2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Transformer-based CNNs: Mining Temporal Context Information for Multi-sound COVID-19 Diagnosis

Yi Chang[1], Zhao Ren[2], Björn W. Schuller[1,2]

Abstract— **Due to the COronaVIrus Disease 2019 (COVID-19) pandemic, early screening of COVID-19 is essential to prevent its transmission. Detecting COVID-19 with computer audition techniques has in recent studies shown the potential to achieve a fast, cheap, and ecologically friendly diagnosis. Respiratory sounds and speech may contain rich and complementary information about COVID-19 clinical conditions. Therefore, we propose training three deep neural networks on three types of sounds (breathing/counting/vowel) and assembling these models to improve the performance. More specifically, we employ Convolutional Neural Networks (CNNs) to extract spatial representations from log Mel spectrograms and a multi-head attention mechanism in the transformer to mine temporal context information from the CNNs' outputs. The experimental results demonstrate that the transformer-based CNNs can effectively detect COVID-19 on the DiCOVA Track-2 database (AUC: 70.0 %) and outperform simple CNNs and hybrid CNN-RNNs.**

I. Introduction

The COronaVIrus Disease 2019 (COVID-19) has emerged as a global pandemic since early 2020. To date, as reported by the Johns Hopkins University's Center for Systems Science and Engineering, there have been more than 3,100,000 deaths and 140,000,000 positive cases globally [1]. Learning how to diagnose COVID-19 early and reliably is critical for rapid isolation of patients, an appropriate prognosis, and successful dissemination risk reduction. However, standard diagnosis methods, such as reverse transcription polymerase chain reaction tests and chest computed tomography, require professionals and lab environments [2] and can produce essential amounts of waste. Therefore, developing cheap, easy, and ecologically friendly processing tools for the pre-screening of COVID-19 is urgently needed.

With the rapid development of machine learning techniques in computer audition, automatic detection of diseases (e. g., cardiological diseases and depression) from body acoustic signals provides a novel perspective for a fast and non-invasive diagnosis [3]. As for COVID-19, respiratory sounds (e. g., cough and breathing) and speech have proven to be promising in detecting COVID-19 [4], [5] and analysing the mental/physical status (e. g., anxiety and fatigue) of COVID-19 patients [6]. Many recent studies [5]–[8] have collected different types of sounds through either well-designed experiments or crowdsourcing platforms. The acoustic features were also analysed [5] and used for developing machine-learning-based diagnosis models [4]–[6]. However, only a few studies considered the potentially complementary information in multiple types of sounds. For instance, both breathing and cough sounds were analysed in [5], [9], and cough, breathing, and speech were explored in [10]. Motivated by these studies, to better exploit the latent features in different sounds, three types of sounds (breathing/counting/vowel) from the crowdsourced DiCOVA Track-2 database [11] are analysed in this work.

In the past years, deep learning has shown its strong capability in extracting high-level representations and beaten traditional machine-learning techniques in the artificial intelligence community [12]. Particularly, Convolutional Neural Networks (CNNs) have been successfully used to process time-frequency representations (e. g., log Mel spectrograms) extracted from audio signals [13]. However, conventional CNNs learn spatial features well, but mostly fail to capture the long-term temporal dependencies in sequential sounds [14]. The temporal dynamics in the sounds may contain valuable information about the disease. Therefore, some research [15] applied Recurrent Neural Networks (RNNs) for COVID-19 detection with respiratory data. Despite RNNs' advantage of learning sequential information, RNNs cannot be developed in parallel due to their sequential processing. Moreover, it is hard for RNNs to attend long dependencies because of the gradient vanishing or decay, even though some bi-directional models widen the dependency range [16]. To solve the aforementioned problems, the transformer [17] takes the input as a whole and retains global dependencies between the input and the output by a multi-head attention mechanism, rather than a recurrent structure along time steps. Through the multi-head attention, the computation parallelisation is achieved. In many studies for audio processing [13], [18], the transformer has shown better performance than RNNs.

To the authors' best knowledge, there is little research in COVID-19 diagnosis training transformer-based models on acoustic data. In this work, to mine the temporal dependencies in the three classes of sounds (a respiratory sound and two speech-related types), the multi-head attention mechanism in the transformer framework is employed to process the high-level representations from a CNN model inspired by a related study [13]. Three transformer-based CNNs trained on the three types of sounds are further assembled for exploiting complementary multi-sound information.

*This work is supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS), and the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

[1]Yi Chang and Björn W. Schuller are with GLAM – the Group on Language, Audio, & Music, Imperial College London, SW7 2AZ London, UK. y.chang20@imperial.ac.uk, schuller@ieee.org

[2]Zhao Ren and Björn W. Schuller are with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany. zhao.ren@informatik.uni-augsburg.de

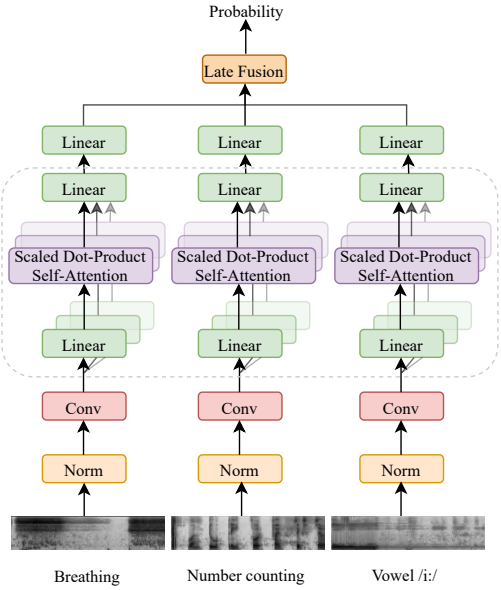

Fig. 1. The framework of the transformer-based CNNs for multi-sound COVID-19 diagnosis. The content inside the dashed line indicates three multi-head attention mechanisms for processing the three types of sounds.

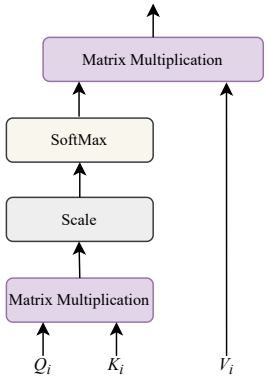

Fig. 2. The pipeline of the scaled dot-product self-attention.

II. Methodology

The framework of our proposed transformer-based CNNs is depicted in Fig. 1. As outlined, three types of sounds are applied in this work: breathing, number counting, and vowel uttering /i:/. For each type of sound, the log Mel spectrogram is fed into a separate transformer-based CNN model. After batch normalisation, we apply $2D$ convolution layers for high-level features and feed the features into the multi-head attention mechanism, as shown in the dashed line circle in Fig. 1. Following the multi-head attention, a fully connected layer is utilised for time-frame-wise COVID-19 positive probabilities, which are then averaged to provide the final probability for the audio sample. Finally, to assemble the predictions of the three sounds, a late fusion method, average or max, is imposed. In the following, the aforesaid CNNs and multi-head attention mechanism will be described.

A. Convolutional Neural Networks

CNNs were originally proposed for image processing tasks [14]. Typically, for classification or regression, CNNs are composed of three classes of layers: convolutional layers for feature extraction, pooling layers for downsampling, and fully connected layers for the final outputs.

For audio processing, CNNs have been successful in extracting effective and highly abstract representations from log Mel spectrograms [19]. In this work, a batch normalisation layer is firstly applied on the (T_{mel}, F_{mel}) log Mel spectrograms, where T_{mel} is the number of time frames, and F_{mel} denotes the number of Mel bins. Afterwards, we employ a series of convolutional blocks, each of which contains several convolutional layers and a local average pooling layer. Every convolutional layer herein is composed of a convolutional operation followed by a batch normalisation and a Rectified Linear Unit (ReLU) activation function. In this way, the (C, T, F) feature maps are obtained, where C is the channel number, T stands for the dimension along the time axis, and F denotes the dimension along the frequency axis. Finally, a global average pooling layer is utilised on the frequency axis, generating representations with a shape of (C, T).

B. Sequence modelling by Transformer

RNNs have been widely applied for sequence modelling in natural language processing [16]. Especially, most RNNs (e. g., Long Short-Term Memory (LSTM) RNNs and Gated Recurrent Unit (GRU) RNNs [16]) consider the sequence information with several hidden layers. In each hidden layer, a hidden state is determined not only by the current input but also by the previous hidden state, limiting the computation efficiency for inputs with longer sequences [16]. To overcome the constraint, the transformer framework was proposed in [17]. The transformer is based on the self-attention mechanism to calculate the weight of each time step among the whole sequence, enabling computation parallelisation.

A transformer is composed of encoder and decoder stacks. Each encoding component generates embeddings from the input, and each decoding component converts the learnt embeddings to the output. Specifically, in this work, the multi-head attention mechanism from the encoder consisting of h (i. e., number of heads) scaled dot-product self-attentions is used. Firstly, the multi-head attention takes the transposed CNNs' output representations X with a shape of (T, C). Next, the query Q, key K, and value V are calculated by dot product of X and three transformer matrices $W^Q \in \mathbb{R}^{C \times d_k}$, $W^K \in \mathbb{R}^{C \times d_k}$, $W^V \in \mathbb{R}^{C \times d_v}$, respectively, where $d_k = d_v = \frac{C}{h}$. Afterwards, the Q, K, and V are split into h parts, generating Q_i, K_i, and V_i, where $i \in [1, h]$. A single scaled dot-product self-attention mechanism is then defined as

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i. \quad (1)$$

The dot product of Q_i and K_i^T is scaled by $\sqrt{d_k}$ to counteract the small gradients issue of the softmax function [17]. The softmax function is applied to generate the attention weights for all time steps. Afterwards, the output tensors from every single dot-product attention are concatenated and fed into a linear layer with a ReLU activation function.

TABLE I

PERFORMANCES [%] OF THE PROPOSED APPROACHES EVALUATED ON THE DICOVA TRACK-2 DATABASE FOR COVID-19 DETECTION. EACH PERFORMANCE OF CROSS-VALIDATION IS THE AVERAGE RESULT OVER THE FIVE FOLDS, FOLLOWED BY A STANDARD DEVIATION (STD).

		Cross-validation			Evaluation		
Sound type	Model	Sensitivity (std)	Specificity (std)	AUC (std)	Sensitivity (std)	Specificity (std)	AUC (std)
Breathing	CNN	**81.1 (0.4)**	**51.8 (1.3)**	**72.1 (0.5)**	81.0 (0.0)	34.5 (11.4)	60.9 (2.8)
	CNN-BiLSTM	81.0 (0.0)	51.2 (2.9)	70.0 (2.7)	**81.0 (0.0)**	**46.1 (10.7)**	**63.1 (2.1)**
	CNN-BiGRU	81.0 (0.0)	50.0 (4.7)	69.4 (2.3)	81.0 (0.0)	33.7 (5.0)	58.1 (2.5)
	CNN-Transformer	81.0 (0.0)	46.8 (4.9)	68.1 (2.3)	81.0 (0.0)	41.7 (3.1)	60.6 (4.2)
Number-counting	CNN	**81.0 (0.0)**	**43.8 (3.4)**	**69.5 (1.2)**	81.0 (0.0)	39.7 (2.8)	64.3 (2.7)
	CNN-BiLSTM	81.0 (0.0)	41.1 (3.3)	68.4 (1.1)	81.0 (0.0)	44.9 (8.3)	65.2 (1.9)
	CNN-BiGRU	81.0 (0.0)	41.6 (1.5)	69.3 (1.1)	81.0 (0.0)	38.7 (5.7)	65.0 (2.2)
	CNN-Transformer	81.0 (0.0)	42.0 (5.3)	68.4 (1.6)	**81.0 (0.0)**	**46.9 (6.0)**	**67.1 (3.3)**
Vowel /i:/	CNN	81.0 (0.0)	26.4 (2.1)	58.6 (0.7)	81.0 (0.0)	39.6 (13.5)	67.6 (3.4)
	CNN-BiLSTM	81.0 (0.0)	29.4 (3.3)	60.3 (1.0)	**81.0 (0.0)**	**40.3 (14.5)**	**67.8 (4.8)**
	CNN-BiGRU	81.0 (0.0)	24.5 (1.9)	58.6 (1.4)	81.0 (0.0)	41.7 (13.2)	66.4 (5.2)
	CNN-Transformer	**81.3 (0.5)**	**30.7 (3.5)**	**60.8 (2.0)**	81.0 (0.0)	36.1 (12.0)	66.5 (2.7)
Late fusion-max	CNN	– (–)	– (–)	– (–)	81.0 (0.0)	42.8 (6.1)	66.4 (5.2)
	CNN-BiLSTM	– (–)	– (–)	– (–)	81.0 (0.0)	44.8 (6.0)	66.3 (0.9)
	CNN-BiGRU	– (–)	– (–)	– (–)	81.0 (0.0)	33.0 (9.2)	63.8 (2.5)
	CNN-Transformer	– (–)	– (–)	– (–)	**81.0 (0.0)**	**36.5 (4.6)**	**68.1 (1.2)**
Late fusion-avg	CNN	– (–)	– (–)	– (–)	81.0 (0.0)	37.3 (13.5)	67.6 (2.4)
	CNN-BiLSTM	– (–)	– (–)	– (–)	81.0 (0.0)	42.2 (5.7)	69.4 (1.6)
	CNN-BiGRU	– (–)	– (–)	– (–)	81.0 (0.0)	38.0 (5.5)	67.0 (2.9)
	CNN-Transformer	– (–)	– (–)	– (–)	**81.0 (0.0)**	**39.5 (6.8)**	**70.0 (2.0)**

III. EXPERIMENTAL RESULTS

A. Database

The crowd-sourced Track-2 dataset of the DiCOVA challenge 2021 [8], [11] is used to verify the effectiveness of our proposed approach. With the target of analysing multiple sounds for COVID-19 diagnosis, three types of sounds were recorded from each subject: (i) deep breathing, (ii) number counting (normal pace), and (iii) vowel uttering /i:/. All recordings were labelled as one of the two classes: *COVID-19 negative* and *COVID-19 positive*. For each type of sound, $1,199$ audio files (negative: $1,118$, positive: 81) were recorded, and re-sampled into 44.1 kHz. The whole dataset was split into a development set (negative: 930, positive: 60) and an evaluation set (negative: 188, positive: 21). The development set was then partitioned according to the officially provided 5-fold cross-validation. In each fold, the training subset consists of 744 negative samples and 39 positive samples, while the validation subset contains 186/21 negative/positive ones. Similar to the DiCOVA challenge paper [11], the AUC, Area under the Receiver Operating Characteristic (ROC) Curve, works as the main evaluation metric and the specificity is calculated at $80\,\%$ sensitivity for more focus on correctly detecting COVID-19 positive samples. Specifically, the AUC is computed using the trapezoidal rule with a granularity of 0.0001.

B. Experimental Setup

All audio files are firstly re-sampled from 44.1 kHz to 16 kHz for potentially faster progression [19]. For log Mel spectrograms sharing a same general shape before feeding into the CNNs, all audio recordings are unified with the same duration of 20 seconds – around the 80^{th} percentile of all audio durations. Specifically, waves after the 20 seconds are cut out of the original audio file, while sound samples shorter than 20 seconds are self-repeated. The log Mel spectrograms are then extracted with a 512-length sliding window, an overlap of 256 time frames, and 64 Mel bins.

With the limited training data, the mixup method [20] is applied to augment the log Mel spectrograms. From two different data samples (x_1, y_1) and (x_2, y_2), a new data sample (x, y) is generated by $x = \lambda x_1 + (1 - \lambda)x_2$ and $y = \lambda y_1 + (1 - \lambda)y_2$, where $\lambda \sim \text{Beta}(1, 1)$. The binary cross-entropy with logits loss is employed to optimise the neural networks. To mitigate the class imbalance, the weight parameter in the above loss function is set as the number of COVID negative samples over the number of positive samples. The optimiser is set to 'Adam' with an initial learning rate of 0.001, which is then reduced by a factor of 0.1 at each 10^{th} epoch for stabilisation. All training procedures are stopped at the 30^{th} epoch. We apply all data from the cross-validation folds to train the model. Afterwards, the generated model is evaluated on the evaluation set.

The CNNs consist of four convolutional blocks, in each of which there are two convolutional layers with a kernel size of $(3, 3)$. The number of the output channels for the four convolutional blocks are 64, 128, 256, and 512, respectively. Every two convolutional layers inside a convolutional block share an equal number of output channels. The first three local average pooling layers have a kernel size of $(2, 2)$, and the final one has a kernel size of $(1, 1)$. For the multi-head attention, we set the head number h as 8.

For the experimental comparison, bidirectional LSTM-RNNs and bidirectional GRU-RNNs are used to compete with the multi-head attention. In each direction of the above RNNs, the number of features in the hidden state is 256, leading to 512-dimensional features at a time step.

C. Results and Discussions

For stabilisation, we run each model in five-fold cross-validation and evaluation five times, after which the average results and the standard deviations are calculated and indicated in Table I. Specifically, for each time running, cross-validation results indicate the average performance on all validation subsets, and evaluation results are calculated on the whole evaluation set. Due to the specificity calculation method and the granularity chosen for the AUC calculation in Section III-A, all proposed approaches have the sensitivity around 81.0 %. We can see that, for each type of sound, the CNN-BiLSTMs perform better than the CNN-GRUs on the evaluation set, perhaps since LSTM-RNNs could remember longer sequences than GRU-RNNs [21]; the performance of the transformer-based CNNs is comparable with that of the other three CNN models (i. e., CNN, CNN-BiLSTM, and CNN-BiGRU) on both the cross-validation and the evaluation set. Particularly, the transformer-based CNNs slightly outperform the other models on the number counting evaluation data and the vowel validation data. Furthermore, after the late fusion strategies, the transformer-based CNNs are improved over the corresponding single-sound transformer models. This indicates that detecting COVID-19 from multiple sounds is worthwhile for a more accurate diagnosis, probably due to the complementary information in the three sound types. Finally, the best result (AUC: 70.0 %) on the evaluation set is obtained by the average late fusion of the transformer-based CNNs. Although the other three models by late fusion are also improved over their single-sound versions, the fusion of the transformer-based CNNs achieves a bigger improvement, which shows the transformer's potential in multi-sound fusion.

IV. Conclusions and Future Work

In this study, the transformer-based Convolutional Neural Networks (CNNs) were employed to detect COronaVIrus Disease 2019 (COVID-19) from multiple sounds. The CNNs extracted spatial representations from log Mel spectrograms, whereas the transformer learnt temporal information in parallel. Three transformer-based CNNs were used to process three types of sounds, and they were further assembled with late fusion methods to generate more precise predictions. In the experiments, the transformer-based CNNs performed better than CNNs and hybrid CNN-RNNs.

In future efforts, more types of respiratory sounds, e. g., coughing, will be applied to train models for better performance. Moreover, up- and down-sampling methods [22] and the focal loss [23] can be utilised to further address the data imbalance issue. Because of the limited data size in this work, large-scale respiratory sound databases (e. g., COUGHVID [7]) should also be explored to develop more robust transformer-based models.

References

[1] Johns Hopkins University, USA. (29.04.2021) Covid-19 case tracker follow global cases and trends (updated daily). [Online]. Available: https://coronavirus.jhu.edu

[2] C. Long, H. Xu, Q. Shen, X. Zhang, B. Fan, C. Wang, B. Zeng, Z. Li, X. Li, and H. Li, "Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?" *European Journal of Radiology*, vol. 126, p. 108961, 2020.

[3] F. Dong, K. Qian, Z. Ren, A. Baird, X. Li, Z. Dai, B. Dong, F. Metze, Y. Yamamoto, and B. Schuller, "Machine listening for heart status monitoring: Introducing and benchmarking HSS – the heart sounds Shenzhen corpus," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2082–2092, 2019.

[4] B. Schuller *et al.*, "The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates," 2021, arXiv preprint:2102.13468, 5 pages.

[5] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proc. ACM SIGKDD*, virtual event, 2020, pp. 3474–3484.

[6] K. Qian, M. Schmitt, H. Zheng, T. Koike, J. Han, J. Liu, W. Ji, J. Duan, M. Song, Z. Yang, Z. Ren, S. Liu, Z. Zhang, Y. Yamamoto, and B. Schuller, "Computer audition for fighting the SARS-CoV-2 corona crisis – Introducing the multi-task speech corpus for COVID-19," *IEEE Internet of Things Journal*, 2021, 11 pages.

[7] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," 2020, arXiv preprint:2009.11644, 11 pages.

[8] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. Chetupalli, N. R., P. Ghosh, and S. Ganapathy, "Coswara – A database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 4811–4815.

[9] B. Schuller, H. Coppock, and A. Gaskell, "Detecting covid-19 from breathing and coughing sounds using deep neural networks," 2020, arXiv preprint:2012.14553, 6 pages.

[10] T. Xia, J. Han, L. Qendro, T. Dang, and C. Mascolo, "Uncertainty-aware COVID-19 detection from imbalanced sound data," 2021, arXiv preprint arXiv:2104.02005, 5 pages.

[11] A. Muguli, L. Pinto, N. R., N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy, and V. Nanda, "DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," 2021, arXiv preprint:2103.09148, 5 pages.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[13] Q. Kong, Y. Xu, and M. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[14] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, no. C, pp. 354–377, 2018.

[15] A. Hassan, I. Shahin, and M. B. Alsabek, "COVID-19 detection system using recurrent neural networks," in *Proc. CCCI*, Sharjah, United Arab Emirates, 2020, 5 pages.

[16] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. CVPR*, Salt Lake City, UT, 2018, pp. 5457–5466.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, 2017, pp. 6000–6010.

[18] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7829–7833.

[19] Z. Ren, A. Baird, J. Han, Z. Zhang, and B. Schuller, "Generating and protecting against adversarial attacks for deep speech-based emotion recognition models," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7184–7188.

[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, Vancouver, Canada, 2018, 13 pages.

[21] M. Golmohammadi, S. Ziyabari, V. Shah, E. Von Weltin, C. Campbell, I. Obeid, and J. Picone, "Gated recurrent networks for seizure detection," in *Proc. SPMB*, Philadelphia, PA, 2017, 5 pages.

[22] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," *Information Sciences*, vol. 384, pp. 174–190, 2017.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.