Pneumonia Diagnosis based on CNN-LSTM-BERT Model

Yingxue Yu, Tianzong Zhang, Qinyue Zheng
School of Computer and Communication Sciences, EPFL, Switzerland
mentored by Annie Hartly from iGH

***Abstract*—Lung auscultation is an established clinical exam for diagnosing respiratory disease, but it's interpretation by the physicians has limited accuracy. Computer-aided interpretation of Digital Lung Auscultation (DLA) has the potential to avoid inter-user bias and automate the process. We applied an existing CNN-LSTM-BERT Model for COVID diagnosis to Pneumonia Diagnosis. To simulate real life scenario, and figure out the importance of audios collected at different thoracic sites, we tested the model's robustness to missing data. We trained the model on two datasets collected from Geneva and Porto Alegre. Using only DLA audios, our model yielded 91.7% accuracy for Porto Algere and 88.9% accurracy for Geneva and correctly classified all symptomatic pneumonia cases. We found that the model is robust to missing data and surprisingly, even performed better with missing audios in general. Our model can greatly aid the diagnosis of pneumonia.**

I. Introduction

Pneumonia is a life-threatening infectious disease affecting one or both lungs in humans. Traditional auscultation technique is often used to diagnose these disorders. Although lung sounds convey relevant information related to pulmonary diseases [1], the limitations of physician expertise and the non-stationary feature of lung sounds may still lead to a wrong diagnosis [2]. Therefore, automatic recognition systems are introduced to deal with these limitations.

In the assessment of respiratory disease, computer-aided assessment of Digital Lung Auscultation (DLA) can better standardise and automate evaluation. Besides, complex deep learning approaches to sound analysis such as convolution neural networks (CNNs), deep belief networks, adversarial networks, and recurrent networks have made major advances in speech recognition and audio processing and are beginning to be used in several medical applications [3].

Here we adapt and apply an existing CNN-LSTM-BERT model for COVID diagnosis to a dataset collected by DLA to distinguish forms of Pneumonia. Taking the anatomic distribution of sound in to consideration, we also explore the effect of various auscultation positions.

II. Models and Methods

A. Dataset

1) Data Collection: Data were collected by doctors and nurses of the HUG in Geneva Switzerland from October, 2018 to January, 2019, and da Crianca Santo Antonio Hospital in Porto Alegre, Brazil from Janurary, 2016 to March, 2018. Lung auscultation recordings were acquired from eight thoracic sites as represented on Figure 1. DLA audios were recorded using a 3MTM Littmann digital stethoscope (model 3200) following a standardised procedure detailed in [4]. Additionally, clinical features were collected including demographic information, medical history, current symptoms, and findings from clinical and paraclinical exams.

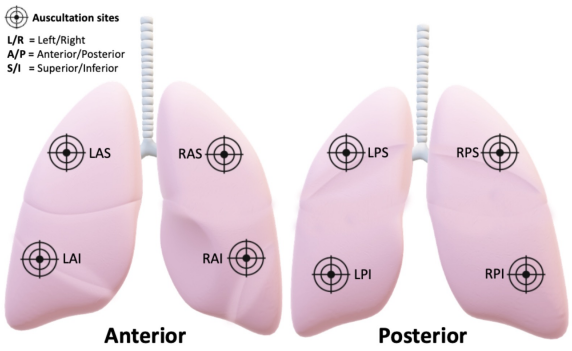

Figure 1. Sites of digital lung auscultation acquisition. Eight sites comprising the quadrants of the anterior and posterior thoracic wall (Figure courtesy to [3].)

2) Population: The dataset we use in our model consists of digital lung ausculation (DLA) audios acquired from suspected pneumonia patients at two location: Geneva and Porto Alegre. The audios are collected from 55 people who have pneumonia (cases), and 23 people who don't (controls) at Geneva (GVA), and from 168 cases, 80 controls at Porto Alegre (POA). The sex ratio of the patients is 1.35:1. Most patients are under the age of 8, with one patient aged 14.

3) Audio Recordings: For each patient, we have 4-8 DLA audios collected at different anatomic positions.The total number of anatomic positions is 8. Audio length ranges from 1.9s to 60s. Typical audio length for cases in Porto Algere is 40s, while the average length for controls is 21.7s. The average audio length for cases and controls in Geneva is 14 and 29, respectively.

B. Audio Transformation

For feature extraction, the audio was converted to the Mel scale: a biomimetic pitch scale whereby frequencies are spaced according to human perception in a linear cosine transformation of a log power spectrum. This transformation has the potential benefit of offering higher resolution in lower frequencies due to the unequally spaced frequency

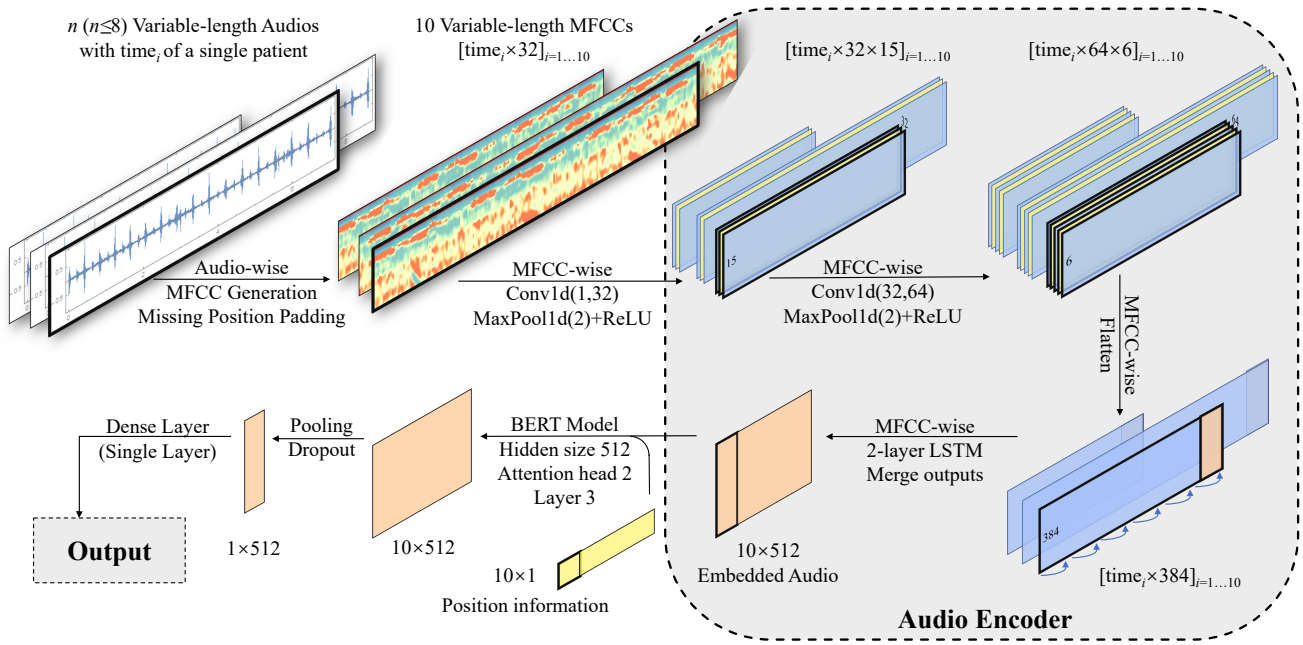

Figure 2. The pipeline of our CNN-LSTM-BERT pneumonia diagnosis model. Read the figure clockwise from the upper-right: Each patient has **1)** at most eight lung audio recordings (see 1 for detail). We transform the audios to MFCCs and pad MFCCs for missing positions. For each patient **2)** 10 MFCCs are obtained, and each MFCC is embedded by the audio encoder, specifically **3)** fed into a 2-layer CNN model followed by **4)** a 2-layer LSTM model. All 10 outpus of the 10 MFCCs are concatenated, yielding the **5)** embedded audio for a single patient. The embedded audios are trained by **6)** a 3-layer BERT model followed by a single linear classifier, finally giving the prediction label. More details are given in the main text.

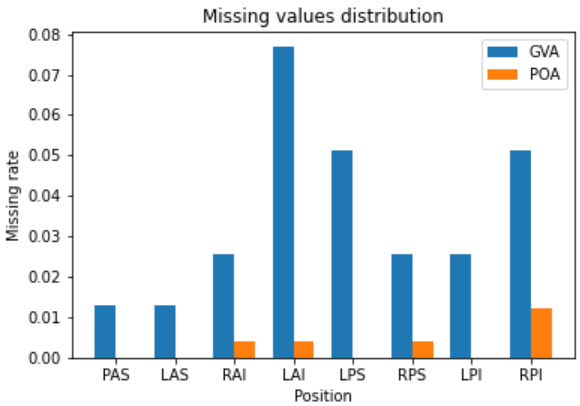

Figure 3. The missingness of each auscultation position in two datasets

bands. A total of 32 MFCC values are used in this analysis. All features extracted from the sound recordings were generated using the python library Librosa [5].

C. Missing Values

As shown in Figure, there are missing samples of each position. Compared to the expected number of samples (326 patients*8 positions = 2608 samples), we received 2580, i.e 28 site-samples were missing. Respectively, 77, 77, 76, 72, 74, 76, 76 and 74 recordings from Geneva were present for RAS, LAS, RAI, LAI, LPS, RPS, LPI, RPI and 248, 248, 247, 247, 248, 247, 248, 245 from Porto Alegre. Effect of different positions for pneumonia diagnosis is presented in section III-C.

Our diagnostic model proposed in Section II-E requires all patients have the audio files (or correspondingly MFCC files) at all the eight positions. To deal with the issue of missing data, we utilize the technique of padding for missing MFCC files. We fill a pseudo-data (1 in out case) into a pseudo-presumed MFCC with size (32×1292 in our case). The padded MFCC files of one single patient are further augmented with a start pseudo-MFCC and an end pseudo-MFCC, which we will find useful when the MFCC files are further concatenated and learned by the BERT model.

D. Statistical Methods

Summary statistics for demographic are reported for each outcome group ($\text{Pneumonia}_{\text{pos}}$ and $\text{Pneumonia}_{\text{neg}}$). The performance of model is assessed with sensitivity, specificity, area under the receiver-operator curve (AUROC) as well as prediction accuracy.

E. Our Diagnostic Model

The pipline of our CNN-LSTM-BERT pneumonia diagnosis model is illustrated in Figure 2. The model is based on deep learning techniques and contains two parts: the CNN-LSTM audio encoder and the BERT training model. We will elaborate each part in detail below.

1) CNN-LSTM audio encoder: this model embeds the MFCC files of each patient and yields an abstract audio embedding of the patient by the joint model of convolutional neural network (CNN) and long short-term memory (LSTM) [6]. CNN models are widely used in pattern recognition problems such as image and audio recognition, since it convolves and abstracts the input and hence has the ability to learn structural features of the input. Specifically in our CNN-LSTM audio encoder a two-layer CNN is used, where for both convolutional layers we perform the one-dimensional convolutions along the frequency axis. The two layers respectively contains 32 and 64 output channels with kernel size 4, followed by a max-pooling summarization with kernel size 2. Since each original MFCC contains 32 features (i.e. frequencies), the two layers respectively contain 15 and 6 features, yielding an output with 64 channels and 6 features. The output is then transformed to a single matrix with $384 = 64 * 6$ features by flattening the output along the 64 channels.

Since for each patient there are 10 MFCCs, implementing the previous CNN model yields 10 matrix with 384 features, but with different time lengths. To deal with the variable time length issue, the model of LSTM is implemented. LSTM is a recurrent neural network (RNN) architecture that deals with input with different lengths. The CNN-LSTM audio encoder uses two serial-connected bidirectional LSTMs with 256 hidden states, and the hidden states are recurred across the timesteps, which gives a vector of size 512 at the final timestep for each MFCC. Concatenating across the 10 MFCCs of a single patient yields a 10×512 matrix, which we refer as the *audio embedding* of the patient. This embedding is used for training the BERT model introduced below.

2) BERT model: Bidirectional Encoder Representations from Transformers (BERT) [7] is a deep learning technique designed for natural language processing tasks. Due to the analogy between natural language processing tasks and pneumonia diagnosis with audio data at different positions, BERT model is potentially powerful in our task. Specifically in our CNN-LSTM-BERT model, a 3-layer BERT model with 2 self-attention heads and 512 hidden units are used. The audio embedding of size 10×512 from the CNN-LSTM audio encoder, added by a positional embedding of size 10×512 embedded from the 10 positions, is fed to the BERT model as input. The BERT model outputs the enhanced embedding of size 10×512, followed by pooling process and a single-layer classifier gives the output of the CNN-LSTM-BERT model.

III. Results

A. Demographics Statistics

The cohort comprised 326 children, 78 from Geneva and 248 from Porto Alegre, among which 57 from Geneva and 169 from Porto Alegre are positive pneumonia patients. The female:male ratio is 1.35:1 and ages range fram 0 months to 177 months (median 13 months). The case dataset (i.e. $\text{Pneumonia}_{\text{pos}}$) and control dataset (i.e. $\text{Pneumonia}_{\text{neg}}$) are matched, so we only state the demographics distribution of positive patients in Table II.

		Acc	Sens	Spec
GVA	Train	86.96%	82%	100%
	Test	88.89%	80%	100%
POA	Train	91.50%	100%	74.63%
	Test	91.67%	100%	69.23%

Table I
THE PERFORMANCE OF MODELS TRAINED FROM SEPARATE DATASET

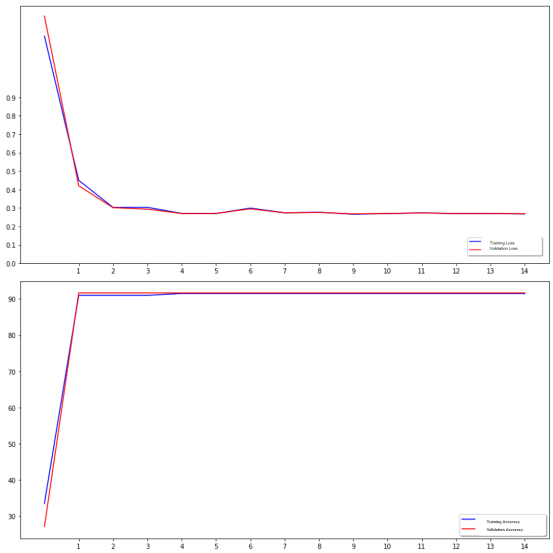

Figure 4. The learning curve of model trained with data from Porto Alegre. The upper figure illustrates the loss curve and the lower part illustrates the accuracy curve.

B. Model Evaluation

Given the environmental factors of different countries and cities, the patterns of pneumonia can be different. We have trained our LSTM-BERT model separately with two datasets, containing suspect patients from Geneva (GVA) and Porto Alegre (POA). The ratio of train and test patients is set to $4 : 1$. Details of these two dataset are described in Section II. Not surprisingly, as shown in Table I, having more input data, the model of Porto Alegre has achieved the accuracy of 91.67%, with 100% sensitivity and 69.23% specificity, while the model of Geneva obtained 88.9% accuracy, 80% sensitivity and 100% specificity. The learning curve of our model of Porto Alegre is presented in Figure 4.

C. Effect of Different Positions

To test the trained model's robustness to missing data, we conducted 9 tests, one with all the audios from the test set, and 8 others leaving out all the audios collected from a

Pnemonia+	**GVA** ($n = 57$)		**POA** ($n = 169$)	
Age (months)	median (range)	IQR	median (range)	IQR
	30 (0-177)	32	8 (0-59)	8
Sex (female)	number	%	number	%
	22	38.60	72	42.60

Table II
DEMOGRAPHIC DISTRIBUTION OF $\text{Pneumonia}_{\text{pos}}$ PATIENTS.

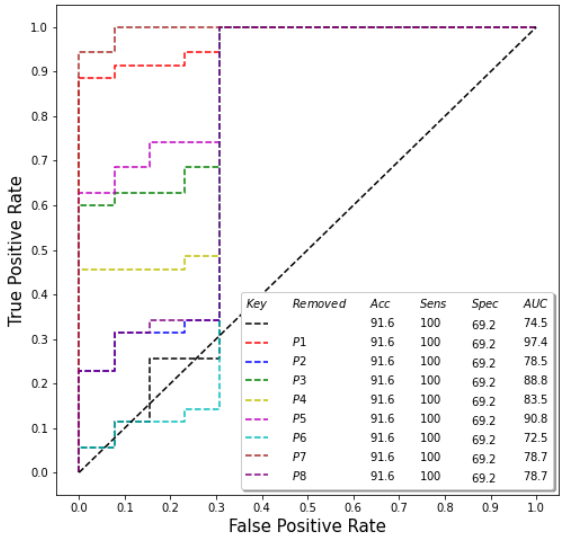

Figure 5. The effect of different positions in pneumonia diagnosis. Audio of different positions are removed to test the robustness and evaluate the importance of the positions.

specific thoracic site, respectively. The Comparison of AUC-ROC curve is shown in Fig. 3. Interestingly, compared to prediction with all the audios, leaving one audio out generally yielded higher AUC. Notably, leaving out position 7 yielded 99.3 AUC. This may due to the ineffectiveness of the audios collected at these positions. One exception is postion 6, which actually improves the the model performance.

IV. DISCUSSION

In this study, we applied CNN-LSTM-BERT model based on deep learning techniques to detecting diagnostic patterns in lung auscultations for pneumonia among 326 suspect patients. We achieved 91.6% accuracy in detecting clinical-confirmed pneumonia cases while maintaining a high specificity of 69.2% and 100% sensitivity. Our model has shown to be effective in detecting pneumonia patterns. Most studies attempt a sound-by-sound analysis, ignoring the anatomic distribution of sound at the patient level. However, audio recorded from different anatomical positions may not be equally informative. To investigate the importance of anatomical origin of the DLA audios, we conducted 9 tests, one with all the audios from the test set, and 8 others leaving out all the audios collected from a specific thoracic site, respectively. It's interesting to find that removing data at certain positions would actually yield higher AUC. The effectiveness of various auscultation sites is not the same. For further study, it's worth digging in which position contains most of the information for pneumonia detection, and how many positions at least do we need for intelligent diagnosis.

V. SUMMARY

We applied an existing CNN-LSTM-BERT Model for COVID diagnosis to Pneumonia Diagnosis. We also tried out And tested the importance of different thoracic sites as well as the model's robustness to missing data. Our model achieved high accuracy and is robust to missing data. We also found that using more audios from different positions doesn't necessarily produce better prediction results. And the importance of different positions vary. In the future, we can test the effectiveness of combining pretrained model, and finding out the best combination of positions for prediction.

ACKNOWLEDGEMENTS

We would like to thank Annie Hartly, Deeksha M Shama and Edoardo Holzl for their help and suggestions. We also appreciate the opportunity provided by Intelligent Global Health Lab for us to delve into this very interesting project and to better understand the application of deep learning techniques in the health field.

REFERENCES

[1] 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.

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

[3] D. M. Shama, A. Glangetas, A. Cantais, T. Chavdarova, E. Holzl, D. Courvoisier, S. Bourquin, J. Dervaux, D. Rivollet, M. Jaggi, A. Gervaix, J. Seibert, and M.-A. Hartley, "Deepbreath: Diagnostic pattern detection for covid-19 in digital lung auscultations," *Preprint*.

[4] N. C. Galli, "Place de l'intelligence artificielle dans la gestion de l'asthme de l'enfant," *Travail de master, UNIGE, Suisse*.

[5] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[6] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.