PneumoNet: Neural networks for the detection of pneumonia from digital lung auscultation audio

Sara Pagnamenta, Etienne Salimbeni, Luka Chinchaladze
EPFL

***Abstract*—Precise lung sounds classification is still an open issue: traditional auscultation methods are limited due to biased human interpretation. This paper introduce two CNN models, which differ in the way audio crops are combined and fed to the network. They recognize specific audio patterns in the STFT spectrograms and classify healthy and unhealthy pediatric patients, suffering from different pulmonary diseases. Two different datasets are investigated and compared rigorously; however, the attempts to find a general model that performed well on both sets were not successful. After a weighted mean aggregation method, both models achieved an accuracy of 94%.**

I. Introduction

Pulmonary disease is a major global health threat: pneumonia alone causes up to 1 million childhood deaths per year[7], and COVID-19 has further revealed the destructive potential of emerging respiratory infections. Lung sound auscultation is a fundamental clinical exam in the diagnosis of respiratory disease but its interpretations suffers from significant subjectivity and inter-user bias[1]. To more objectively discriminate diagnostic patterns in lung sounds, the iGH at EPFL has developed a convolutional neural network (CNN) to identify COVID-19 with 90% accuracy[8]. A similar approach could be extended to other lung pathologies.
The following work aims to analyze such approach: we present several methods to characterize the dataset, assemble several recordings from a single patient, efficiently train a CNN and derive a final diagnosis.

II. Data and Pre-processing

The dataset is composed of lung sound recordings acquired at 22050 Hz, using the Littmann 3200 digital stethoscope at 8 different thoracic sites (Fig. 1) spanning from few seconds to more than a minute.

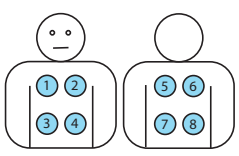

Figure 1: Location of the different auscultation site on a patient

A total of 318 patients are considered from a pediatric outpatient population aged 0 to 15 years and recruited at the outpatient departments of Geneva University Hosptial, Switzerland (GVA, n=78) and Porto Alegre (POA, n=240) in Brazil. The recruitment period for both was pre-COVID from 3.10.2018 to 28.01.2019 for POA and from 06.01.2016 to 02.03.2018 for GVA. The population is divided in controls (healthy, n=103) and cases (pathological, n=215) (table I), and cases have four subcategories: bacterial pneumonia, viral pneumonia, bronchiolitis and asthma (n=49, 9, 137 and 17 respectively).

	Nb.patients	Total recording
GVA Cases	55	1h40min
GVA Controls	23	1h35min
POA Cases	160	15h10min
POA Controls	80	3h50min

Table I: Overview of the dataset, divided in healthy and unhealthy patients

A. Data augmentation

To improve the performance of the CNN, several methods for data augmentation have been tested:

- Crop audios into smaller segments of 5 seconds each, With a 50% overlap;
- Introduce random noise;
- Change the loudness and/or the pitch;
- Shift spectrograms along the time axis.
- Sample-wise normalization
- Feature-wise normalization

B. Audio transformation

Three traditional approaches of sound transformation were compared [9].

- Short-Term Fourier Transform (STFT) power spectrum, rescaled in decibels;
- Mel-spectrogram derived from the Mel filter banks;
- Mel-Frequency Cepestral Coefficients (MFCC), obtained after a linear cosine transformation.

The Mel scale reflects human like audio perception. It offers a higher resolution in lower frequency compared to normal spectrograms[4].

C. Generalization

Because POA and GVA audios were acquired in different conditions, changes in background noise are noticeable (Fig. 2). It also seems that POA audios were previously low-pass filtered. These factors will

influence the generalisation of our model. The addition of white noise or cropping both sets to lower frequency range did not improve inter set performance, thus it will not be included in the discussion of this paper.

To improve inter-data set performance and to understand the importance of each position for determining the disease the following techniques were used:

- Set high dropout rate while training to improve generalization
- Train the model on one data set and transfer learn on another. The initial model was firstly trained 10 times, then the best model was selected for transfer learning on a different data set.

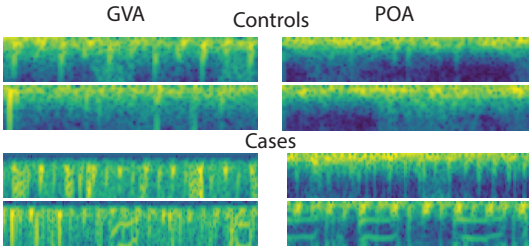

Figure 2: Spectrogram examples (notice the similarities between POA cases and GVA Controls)

D. Training, Validation and Testing

From the 78 GVA patients, 9 were used for validation, 69 for training and no testing (insufficient data). From the 240 POA patients, 8 were used for validation, 90 for batch 1 and 66 for batch 2 (balanced batches), the rest 84 were left out in batch 3 (only disease cases) and were only used for testing but not training.

III. Methods

A. Model by Position (MPO)

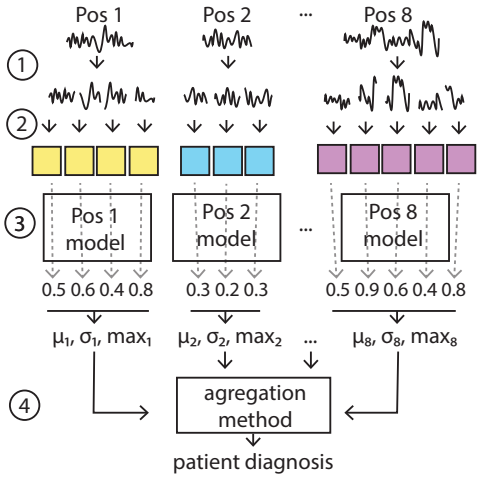

Figure 3: Model by position diagram. 1. Crop audio data with overlay, 2. map crops to spectrograms, 3. estimate the diagnosis of each spectrogram with their corresponding model, 4. combine the 8 means, std and max of each position prediction to estimate the final diagnosis of the patient

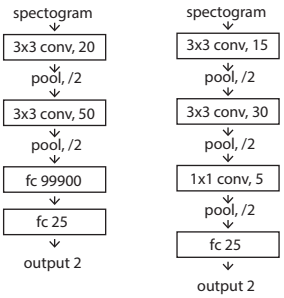

Figure 4: CNN model, right: model by position, left: model by patient

To avoid filling missing positions audios with 0s we train 8 different models (see Fig.4 and Fig.3), one per position. We tested the following aggregation methods to combine the probabilistic output of each model:

- simple average of probabilities
- weight the mean with a factor of the inverse loss for each model

B. Model by Patient (MPA)

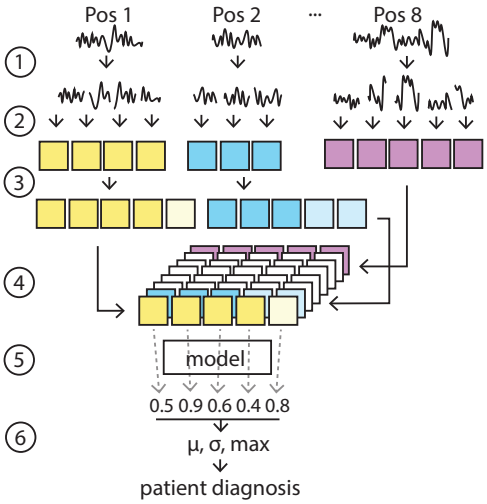

Figure 5: MPA. 1. Crop audio with overlay, 2.map crops to spectrogram 3. duplicate random crops in order to have the same number of crops for each position, 4. stack all crops depth-wise by position, 5. estimate each of the stacked crops tensors, 6. combine the estimations to get the final patient diagnosis

The MPO does not take into consideration correlations across positions due to the late fusion. The MPA uses early fusion and combines all the spectrograms for individual positions depthwise in preprocessing.Each patient would produce N number of data samples with N being the number of crops for the longest audio for any position. For the other positions, all the available crops were used first, then any random crop was sampled. If the recording for a certain position was missing altogether the layer was filled with 0s. (see Fig. 5).

To assess the importance of each position, we trained 8 different models each time with one position removed and compared it to the full model

C. Pneumonia vs Bronchitis

The way the data is collected between unhealthy and healthy patients is fundamentally different. To check that the network is not simply detecting the difference in the data collection, the model was tested whether it could learn to differentiate between bronchiolitis and pneumonia.
This condition was tested on POA patients, with undersampling of bronchiolitis cases. In total, 60 patients (52 for train and 8 for test) were used.

IV. RESULTS

A. Pre-processing

In Table II the effect of the various pre-processing techniques on the GVA dataset in the MPA are compared. The best forming model was from Mel.

	Loss		Accuracy		F1	
	μ	σ	μ	σ	μ	σ
STFT	0.36	0.14	0.86	0.071	0.82	0.075
MFCC	0.40	0.17	0.82	0.11	0.76	0.17
MEL	**0.34**	0.07	**0.88**	0.03	**0.89**	0.04
Sample Wise	0.47	0.23	0.75	0.18	0.74	0.17
Feature Wise	0.49	0.14	0.74	0.16	0.75	0.16

Table II: All of the results were calculated using the MPA on GVA dataset. Normalization techniques were performed on STFT spectrograms. The statistics are calculated for 5 second crops not for individual patients. Each method was run 10 times with maximum of 200 epochs and early stopping if the validation loss stopped improving. For each run the minimum validation loss and corresponding value of validation accuracy and F1 score were selected. μ and σ represent mean and std respectively. The best performing model is in bold.

B. Model by Position (MPO)

To validate the GVA model, 10-fold cross validation was used (see Table III) whereas for POA_1, it was POA_2 + POA_3 sets (see Table IV). The notation of GVA and POA_1 in the table means that the model was **trained** on these sets. From the the 10-fold cross validation it is evident that the GVA data is roughly homogeneous, thus due to lack of data, entirety of GVA set (train+test) was used for testing.

	loss		acc		f1	
	mean	std	mean	std	mean	std
pos1	0.40	0.04	0.83	0.04	0.83	0.04
pos2	**0.31**	0.07	**0.87**	0.05	**0.89**	0.04
pos3	0.44	0.08	0.80	0.08	0.81	0.09
pos4	0.44	0.08	0.79	0.06	0.79	0.06
pos5	0.37	0.09	0.82	0.05	0.79	0.05
pos6	0.39	0.06	0.84	0.05	0.86	0.06
pos7	0.40	0.04	0.81	0.05	0.80	0.06
pos8	0.39	0.06	0.83	0.05	0.82	0.04

Table III: 10-fold cross validation on GVA

C. Model by Patient (MPA)

The table V shows the confusion matrix using MPA. The POA_1 →GVA means that the model was initially trained on the POA_1 data set and then transfer learned on the GVA data etc. The models GVA, POA_1 →GVA were tested on the entire GVA set (since we did not have test set). On the other hand the models POA_1 and GVA→POA_1 were tested on POA_2 + POA_3. The predictions are aggregated over multiple crops with simple mean.

		TP	TN	FP	FN	Acc.
GVA	mean	47	23	1	7	0.89
	w. inv.loss mean	**50**	**23**	**0**	**4**	**0.94**
POA_1	mean	112	24	2	11	0.91
	w. inv.loss mean	**116**	**25**	**1**	**8**	**0.94**

Table IV: Confusion matrix for POA and GVA , using 2 aggregation methods (simple mean and weighted mean using the inverse of the loss of each loss). Th best model in bold.

	TP	TN	FP	FN	Accuracy
GVA	**55**	**23**	**0**	**0**	**1.00**
POA_1 → GVA	52	21	2	3	0.94
POA_1	108	23	3	16	0.87
GVA → POA_1	**116**	**24**	**2**	**8**	**0.93**

Table V: Prediction for individual patients. Predictions were done on best models only (one with the lowest loss). The Best model in bold.

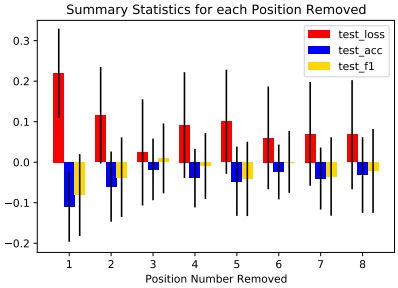

Figure 6: Performance of the individual models when we removed each position during training. For each model the statistics of the full model were subtracted. The positive loss means the increase in loss when the given position was removed.

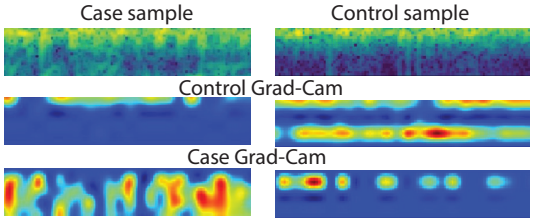

Figure 7: Grad-Cam [5] analysis of the MPA on POA. Frequencies increase from top to bottom. The second row shows how many features the model deemed to be similar to what it thinks control should have. Thus, for the case sample it didn't find to have many features similar to control but found more features to be similar to control in control sample. Similarly for the third row.

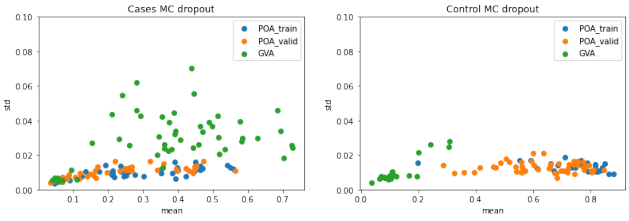

Figure 8: Monte Carlo Dropout accuracies mean(x-axis) and std(y-axis) over 20 estimations with 0.2 dropout rate. The smaller the std the more certain is the prediction of the model. Left estimation of the POA model for Cases , right for controls. GVA patients (Green) , POA_1 (blue) and POA_2 (Orange). Note, in this graph the labels are reversed. Cases are between 0 and 0.5 and controls 0.5 and 1.

D. Pneumonia VS Bronchiolitis

The training was performed using the MPA and MPO. The best of the 10 MPA runs (with an accuracy of 81%, loss of 47% and F1 of 82%) has been chosen for prediction. The predictions found in table VI are computed on the test and training sets together, since there was no test set.

POA-Pneu-Bronch	TP	TN	FP	FN	Accuracy
Model by positions (MPO)	19	26	4	11	**0.75**
Model by patient (MPA)	22	22	8	8	0.73

Table VI: Prediction for individual patient. In this case, TP are patients with pneumonia and TN patients with bronchiolitis. For aggregation, the max method was used.

V. Discussion

The results shown in table II were used to decide which strategies to employ to perform the analysis. STFT performed better in most cases (even though MEL has slightly higher accuracy on just the MPA) and it was chosen as the default spectrogram. Normalization worsened performance and it was excluded. Data augmentation seemed promising; however, due to computational limitations, it was left out.
Transfer learning did not improve inter set performance but increased the base accuracy when transferred onto POA but not vice versa.

When comparing two approaches, the MPA outperforms the MPO on GVA set. On the other hand, the MPO yields higher accuracies on POA. MPO also performs better when trying to differentiate pneumonia and bronchiolitis. The accuracy obtained, even if low, suggests that the predictions are made beyond simple differentiation of the data collection methods.

Overall, both models offer roughly the same performance. Both have pros and cons. The MPA is simple to use, does not require complicated post-training aggregation methods, but suffers when patients have missing positions and is computationally heavy.

On the other hand, MPO is flexible (new positions can be added without retraining), lightweight (fewer data to train a model) but needs 8 separate models to train and requires aggregation of outcomes, which is not always obvious, as seen in table 4.

The MPA pays most attention to position 1 followed by 5 and 2 (see Fig.6). On the other hand, MPO has the lowest loss for position 2 and 5 (see table III), hinting that the both models look for similar features in the data.

The Grad-Cam class activation visualisation (see Fig. 7), confirms that the features for healthy patients in POA data are contained at lower frequencies (frequency in the figure increases from top to bottom) and features for unhealthy at high frequencies. It explains why the model is only valid within its dataset as POA and GVA differ at high frequencies. Looking at Fig.2, one can see that both POA cases and GVA controls contain dense information at high frequencies, thus the model confuses GVA controls as cases. On the other hand, GVA cases have almost no information at very low frequencies, this is very unusual for the model trained on POA, and since the information in POA controls is concentrated at low frequencies then the model predicts them as cases but with low certainty. The prediction and certainty of prediction is visualised in Monte Carlo Dropout estimation in Fig.8

VI. Conclusion

In summary it can be said that both proposed models offer suitable ways of differentiating between healthy and unhealthy patients using digital lung auscultation audios. The final prediction accuracies of both are above 90%. However, the GVA results are less robust due to being tested on the data that it was trained on, therefore may be biased. On the other hand 93% accuracy of the model trained on POA and tested on previously unseen 150 patients indeed seems very promising. Relatively low but still significant accuracy of the MPA model on the Bronchiolitis vs Pneumonia (75%) also confirms that the model is learning true differences in pathological sound patterns rather than potential biases in acquisition quality between pathological and healthy classes (for example clinically irrelevant background noise).

One must acknowledge the limitation of the models, as both only offer very broad usage of discriminating between controls and cases. However, when teamed up with clinical insight, such broad distinctions may be sufficient to guide objective clinical decisions. The failure to generalize a single model for both data sets suggests that the models will only perform well if the data collection is systematic and there are no drastic differences present as there were between GVA and POA.

Transfer learning was a limited success achieving higher accuracy when performed on POA but not vice versa (see Fig. VIII in appendix). Therefore, further use of the models trained in this work can be performed such as transfer learning to predict healthy vs COVID patients.

The improvements for the future endeavors can be the following: deeper understanding of the data and data collection methods between locations. Better post-training aggregation methods for the MPO such as feeding the outcomes of 8 models into small NN, or SVM. Better pre-training combination methods of different crops for the MPA, such as avoiding filling the layers with 0s but replacing with the mean of the entire data set for that position.

VII. Acknowledgments

We would like to thank Mary-Anne Hartley, Edoardo Holzl, Deeksha Shama from MLO iGH for being such an amazing supervisors through out this project, who provided us with the best feedback and a weekly dose of emojis on Slack.

References

[1] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. "Lung sounds classification using convolutional neural networks". In: *Artificial Intelligence in Medicine* 88 (June 1, 2018), pp. 58–69. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2018.04.008. URL: http://www.sciencedirect.com/science/article/pii/S0933365717302051 (visited on 12/17/2020).

[2] Francois Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[3] Aakash Goel. *How to add user defined function (Get F1-score) in Keras metricsnbsp;?* June 2020. URL: https://medium.com/@aakashgoel12/how-to-add-user-defined-function-get-f1-score-in-keras-metrics-3013f979ce0d.

[4] Ali Imran et al. "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app". In: *Informatics in Medicine Unlocked* 20 (Jan. 1, 2020), p. 100378. ISSN: 2352-9148. DOI: 10.1016/j.imu.2020.100378. URL: http://www.sciencedirect.com/science/article/pii/S2352914820303026 (visited on 12/17/2020).

[5] Yasuhiro Kubota. *visual analysis for Keras*. 2020. URL: https://github.com/keisen/tf-keras-vis.

[6] Brian McFee et al. *librosa/librosa: 0.8.0*. Version 0.8.0. July 22, 2020. DOI: 10.5281/ZENODO.591533. URL: https://zenodo.org/record/591533 (visited on 12/17/2020).

[7] *Pneumoscope – An intelligent stethoscope*. URL: https://pneumoscope.ch/ (visited on 12/09/2020).

[8] *PREPRINT - DeepBreath: Diagnostic Pattern Detection for COVID-19 in Digital Lung Auscultations*. URL: https://www.epfl.ch/labs/mlo/deepbreath/ (visited on 12/17/2020).

[9] X. Zheng et al. "A CRNN System for Sound Event Detection Based on Gastrointestinal Sound Dataset Collected by Wearable Auscultation Devices". In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 157892–157905. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3020099.

APPENDIX

	TP	TN	FP	FN	Accuracy
GVA	55	23	0	0	1.0000
$POA_1 \rightarrow GVA$	52	21	2	3	0.9359
$POA_2 \rightarrow GVA$	39	22	1	16	07821
POA_1	108	23	3	16	0.8733
$GVA \rightarrow POA_1$	116	24	2	8	0.9333
POA_2	117	21	29	7	0.7931
$GVA \rightarrow POA_2$	114	27	23	10	0.8103

Table VII: Similar to table V but with the added model trained on POA_2.

	Loss		Accuracy		F1	
	μ	σ	μ	σ	μ	σ
GVA	0.31	0.09	0.88	0.05	0.86	0.06
$POA_1 \rightarrow$GVA	0.36	0.02	0.81	0.03	0.83	0.02
$POA_2 \rightarrow$GVA	0.42	0.05	0.80	0.02	0.79	0.03
POA_1	0.22	0.10	0.93	0.05	0.94	0.06
GVA$\rightarrow POA_1$	0.26	0.04	0.89	0.02	0.86	0.01
POA_2	0.33	0.10	0.87	0.06	0.88	0.08
GVA$\rightarrow POA_2$	0.30	0.04	0.85	0.05	0.85	0.05

Table VIII: The validation accuracies during training. The first 2 rows are the models that were simply trained on the data outlined. The GVA$\rightarrow POA_1$ means that the model was initially trained on the GVA data set and then transfer learned on the POA data set batch 1 etc. Each model was ran 10 times. The statistic are calculated for 5 second crops not for individual patients. The statistics are calculated only on corresponding validation sets(no test).

	GVA						POA					
	loss		accuracy		f1		loss		accuracy		f1	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
position 1	0.3780	0.0203	0.8628	0.0051	0.8641	0.0057	0.2764	0.0956	0.8866	0.1049	0.8749	0.1002
position 2	0.4599	0.0298	0.8758	0.0045	0.8794	0.0078	0.2879	0.0822	0.9166	0.0473	0.9166	0.0473
position 3	0.4741	0.0394	0.8095	0.0585	0.8058	0.0639	0.2226	0.0138	1.	0.	1.	0.
position 4	0.4997	0.1369	0.8223	0.0841	0.8254	0.0817	0.2980	0.0209	0.9456	0.0088	0.9479	0.0085
position 5	0.4380	0.0286	0.8405	0.0241	0.8404	0.0241	0.4887	0.0910	0.7862	0.1028	0.7842	0.1002
position 6	0.4598	0.0262	0.8049	0.0494	0.8142	0.0622	0.3701	0.0326	0.8576	0.0419	0.8576	0.0419
position 7	0.3122	0.0981	0.8701	0.0682	0.8695	0.0679	0.5784	0.0367	0.6527	0.0383	0.6527	0.0383
position 8	0.2928	0.1415	0.8771	0.1112	0.8765	0.1138	0.7084	0.0877	0.5902	0.0512	0.5902	0.0512

Table IX: Each model was ran 5 times. The statistic are calculated for 5 second crops not for individual patients. The statistics are calculated only on corresponding test sets (no validation).