024.

[4] Abraham Bohadana, Gabriel Izbicki, and Steve S. Kraman. Fundamentals of lung auscultation. *New England Journal of Medicine*, 370(8):744–751, 2014.

[5] Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. *arXiv preprint arXiv:2305.14032*, 2023.

[6] Yuan Gong, Yu-An Chung, and James R. Glass. AST: audio spectrogram transformer. *CoRR*, abs/2104.01778, 2021.

[7] Ilyass Moummad and Nicolas Farrugia. Pretraining respiratory sound representations using metadata and contrastive learning. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2023.

[8] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[9] Meta Research. fvcore. [Online]. Available at: `https://github.com/facebookresearch/fvcore`. Accessed 13 February 2025.