CONVOLUATIONAL TRANSFORMER WITH ADAPTIVE POSITION EMBEDDING FOR COVID-19 DETECTION FROM COUGH SOUNDS

Tianhao Yan[1,2]*, Hao Meng*[1]*, Shuo Liu*[2]*, Emilia Parada-Cabaleiro*[3]*, Zhao Ren*[4]*,Björn W. Schuller*[2,5]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, 150001, China
[2] EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[3]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[4] L3S Research Center, Leibniz Universität Hannover, Germany
[5] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

menghao@hrbeu.edu.cn

ABSTRACT

Covid-19 has caused a huge health crisis worldwide in the past two years. Although an early detection of the virus through nucleic acid screening can considerably reduce its spread, the efficiency of this diagnostic process is limited by its complexity and costs. Hence, an effective and inexpensive way to early detect Covid-19 is still needed. Considering that the cough of an infected person contains a large amount of information, we propose an algorithm for the automatic recognition of Covid-19 from cough signals. Our approach generates static log-Mel spectrograms with deltas and delta-deltas from the cough signal and subsequently extracts feature maps through a Convolutional Neural Network (CNN). Following the advances on transformers in the realm of deep learning, our proposed architecture exploits a novel adaptive position embedding structure which can learn the position information of the features from the CNN output. This make the transformer structure rapidly lock the attention feature location by overlaying with the CNN output, which yields better classification. The efficiency of the proposed architecture is shown by the improvement, w. r. t. the baseline, of our experimental results on the INTERPSEECH 2021 Computational Paralinguistics Challenge CCS (Coughing Sub Challenge) database, which reached 72.6 % UAR (Unweighted Average Recall).

Index Terms— SARS-CoV2 Detection, Computer Audition, Convolutional Neural Network, Adaptive Position Embedding Transformer, log-Mel Spectrogram.

1. INTRODUCTION

The World Health Organization (WHO) declared Covid-19, caused by the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV2), a global pandemic on March 11, 2020. Causing tens of thousands of lives since then, great efforts and sacrifices are still ongoing in order to overcome the virus. Amongst these, developing methods to detect and isolate the disease at early stages has been identified as one of the most effective means to fight the pandemic [1]. Since the virus mainly attacks the human respiratory system, cough and shortness of breath are some of the most salient symptoms of Covid-19 [2]. Unlike other signals, such as X_ray, which are only available through professional instruments, coughing can be easily retrieved trough widely available devices, e. g. , mobile-phones. Hence, developing solutions to automatically identify Covid-19 disease from people's cough is a plausible solution that would considerably reduce the time and costs of the detection process.

Even though attempts to automatically detect the presence of Covid-19 infection from coughing have been developed, existing methods still need to be improved. For instance, the mainstream transformer position encoding structures broadly used in Natural Language Processing (NLP) has no yet presented major breakthroughs in speech technology. This relates, to some extent, to the fact that the general transformer structure relies on features' dimension to adopt the attention mechanism algorithm, thus ignoring the position information of the effective features. When working with log-Mel spectrograms, i. e. , features maps extracted from audio signals which contain information in both temporal and frequency domain, the fact that transformers ignore the position information could result on the extraction of less accurate features from the spectrograms [3, 4, 5].

Inspired by the work by Carion et al. [6], we exploit a novel convolution transformer structure with the adaptive position embedding feature map of time and filter dimension. Firstly, in order to form three channels spectrogram feature maps, we extract the log-Mel spectrogram, which includes static, delta, and delta-deltas, from the coughing samples. These are subsequently inserted into a Convolutional Neural Network (CNN) to further capture information through a skip connection framework. Finally, the feature maps are fed into the structure of a transformer along to the adaptive position information. Due to the time and frequency domain characteristics of audio, the addition of adaptive positional embedding allows us to find the relevant feature areas of the spectrogram, which enhances the final classification result. The main novelty of our work is designing an adaptive position embedding structure with a transformer which integrates the log-Mel spectrogram's feature maps into a standard architecture. To validate the effectiveness and robustness of the proposed architecture, we present experimental results on the INTERPSEECH 2021 Computational Paralinguistics Challenge CCS (Coughing Sub Challenge) database [7].

The rest of the manuscript lays out as follows: in Section 2, the state-of-the-art research on the topic is outlined; in Section 3, the core aspects of the architecture are described; in Section 4 and 5, the experimental setup and results are discussed; finally, in Section 6, conclusions and future works are given.

2. RELATED WORK

For addressing the question of how to detect the presence of Covid-19 infection timely, a variety of datasets has been presented. Unlike corpora collected through professional devices [8], coughing-based

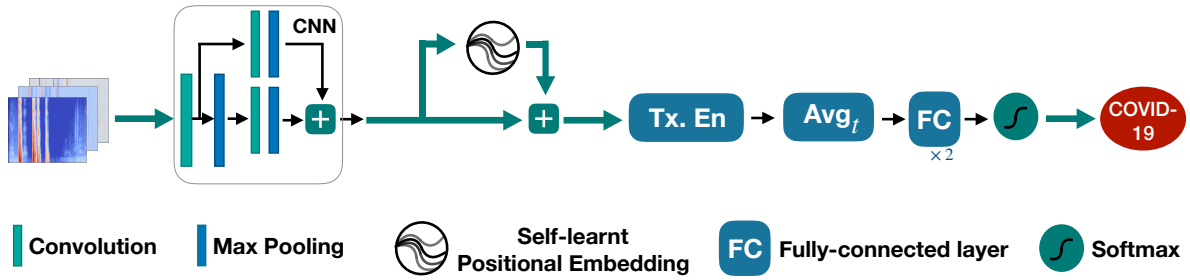

Fig. 1: The 3-channel log-Mel spectrograms are input on a 2-layer CNN with residual structure. The position information of feature maps is learnt in terms of adaptive position embedding networks superimposed with the CNN output and connected with a Transformer(Tx)–Encoder(En) framework followed by an AveragePooling2D layer. After 2 fully connected layers a softmax layer yields the results.

datasets gathered in realistic conditions [7, 9, 10] are a great resource to promote the development of artificial systems for in-the-wild applications. To this end, Artificial Intelligence (AI) algorithms offer a broad umbrella of audio-based potential solutions. Imran et al. leverage a Deep Transfer Learning-based Multi Class classifier (DTL-MC) to develop and test an AI-powered screening cough-based method to identify patients affected by Covid-19, pertussis, and bronchitis [11]. Casanova et al. achieved promising results on cough recognition by applying transfer learning on pre-trained networks [12]. CNNs have also been commonly used for diagnosing Covid-19, not only trough X-ray image recognition [13], but also through the identification in audio signals [14, 15]. Because of the temporal nature of the speech signal, Recurrent Neutral Networks (RNN) and Long-Short Term Memory (LSTM), able to further capture temporal information, were also successfully applied for the task at hand [16]. For instance, Yan et al. designed a novel Spatial Attentive ConvLSTM Neural Network (SACRNN) structure that has the ability to further extract the temporal and frequency domain feature information based on log-Mel spectrograms [17]. Finally, with the development of NLP, the structure of an attention mechanism represented by multi-head attention achieved great success in various fields [18], being applied in the context of cough recognition too [19].

3. ARCHITECTURE

As input features for the proposed architecture, we extract information from 3-channel log-Mel spectrograms exploiting a convolutional layer. Subsequently, an adaptive layer aimed to retrieve the position from the feature maps is applied. Finally, to reconcile the features with the positions while discovering the best features via the transformer structure, we employ 2 fully-connected layers. The flow chart of the whole network structure is shown in Fig. 1. For reproducibility purposes, the source code to recreate the experimental results presented in this article is made freely available.[1]

3.1. Log-Mel Spectrograms Generation

For each audio sample, we divide the raw signal into frames and apply a window function. Subsequently, we perform Fast Fourier Transform (FFT) on each frame. The whole process transforms the time domain signal into a frequency domain signal. After applying 40 Mel filterbanks to the energy spectrum under the role of Eq. (1), where the samples are set to a sample rate of 16 kHz and a Hamming window of 25 ms with a shift of 10 ms, the log-Mel spectrograms are obtained by

$$MelSpec(m) = \sum_{k=f(m-1)}^{f(m+1)} log(H_m(K) * \left|X(k)^2\right|), \qquad (1)$$

where $|X(k)|^2$ denotes the energy of the k-th point in the energy spectrum, $H_m(k)$ describes the m-th Mel-filterbank, and m indicates the number of filterbanks.

Due to the non-linear human perception of sound intensity, it is reasonable to compute log-Mel spectrograms, as well as to collect dynamic information on the change of MFCC over time. Hence, we extract delta and delta-deltas based on the static spectrogram and jointly form a 3-channel input feature $M \in \mathbb{R}^{t,f,c}$ [20, 21], where t denotes time, f describes the number of filters, and c represents the number of channels, which is set to three:

$$M(m)^d = \frac{\sum_{n=1}^{N} n \left(M(m)_{t+n} - M(m)_{t-n}\right)}{2\sum_{n=1}^{N} n^2}, \qquad (2)$$

where $\mathbf{d}$ stands for the number of iterations of delta, which is set to 1 and 2 for delta and delta-deltas, respectively. $\mathbf{t}$ symbolises the position of the frame, $\mathbf{n}$ indicates the number of differences between the current frame and the previous-next frames, and $\mathbf{N}$ is set to 2 in general. Thus, the 3-channel log-Mel spectrograms are obtained.

3.2. CNN Model

After obtaining the log-Mel spectrogram group, we designed two standard convolution layers with a residual structure as shown in Figure 1. Due to the reduced size of the samples, we chose 5*5 convolutional kernels as suitable in the main structure of the CNN. This enables also to increase the perceptual field by using fewer layers, by this fully covering the feature maps and reducing the risk of overfitting due to the overly redundant and high complex nature of the model. Finally, a Maxpooling layer was also added in order to reduce the size of the feature maps, so that the information could be focused on the channel dimension after the convolution layers. It is a remarkable fact that we adopt a convolutional kernel of 3x3 size which may contain some valuable feature information for supplementing the missing details from the subject structure. The whole structure is designed to extract as much feature information as possible without raising the complexity of the model. By this, we aim to reduce the dimensionality and pave the way for a further feature extraction step carried out by using the adaptive position embedding transformer. The parameters of the whole structure are listed in Table 2.

3.3. Adaptive Position Embedding Transformer

Since the model complexity is impaired by the limited number of samples, in order to ensure its convergence and further extract remaining information, an adaptive position embedding transformer

[1] https://github.com/EIHW/CCS_SLPETX/

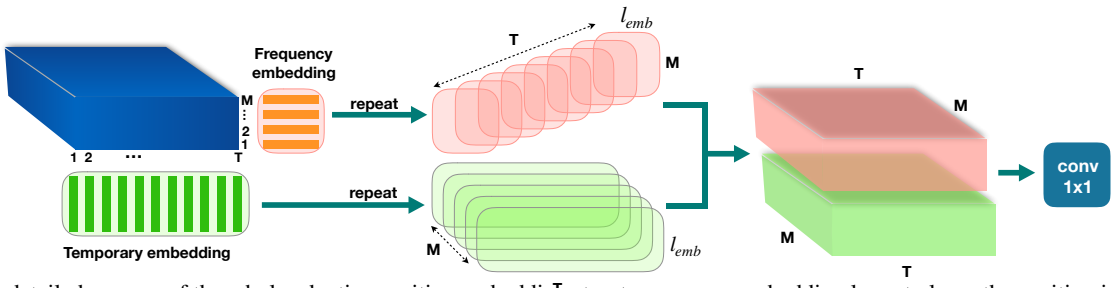

Fig. 2: The detailed process of the whole adaptive position embedding structure uses an embedding layer to learn the position information of the Time_Dimension and the N_Mel_Dimension respectively, and then further enhances learning based on the convolutional layer after merging.

Table 1: Hyperparameters of the whole CNN model.

Layer	In_Ch	Out_Ch	Kernel_Size	Stride	Padding
Conv1	3	32	(5,5)	(1,1)	(2,2)
MaxP1	32	32	(2,2)	(2,2)	(0,0)
Conv2	32	64	(5,5)	(1,1)	(2,2)
MaxP2	64	64	(1,2)	(1,2)	(0,0)
Conv_skip	32	64	(3,3)	(1,1)	(1,1)
MaxP_skip	64	64	(2,4)	(2,4)	(0,0)

was designed (cf. Figure 2). By taking as starting point the feature maps obtained from the CNN [Batch_size, Output_channels, Time, N_mels], the disadvantage of traditional position encoding is that only the position encoding in the direction of time_steps is considered, while Spectrogram is a 2-D feature map [22]. Therefore, we encode the Time and N_mels dimension according to their own number, respectively. For instance, if the number of the Time dimension is 150, we gain 0, 1, 2, ..., 149, position information after encoding it. The same procedure applies for N_mels. Subsequently, we utilise two embedding layers to map the Time and N_mels dimensions to be 2-dimensional vectors, separately, and the number of mapped dimensions is set to half of the number of Output_channels dimension. Besides, we also initialise the embedding layer weights to make them conform to the Gaussian distribution so that the transformer can find the valid feature information better and faster, instead of just executing a simple flattening like 'traditional' location coding method. Finally, after memorising the position content, we employ an unsqueeze layer to reshape the obtained feature map into the same size of the CNN output, by this making it possible to learn contents of the two dimensions together. Subsequently, we retrieve the final position information through a convolutional layer. Note that the purpose of augmenting the CNN is to strengthen the learning of the position again, so that the learnt information is more evenly distributed in the Output_channels which contain holistic feature information.

After passing the log-Mel spectrogram group through the adaptive position embedding layer, the feature map is turned into a 3D tensor [Batch_size, Time, Output_channel*N_mel] and then fed into the transformer algorithm. Instead of an RNN, which cannot process feature information in parallel to a certain extent, or a multi-stacked multilayer CNN, which causes parameter redundancy and increases the training cost, we employ the encoder part of the transformer for the classification. Firstly, the self-attention computes three different transformation matrices W_q, W_k, W_v with linear projections to obtain queries(Q), keys(K), and values(V) from the feature dimension, which is made up of the multiplication of Output_channel and N_mel. The attention output matrix is obtained using Q, K, V by

$$Attention(Q, K, V) = \mathbf{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

Table 2: Hyperparameters of the transformer encoder structure with adaptive position embedding as well as the transformer layer (N).

Adaptive Position Embedding

parameters	Size	Kernel
Pos_T_Emb	(150,32)	-
Pos_M_Emb	(10,32)	-
Pos_Conv	(64,64)	(1,1)

Transformer (**N** = 1)

parameters	values
Dim_features	640
heads	8
Dim_feed_forward	512

Fully Connected Layers

Block	In_features	Out_features
FC1	640	64
FC2	64	2

where Q and K are scaled inner products in order to obtain attention representation at multiple scales. d_k is the dimension of Q and K. Since the value of the inner product raises as the dimension boosts, the normalisation effect is achieved by $\sqrt{d_k}$. Besides, it is crucial to exploit multi-dimensional feature characterisation at different position. Therefore, a multi-head transformer is further proposed which is composed of concatenating the attention output matrix of all heads and multiplying it by an output weight matrix W_o with dimension to attain the final output [23]:

$$\begin{aligned} MultiHead(Q, K, V) &= \mathbf{Concat}(head_1, ..., head_n)W^o, \\ \mathbf{where}\ head_n &= \mathbf{Attention}(QW_n^Q, KW_n^K, VW_n^V). \end{aligned} \tag{4}$$

At the same time, we also exert other techniques from the transformer structure include adding an Add & Norm Layer, where Add stands for residual connectivity (aimed to prevent network degradation), and Norm for Layer Normalisation (used to normalise the activation values of each layer). In addition, it is also essential to operate the Feed Forward Network (FFN), which introduces the ReLU activation function and then transforms the space of the attention output. By this, the FFN only has the ability to further integrate its own features at each time step, i. e. , it is independent of other time steps, which can enhance the expressiveness of the model. The output of the transformer encoder generates a one dimensional tensor by means of a mean reshape layer and two fully connected layers, by this generating the final binary classification results. The specific parameters of the whole structure are shown in Table 2.

4. EXPERIMENTAL DESIGN

We utilise the Covid-19 Cough Sub-Challenge (CCS) database from the INTERSPEECH 2021 Computational Paralinguistics ChallengE (ComParE) [7], which provides audio samples and the corresponding COVID-19 test labels. The corpus contains 929 cough recordings from 397 participants with a total duration of 1.63 hours. Participants

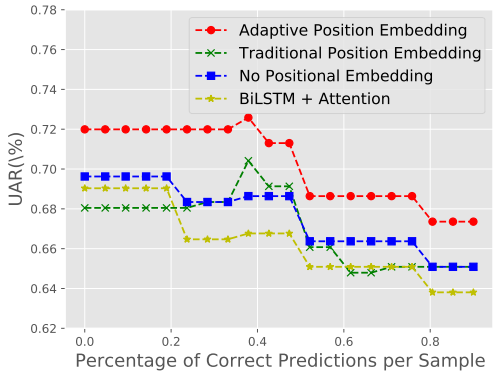

Fig. 3: UAR(%) results based on the proportion of correctly predicted segments to total segments in each sample.

produced one to three forced coughs, which were binary annotated (either positive or negative). A 3-set partitioning of the samples is given: training (286 samples), validation (231), and test (208). All the experiments are conducted in the Pytorch deep learning framework, and the model hyperparameters are set as follows: Batch_size is set to 8; cross_entropy is chosen as loss function; the learning rate is set to 1e-2; Stochastic Gradient Descent (SGD) is used as the optimiser, where weight-decay and momentum are set to 1e-3 and 0.8, respectively. The model hyper-parameters are optimised on the validations set. Due to the unbalanced distribution of the samples across the two labels, we will interpret the experimental results in terms of Unweighted Average Recall (UAR) as in the challenge. The additional metrics accuracy, specificity, sensitivity, and confidence intervals, will be also reported.

In order to accelerate the mini-batch training procedures, it is necessary to unify the input data length; thus, each entire sample is lifted according to the criterion of intercepting a fixed 300-frame segment. For samples containing less than 300 frames, it is compulsory to carry out a zero padding operation. Differently, for samples with more than 300 frames, we divide the number of frames by 300, take the quotient of the total number of segments intercepted by the sample, and then judge the remainder. If the remainder is greater than or equal to 100 frames, a zero padding operation is added to the back, otherwise, the remainder is discarded. This operation is applied on both the validation and test sets. We set a majority voting function to aggregate the predictions of segments within one sample for the test set in order to better compare with other structures. Since our model achieves around 100 % recognition rate on the negative Covid-19 labels in the test set, the voting function determines the final result by judging the proportion of correct predictions. In Figure 3, the line graph of the proportion is shown. We can see that when about 30 %–50 % of the segments in the test set are accurately predicted, the test set sample is identified as positive, reaching the highest UAR with adaptive position embedding (72.6 %).

5. RESULTS AND DISCUSSION

In this work, we aim to demonstrate the effectiveness and generalisation of our designed adaptive position embedding transformer architecture. To evaluate the experimental result, we comparatively assess the outcomes of the proposed framework by keeping the CNN structure parameters unchanged (cf. the lower part of Table 3). In addition, we also indicate the results from previous works as baseline for the discussion (cf. upper part of Table 3). We can observe that with the classical combination of bidirectional LSTM (BiLSTM) + traditional attention mechanism, the structure only obtains 66.8 % UAR (cf. ACRNN). Similarly, when considering the transformer structure without adding position information, the UAR still remains below 70.0 % (cf. 68.6 % for C-Tx). Our results also show that adding position encoding to the traditional transformer clearly improves the model's performance, reaching 70.4 % UAR (cf. C-TPETx). However, the best outcomes are achieved by the proposed structure, i. e. , our designed adaptive position embedding transformer architecture, which reaches a meaningful improvement w. r. t. the other models and baselines: 72.6 % UAR, 96.45 % Specificity, and 48.72 % Sensitivity (cf. C-APETx).

Table 3: Overall results for the four evaluated methods. Unweighted Average Recall (UAR), Accuracy (Acc.) , Specificity (SP), Sensitivity (SE), and Confidence Intervals (CI) for UAR, are given (%).

Methods	UAR	Acc.	SP	SE	CI
End2You [7]	64.7	–	–	–	–
Fusion [7]	73.9	–	–	–	–
E. Casanova et al. [12]	75.9	–	–	–	–
S. Illium et al. [19]	72.0	–	–	–	–
T. Yan et al. [17]	73.2	83.65	89.94	56.41	±15.10
ACRNN	66.8	86.86	97.63	35.90	±15.67
C-Tx	68.6	87.50	98.82	38.46	±14.71
C-TPETx	70.4	85.57	94.67	46.15	±16.55
C-APETx	**72.6**	**87.50**	**96.45**	**48.72**	±15.01

In comparison with the results by Casanova et al. which show a higher UAR (75.9 %), it is important to mention that their approach, unlike the herein presented one, is based on a large-scale transfer learning model, whose complexity implies a longer training time and higher computational effort [12]. Similarly, although the difference between our results and the ones by Yan et al. is minimal (73.2 % UAR) [17], they proposed a solution based on the fusion of two models, which yields a more redundant and complex architecture. The same applies for the baseline fusion framework from the CCS Sub-Challenge (73.9 % UAR) that fuses multiple models [7]. Finally, w. r. t. the results proposed by Illium et al. [19] and the End2You baseline [7], our model is slightly superior in terms of both results and structural parameters. By the above comparative analysis, we can conclude that the structure of the proposed model shows to be particularly effective for cough-based recognition of Covid-19.

6. CONCLUSION

We presented a novel adaptive position embedding transformer structure able to outperform the baseline for the INTERSPEECH 2021 ComParE coughing database. Unlike previous works, which rely on complex and redundant models, our architecture minimises the dimension size of log-Mel spectrograms (without boosting the depth of the model) by utilising a CNN with skip connections. This is particularly useful since it enables to retain the most relevant feature information, something especially important when working with small datasets, as those for Covid-19 recognition. Furthermore, another advantage of the presented framework is that it can also identify the most relevant feature areas through an adaptive position embedding transformer.

One of the limitations of our approach is that the outcomes in terms of sensitivity are not very satisfactory. This may be caused on the one side by the very small size of the dataset, on the other side by the imbalanced distribution of samples across the classes. Hence, further research should be oriented towards the collection of bigger and more balanced corpora for the presented task. In future works, the algorithm's generalisation ability should be tested on other databases, by this further validating the presented outcomes.

7. REFERENCES

[1] Kranthi Kumar Lella and PJA Alphonse, "A literature review on covid-19 disease diagnosis from respiratory sound data," *AIMS Bioengineering*, vol. 8, no. 2, pp. 140–153, 2021.

[2] Hanie Esakandari, Mohsen Nabi-Afjadi, Javad Fakkari-Afjadi, Navid Farahmandian, Seyed-Mohsen Miresmaeili, and Elham Bahreini, "A comprehensive review of covid-19 characteristics," *Biological procedures online*, vol. 22, pp. 1–10, 2020.

[3] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.

[4] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[5] Yi Chang, Zhao Ren, and Björn Schuller, "Transformer-based CNNs: Mining temporal context information for multi-sound COVID-19 diagnosis," in *Proc. EMBC*, Virtual, 2021, pp. 2335–2338.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[7] Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, et al., "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[8] Ali Haider Khan, Muzammil Hussain, and Muhammad Kamran Malik, "Ecg images dataset of cardiac and covid-19 patients," *Data in Brief*, vol. 34, pp. 106762, 2021.

[9] Lara Orlandic, Tomas Teijeiro, and David Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[10] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Prasanta Kumar Ghosh, Sriram Ganapathy, et al., "Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.

[11] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, MD Iftikhar Hussain, and Muhammad Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, pp. 100378, 2020.

[12] Edresson Casanova, Arnaldo Candido Jr, Ricardo Corso Fernandes Jr, Marcelo Finger, Lucas Rafael Stefanel Gris, Moacir A Ponti, and Daniel Peixoto Pinto da Silva, "Transfer learning and data augmentation techniques to the covid-19 identification tasks in compare 2021," 2021.

[13] Tarik Alafif, Abdul Muneeim Tehame, Saleh Bajaba, Ahmed Barnawi, and Saad Zia, "Machine and deep learning towards covid-19 diagnosis and treatment: survey, challenges, and future directions," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, pp. 1117, 2021.

[14] Jivitesh Sharma, Ole-Christoffer Granmo, and Morten Goodwin, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network.," in *INTERSPEECH*, 2020, pp. 1186–1190.

[15] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht, "Deep learning applications for covid-19," *Journal of big Data*, vol. 8, no. 1, pp. 1–54, 2021.

[16] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler, "Covid-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, p. 104572, 2021.

[17] Tianhao Yan, Hao Meng, Emilia Parada-Cabaleiro, Shuo Liu, Meishu Song, and Björn W Schuller, "Coughing-based recognition of covid-19 with spatial attentive convlstm recurrent neural networks," *INTERSPEECH*, pp. 4154–4158, 2021.

[18] Micheal Lanham, "Attention is all we need!," in *Generating a New Reality*, pp. 195–222. 2021.

[19] Steffen Illium, Robert Müller, Andreas Sedlmeier, and Claudia-Linnhoff Popien, "Visual transformers for primates classification and covid detection," *INTERSPEECH*, pp. 451–455, 2021.

[20] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[21] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125868–125881, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need (2017)," *arXiv preprint arXiv:1706.03762*, 2021.

[23] Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6675–6679.