A Temporal-oriented Broadcast ResNet for COVID-19 Detection

1st Xin Jing
Chair of Embedded Intelligence
for Health Care & Wellbeing
University of Augsburg
Augsburg, Germany
xin.jing@uni-a.de

2nd Shuo Liu
Chair of Embedded Intelligence
for Health Care & Wellbeing
University of Augsburg
Augsburg, Germany
shuo.liu@uni-a.de

3rd Emilia Parada-Cabaleiro
Institute of Computational Perception
Johannes Kepler University Linz
Linz, Austria
emilia.parada-cabaleiro@jku.at

4th Andreas Triantafyllopoulos
Chair of Embedded Intelligence
for Health Care & Wellbeing
University of Augsburg
Augsburg, Germany
andreas.triantafyllopoulos@uni-a.de

5th Meishu Song
Educational Physiology Laboratory
University of Tokyo
Tokyo, Japan
meishu@p.u-tokyo.ac.jp

6th Zijiang Yang
Chair of Embedded Intelligence
for Health Care & Wellbeing
University of Augsburg
Augsburg, Germany
zijiang.yang@ieee.org

7th Björn W. Schuller
Group on Language, Audio, & Music
Imperial College
London, United Kingdom
bjoern.schuller@imperial.ac.uk

***Abstract*—Detecting COVID-19 from audio signals, such as breathing and coughing, can be used as a fast and efficient pre-testing method to reduce the virus transmission. Due to the promising results of deep learning networks in modelling time sequences, we present a temporal-oriented broadcasting residual learning method that achieves efficient computation and high accuracy with a small model size. Based on the EfficientNet architecture, our novel network, named Temporal-oriented ResNet (TorNet), constitutes of a broadcasting learning block. The network obtains useful audio-temporal features and higher level embeddings effectively with much less computation than Recurrent Neural Networks (RNNs), typically used to model temporal information. TorNet achieves 72.2% Unweighted Average Recall (UAR) on the INTERPSEECH 2021 Computational Paralinguistics Challenge COVID-19 cough Sub-Challenge, by this showing competitive results with a higher computational efficiency than other state-of-the-art alternatives.**

***Index Terms*—SARS-CoV2 detection, deep neural network, efficient neural network, efficient CNN, residual learning**

I. Introduction

COVID-19 pandemic is one of the main health challenge for our world [1]. Although there are rapid testing methods, their efficiency is often limited by the capacity of the testing equipment. Indeed, ubiquitous *low-cost* methods for detecting COVID-19 are still being explored. Deep Neural Networks (DNNs) have been growing in popularity in recent years, setting the state-of-art in a variety of tasks, including COVID-19 detection from audio signals, e. g., patients' breathing and coughing [2], [3].

The temporal component is an essential characteristic of audio signals. And the successful application of RNNs [4]–[6] and transformers [7]–[9] to audio data illustrates the importance of temporal features for audio tasks. Nevertheless, the complexity of these network structures, unlike CNNs, increases the computational complexity and reduces the training stability. In the present work, we propose a temporal broadcast residual convolution block, i. e., the Alternating Broadcast Block (AB Block), in which we average the 2D features in the frequency dimension to guide the network's focus on the temporal features. Inspired by the EfficientNet [10] architecture , we introduce a new deep learning network named Temporal-oriented ResNet (TorNet) that contains several AB Blocks to make full use of the temporal information in the audio segments. Furthermore, we also adopt Instance Normalisation [11] to assist the network to find the relevant feature areas of the Mel-spectrogram. We evaluate the efficiency of TorNet on the detection of COVID-19 from coughing signals, using the audio dataset from the INTERSPEECH 2021 Computational Paralinguistics Challenge's COVID-19 cough sub-challenge (CCS) [12].

The remainder of our paper is organised as follows: We summarise the related research in Section II. Then, we present our network architecture and describe the experimental settings in Sections III and IV, respectively. In Section V, we discuss the results. Finally, in Section VI, we conclude with a brief summary and outline future directions.

II. Related Works

Data representations such as Mel-Spectrograms can be seen from two different perspectives: either as an image, or as an audio sequence. This duality leads to the use of a variety of DNN architectures typical of both Computer Vision (CV) and the audio domain [13]–[15].

On the one side, previous work has shown that with Mel Frequency Cepstral Coefficients (MFCC) and log Mel-Spectrograms, 1D audio data can be transformed into 2D matrices [16]–[18]. This makes it possible to directly apply CNNs, typically from CV, and which have become the mainstream in Computer Audition [19]–[21]. In the task of COVID-19 detection, Chang et al. [9] studied the performance of classical CNNs pretrained on the FluSense [22] database, collected to track influenza-related indicators, such as cough and sneezes [22]. Similarly, Casanova et al. [23] employed transfer learning from pretrained audio neural networks with different data augmentation techniques.

On the other side, as audio data is inherently a type of temporal sequence [24], RNNs [6] and LSTM [25] have been fully adopted to handle the temporal information in several tasks. For instance, Hassan et al. [26] and Pahar et al. [27] evaluated the role of different audio features as input for LSTM-based classification of COVID-19. Similarly, Yan et al. [5] introduced the Spatial Attentive ConvLSTM-RNN (SACRNN), able to identify the most valuable features through an embedded temporal attention. Various efforts have also explored more efficient CNNs using residual network approaches and ensembles on audio data [28]–[30]. In particular, Byeonggeun et al. [31] used a residual broadcast block to retrieve temporal features by averaging the frequency features. Finally, Zhang et al. [32] proposed a hierarchical structure called pyramidal temporal pooling (PTP), which can retrieve temporal information by stacking a global PTP layer on multiple local ones.

III. Proposed Method

In this section, we propose Temporal-Oriented ResNet (TorNet), which we present for COVID-19 recognition. In addition, we also propose an Alternating Broadcast Block (AB Block), which contains several Broadcast Residual Blocks (BC ResBlock) [31] Finally, we use Frequency-wise Instance Normalisation for better domain generalisation [33].

A. Broadcast Residual Block

The original ResNet [34] block is described by $y = x + f(x)$, with $f(x)$ being the residual function, and x and y denoting the input and output features, respectively. To highlight the frequency convolution over all blocks, an auxiliary 2D residual connection is added from 2D features. To summarise, the BC-ResBlock can be presented as:

$$y = x + f_2(x) + BC(f_1(avgpool(f_2(x)))). \qquad (1)$$

In Equation 1, the 2D feature part f_2 consists of a 3x1 frequency depth-wise convolution followed by SubSpectral normalisation (SSN) [35], which splits the input frequency into multiple groups and normalises them separately. Finally, 2D features are averaged over the frequency dimension. And the BC, which indicates the Broadcasting, im- plies the expanding operation to frequency dimension

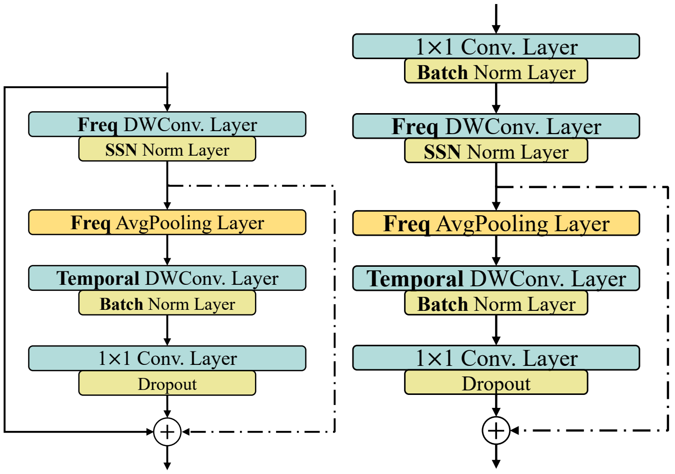

Fig. 1. **Left: Normal BC ResBlock.** A ResNet structure which firstly compresses the frequency dimension of the feature map by average pooling, and then broadcasts the temporal feature to the original feature map. **Right: Transition BC Block.** It contains two additional 1×1 convolution layers to change the channel number without the identity shortcut.

f_1 is a combination between a 1x3 temporal convolution with Batch Norm and Swish activation [36] followed by a 1x1 point-wise convolution. It will expand the feature map from $\mathbb{R}^{1\times w}$to $\mathbb{R}^{h\times w}$.

A normal BC ResBlock (cf. left in Figure 1) remaps the temporal information to the original feature map, so it has the same input and output dimensions. Meanwhile, a transition block is used, with the following modifications:

1) When channels do not have the same size, we add a transition block with Batch Norm and ReLU activation;
2) There is no identity shortcut.

B. Alternating Broadcast Block

With the BC ResBlock, it is possible to turn the features into a higher dimensionality while broadcasting the temporal information to the whole feature map. As shown in Figure 2, we propose a flexible structure of the Alternating Broadcast Block (AB Block), which mainly contains a set of BC ResBlocks and a convolution layer. The AB Block can be easily widened or deepened by simply adding a larger number of Normal BC ResBlocks.

As shown in Figure 1 (left), average pooling is used before temporal depth-wise (DW) convolution, which yields information loss in the frequency dimension (an inevitable side effect of using the BC ResBlock). In order to reduce the impact of information loss, the last layer of the AB Block is set to a convolution layer, followed by a Batch Norm layer and a ReLU activation layer. The main task of the convolution layer is to capture the global information of the temporal-based feature map, while retaining the local information learnt in the

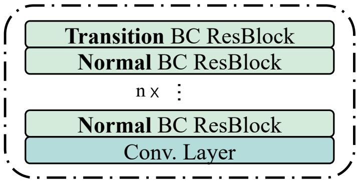

Fig. 2. The Alternating Broadcast Block (AB Block).

previous layer and projecting them to the higher dimensions of the original inputs. Thus, by using the proposed block, we can achieve enhanced frequency-aware temporal 2D features.

To achieve a better domain generalisation, we apply Instance Normalisation [11] (IN), an approach that normalises across each channel in each training example. Since IN does not rely on batch information, its implementation is kept the same for both the training and the testing phases.

C. Temporal-oriented ResNet (TorNet)

Finally, we design the Temporal-oriented ResNet (TorNet) for the COVID-19 detection task as shown in Figure 3. Details on the TorNet structure are given in Table I. As shown in Figure 3, TorNet contains four main stages. The first stage has a 3×3 convolution layer with a 2×2 max-pooling layer on the front to downsample both the time and frequency dimensions. The second stage is a typical residual block with two AB Blocks, where every AB Block will double the channel while halving the frequency dimension to get a higher-level embedding. In the residual shortcut, we added a batch norm layer and used maxpooling to control the size of the receptive field. This is followed by an Instance Normalisation layer between stage 2 and stage 3. Stage 3 shares the same structure as stage 2 with minor differences, i. e., the number of channels is doubled, and the dimension of the feature map does not change. After the second IN layer, the feature map is turned into a 3D tensor $[batch\_size, time, out\_channel \times N\_mel]$. Finally, two fully connected layers are added as classification layers.

IV. Experiments

A. Dataset

The CCS database [12] consists of 929 cough recordings (1.63 hours) from 397 participants presenting either a positive or negative COVID-19 test. Participants were asked to provide one to three forced coughs in each recording[1]. All recordings in the CCS database were resampled and converted to 16 kHz and mono/16 bits.

The official training, validation, and test sets used in the ComParE challenge are used in all our experiments.

B. Experiment settings

For data pre-processing, we standardise the length of the audio data to 10 seconds. The shorter samples are repeated until they match the target length. As input features, we use 40-dimensional log Mel-Spectrograms with a 64 ms window length and a 16 ms frame shift. We also extract deltas and delta-delta of log Mel-Spectrograms and concatenate them as input features. For all models, we use the Adam optimiser with an epsilon value of 10^{-8}, a mini-batch size of 16, and a learning rate of 10^{-5}. As indicated by [35], the sub-bands of SSN in AB Blocks were all set to 5 and the dropout rate was always $p = 0.1$ except for the last layer, where it was set to $p = 0.5$. We also tried data augmentation (mixup, Spec augmentation) methods, but there was no noticeable performance gain in our task, thus omitting them for brevity. All the models were developed on Pytorch 1.8.1 and trained on a single Nvidia RTX 3090 GPU.

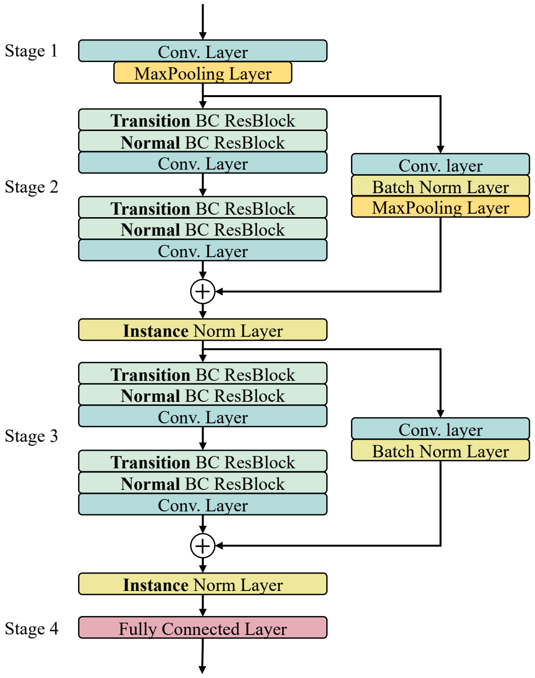

Fig. 3. TorNet for COVID-19 detection.

TABLE I
First 3 stages in TorNet. Each row is a sequence of one module with an input shape of $channel \times frequency \times time$.

#	Input	Operator	Stride	Output
Stage 1	1 × 40 × 512	conv2d 3x3	1	32
	32 × 40 × 512	maxpool 2x2	2	32
Stage 2	32 × 20 × 256	AB Block	(2, 1)	64
	64 × 10 × 256	AB Block	(2, 1)	128
	128 × 5 × 256	IN	-	128
Stage3	128 × 5 × 256	AB Block	1	256
	256 × 5 × 256	AB Block	1	512
	512 × 5 × 256	IN	-	512

C. Proposed Experiments

To verify the efficiency of the proposed TorNet, as well as the effectiveness of the temporal features, we developed four additional ResNet-based methods for comparison:

- ResNet-10: has an identical structure as TorNet but uses a convolution layer and maxpooling to control the size of

[1] https://www.COVID-19-sounds.org/; retrieved 12 March 2022

TABLE II
UNWEIGHTED AVERAGE RECALL (UAR), NETWORKS' PARAMETERS AND THE OVERALL RESULTS WITH 95% BOOTSTRAP CONFIDENCE INTERVALS (CI) USING 1000 SAMPLES (WITH REPLACEMENT). THE BEST RESULT IS MARKED IN **BOLD**, AND THE SECOND BEST ONE IS UNDERLINED.

Method	# Param(M)	UAR (%)	CI on Test(%)
End2You [12]	-	64.7	56.2 - 73.5
Fusion [12] (official baseline)	-	73.9	66.0 - 82.6
CNN14 [23]	79.67	75.9	-
The Vision Transformer (ViT) [37]	-	72.0	-
ResNet-10 (baseline)	5.12	66.9	57.7 - 71.8
ResNet-10+LSTM	6.17	66.7	64.2 - 81.6
ResNet-10+LSTM+attention	6.23	**70.5**	63.9 - 80.1
ResNet-10+Transformer	58.90	68.0	63.3 - 79.6
TorNet (AB Block without last conv)	**1.32**	65.5	60.1 - 77.6
TorNet (only Transition Block)	4.09	69.4	58.5 - 72.6
TorNet w/o InstanceNorm	4.46	70.2	59.4 - 72.8
TorNet w InstanceNorm	4.46	**72.2**	71.5 - 88.6

the feature map. It is used as a baseline model.

- ResNet-10 + LSTM: introduces a layer of a standard LSTM network at the output of the ResNet-10.
- ResNet-10 + LSTM + Attention: adds a 4-head multihead attention module to extract more centralised temporal-frequency feature maps.
- ResNet-10 + Transformer: adds 2 transformer encoder layers for locating and re-extracting the most relevant features of the audio segments.

V. RESULTS AND DISCUSSION

Our experimental results obtained on the binary task of COVID-19 detection are presented in Table II. The upper part of Table II displays the results from previous works on the CSS dataset. The middle part shows the results obtained from the four additional ResNet-based methods presented for comparison. Finally, the results for the series of experiments with same network hyperparameters and different modules are given in the lower part. Unweighted Average Recall (UAR) is reported as the evaluation metric. For each method, the results on the test set are obtained by using the model achieving the highest UAR on the validation set.

Our proposed TorNet, based on the combination of AB Block and residual learning's results, reached up to 72.2% UAR. The results show that all experiment results on the Tornet outperform the official baseline (End2You 64.7%) but still lag behind other approaches. Unlike the herein presented one, [23] achieved 75.9% UAR based on a large-scale transfer learning model. Their CNN14 model is pre-trained on Audioset, which means a longer training time and higher computational effort. Indeed, the parameters of CNN14 (79.67 millions) are almost 18 times more than the number of parameters compared to TorNet (4.46 millions) – thus showing that TorNet has a higher computational efficiency. Similarly, the baseline fusion framework for the CCS Sub-Challenge fuses multiple best models to obtain the final results (73.9% UAR), which also results in a far higher computational complexity than our TorNet. Overall, this shows that TorNet can achieve competitive performance without pre-training or fusion while also using far lower computational resources.

Meanwhile, the results show that the extraction of temporal information can improve the final UAR results, as shown by the combination of ResNet + LSTM (cf. 70.5%, in the bold, in the middle part of Table II).

Since our goal is to investigate to which extent it is possible to model temporal information while improving computational efficiency with DNNs, we also set up four comparison experiments based on TorNet. In these, we keep all training parameters consistent in order to assess the impact of different modules, i. e., the use of a convolution layer in the AB Block, the Normal BC ResBlock, and Instance Normalisation on the overall performance of TorNet.

The lower part of table II contains an ablation study of the components introduced in this work. TorNet without the last convolution layer in the AB block achieves only 65.5% UAR, while when convolution layers are introduced, there is a performance improvement of nearly 5.0% (cf. 70.2%, underlined in the lower part of Table II). This is because in each AB Block, the BC ResBlock has the ability to broadcast the temporal features to the original feature map, but loses a portion of the frequency features. By introducing an extra convolution layer, we eliminate the influence that this loss of granular details entails, thus obtaining a better overview for the feature map, which results in a sizable performance increase.

VI. CONCLUSION

In this work, we proposed an AB Block that can efficiently exploit the temporal information in audio sequences. Based on the AB Block with residual learning, we proposed a flexible, lightweight, and time-oriented network – TorNet. TorNet has a typical ResNet structure, but we replaced the convolution module with the AB Block. Competitive results highlight the high computational efficiency and robustness of TorNet, a promising architecture that offers new insights for the detection of COVID-19.

Benefiting from the AB Block structure and the Residual Learning architecture, the integration of TorNet into the latest frameworks, e. g., self-supervised learning, is also a promising direction to be further investigated.

VII. ACKNOWLEDGEMENTS

This work was funded by the China Scholarship Council (CSC), Grant #202006290013 and by the DFG (German Research Foundation), Reinhart Koselleck-Project AUDI0NOMOUS (Grant No. 442218748).

References

[1] "WHO coronavirus (covid-19) dashboard," 2022. [Online]. Available: https://covid19.who.int/

[2] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the sars-cov-2 corona crisis," *Frontiers in digital health*, vol. 3, no. 14, 2021.

[3] J. Andreu-Perez, H. Pérez-Espinosa, E. Timonet, M. Kiani, Giron-Perez *et al.*, "A generic deep learning based cough analysis system from clinically validated samples for point-of-need covid-19 test and severity levels," *IEEE Transactions on Services Computing*, no. 13, pp. 1–7, 2021.

[4] A. Baird, S. Amiriparian, M. Milling, and B. W. Schuller, "Emotion recognition in public speaking scenarios utlising an lstm-rnn approach with attention," in *Proceeding of the IEEE Spoken Language Technology*, virtual, 2021, pp. 397–402.

[5] T. Yan, H. Meng, E. Parada-Cabaleiro, S. Liu, M. Song, and B. W. Schuller, "Coughing-based recognition of covid-19 with spatial attentive convlstm recurrent neural networks," in *Proceedings of Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021, pp. 4154–4158.

[6] A. A. Alvarez and F. Gómez, "Motivic pattern classification of music audio signals combining residual and lstm networks," *International journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 6, 2021.

[7] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Proceedings of Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021, pp. 4154–4158.

[8] M. Song, E. Parada-Cabaleiro, Z. Yang, K. Qian, B. W. Schuller, and Y. Yamamoto, "Parallelising cnns and transformers: A cognitive-based approach for automatic recognition of learners' english proficiency," in *Proceedings of the Intelligent Human Systems Integration*, Venice, Italy, 2022, 6 pages, to appear.

[9] Y. Chang, X. Jing, Z. Ren, and B. W. Schuller, "CovNet: A transfer learning framework for automatic covid-19 detection from crowd-sourced cough sounds," *Frontiers in Digital Health*, vol. 3, 2022, 10 pages.

[10] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016, 6 pages.

[12] B. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, Lefter *et al.*, "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021, 5 pages.

[13] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[14] Z. Yang, K. Qian, Z. Ren, A. Baird, Z. Zhang, and B. Schuller, "Learning multi-resolution representations for acoustic scene classification via neural networks," in *Proceedings of the Sound and Music Technology*, Harbin, China, 2020, pp. 133–143.

[15] X. Jing, M. Song, A. Triantafyllopoulos, Z. Yang, and B. W. Schuller, "Redundancy reduction twins network: A training framework for multi-output emotion regression," in *Proceedings of the ICML Expressive Vocalizations Workshop and Competition*, 2022, 5 pages, to appear.

[16] M. Song, A. Mallol-Ragolta, E. Parada-Cabaleiro, Z. Yang, S. Liu, Z. Ren, Z. Zhao, and B. W. Schuller, "Frustration Recognition from Speech during Game Interaction Using Wide Residual Networks," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 76–86, 2021.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] Z. Yang, S. Liu, M. Song, E. Parada-Cabaleiro, and B. Schuller, "Adventitious Respiratory Classification using Attentive Residual Neural Networks," in *Proceedings of Annual Conference of the International Speech Communication Association*, Shanghai, China, 2020, pp. 2912–2916.

[20] A. Triantafyllopoulos, S. Liu, and B. Schuller, "Deep Speaker Conditioning for Speech Emotion Recognition," in *Proceedings of IEEE International Conference on Multimedia and Expo*, Shenzhen, China, 2021, pp. 1–6.

[21] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *arXiv preprint arXiv:2203.01205*, 2022.

[22] F. Al Hossain, A. A. Lover, G. A. Corey, N. G. Reich, and T. Rahman, "Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–28, 2020.

[23] E. Casanova, A. Candido Jr., R. C. Fernandes Jr., M. Finger, L. R. S. Gris, M. A. Ponti, and D. P. Pinto da Silva, "Transfer learning and data augmentation techniques to the covid-19 identification tasks in compare 2021," in *Proceedings of Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021, pp. 446–450.

[24] L. Vrysis, N. Tsipas, C. Dimoulas, and G. Papanikolaou, "Extending temporal feature integration for semantic audio analysis," in *Proceedings of the Audio Engineering Society*, no. 9808, Berlin, Germany, 2017.

[25] J. Deng, B. Schuller, F. Eyben, D. Schuller, Z. Zhang, H. Francois, and E. Oh, "Exploiting time-frequency patterns with lstm-rnns for low-bitrate audio restoration," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1095–1107, 2020.

[26] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 detection system using recurrent neural networks," in *Proceedings of the Communications, Computing, Cybersecurity, and Informatics*, Sharjah, UAE, 2020, pp. 1–5.

[27] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *arXiv preprint arXiv:2012.01926*, 2020, 13 pages.

[28] T. V. Kumar, R. S. Sundar, T. Purohit, and V. Ramasubramanian, "End-to-end audio-scene classification from raw audio: Multi time-frequency resolution cnn architecture for efficient representation learning," in *Proceedings of the IEEE Signal Processing and Communication*, Bangalore, India, 2020, pp. 1–5.

[29] L. Vrysis, N. Tsipas, I. Thoidis, and C. Dimoulas, "1D/2D deep cnns vs. temporal feature integration for general audio classification," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 66–77, 2020.

[30] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *Proceedings of the the International Conference on Acoustics, Speech, & Signal Processing*, Toronto, Ont., Canada, 2021, pp. 905–909.

[31] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," *arXiv preprint arXiv:2106.04140*, 2021, 5 pages.

[32] L. Zhang, Z. Shi, and J. Han, "Pyramidal temporal pooling with discriminative mapping for audio classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 770–784, 2020.

[33] H. Nam and H.-E. Kim, "Batch-Instance normalization for adaptively style-invariant neural networks," in *Proceedings of the Neural Information Processing Systems*, Montréal, QC, Canada, 2018, 12 pages.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016.

[35] S. Chang, H. Park, J. Cho, H. Park, S. Yun, and K. Hwang, "Subspectral normalization for neural audio data processing," Toronto, Ont., Canada, 2021, pp. 850–854.

[36] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017, 13 pages.

[37] S. Illium, R. Müller, A. Sedlmeier, and C.-L. Popien, "Visual transformers for primates classification and covid detection," in *Proceedings of Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021, pp. 451–455.