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

COVID-19 Detection with a Novel Multi-Type Deep Fusion Method using Breathing and Coughing Information

Shuo Liu[1*], Adria Mallol-Ragolta[1] and Björn W. Schuller[1,2]

Abstract— **This study explores the use of deep learning-based methods for the automatic detection of COVID-19. Specifically, we aim to investigate the involvement of the virus in the respiratory system by analysing breathing and coughing sounds. Our hypothesis resides in the complementarity of both data types for the task at hand. Therefore, we focus on the analysis of fusion mechanisms to enrich the information available for the diagnosis. In this work, we introduce a novel injection fusion mechanism that considers the embedded representations learned from one data type to extract the embedded representations of the other data type. Our experiments are performed on a crowdsourced database with breathing and coughing sounds recorded using both a web-based application, and a smartphone app. The results obtained support the feasibility of the injection fusion mechanism presented, as the models trained with this mechanism outperform single-type models and multi-type models using conventional fusion mechanisms.**

I. INTRODUCTION

At the time of writing, more than 125.1 M cases of the *Coronavirus Disease 2019* (COVID-19) have been confirmed worldwide, according to the *World Health Organization* (WHO). Despite the vaccines, massive population screenings will still play an important role to control the spread of COVID-19. The current medical instruments used for the detection of the virus are expensive, and burden public expenditures. Thus, there is a need to develop new digital, cost-efficient tools to diagnose and monitor this disease.

The symptomatology of COVID-19 includes serious involvements in the respiratory system. As a consequence, researchers have investigated the automatic detection of COVID-19 analysing chest x-ray images [1], chest CT scans [2], or signals produced by the respiratory system, such as breathing, coughing, or even the human voice [3], [4], [5], [6], [7]. Imran et al. [8] proposed an AI-powered smartphone app to detect COVID-19 from coughing information. In their solution, the recorded audio samples are transferred to an engine, which implements a cough detector followed by a COVID-19 diagnosis component. Faezipour and Abuzneid [9] motivated the study of the time and frequency components of the breathing sounds to detect the presence of the virus.

In this work, we aim at investigating the complementarity between breathing and coughing signals for the automatic detection of COVID-19. As baselines, we train single-type neural networks exploiting information from either breathing or coughing samples. The focus of this study resides in the comparison of four different methods to fuse breathing and coughing information when training multi-type models for the task at hand. Specifically, we propose a novel injection fusion mechanism, which considers the embedded feature representations learned from one data type to extract the embedded feature representations from the other data type.

The rest of the paper is laid out as follows: Section II describes the dataset analysed in this work, while Section III details the methodology followed. Section IV compiles and analyses the results obtained from the experiments performed, and Section V concludes the paper, and suggests potential directions for future works.

II. DATASET

We use the crowdsourced breathing and coughing recordings in [10], which were collected using both a web-based app, and an Android app. The participants were asked to cough three times, and to take three to five deep breaths to the app. Both apps additionally prompted users to report whether they tested positive for COVID-19, so this information can be used as the ground truth. A subset of the overall data collected has been released, so the research community can investigate the use of digital health-oriented systems for the automatic diagnosis of patients with COVID-19. This subset contains 62 COVID-positive patients, and 220 healthy patients, providing a total of 141 and 298 audio samples, respectively, including breathing and coughing samples.

As part of this work, we carefully listened to the audio samples available from this subset. We identified some healthy participants who submitted silent audio samples, or samples with a single respiratory sound. The participants who did not submit suitable recordings were excluded from our experiments with the aim to improve the quality of the data itself, and the models to be trained. Consequently, 210 healthy patients providing a total of 288 audio samples are considered in our experiments. The audio samples are split into participant-independent training, validation, and test partitions, containing 70 %, 10 % and 20 % of the total number of participants, respectively. In addition, we balance the COVID positive and negative patients in both the validation and test partitions (cf. Table I) to assess the models under similar conditions. For reproducibility purposes, the samples used, and the data partitioning are publicly available[1].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

[1] Shuo Liu, Adria Mallol-Ragolta, and Björn W. Schuller are with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany.

[2] Björn W. Schuller is also with GLAM – the Group on Language, Audio, & Music, Imperial College London, London, UK.

[*] Corresponding author: shuo.liu@uni-a.de

[1] https://github.com/EIHW/MultiTypeFusionForCOVID19Detection

TABLE I: Summary with the distribution of the data available over the training, validation, and test partitions. In this table, we depict the number of patients populating each partition, the total number of breathing (B) and coughing (C) frames available, and the total number of breathing-coughing frames pairs (B+C) combined. The information from each partition is provided independently for both COVID-19 (Pos) and healthy (Neg) patients.

	Train		**Validation**		**Test**		**Total**	
COVID-19	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
Patients	22	170	13	13	27	27	62	210
B	464	938	162	52	260	304	886	1294
C	207	410	55	27	109	165	371	602
B+C	1337	2047	370	126	678	695	2385	2868

Previous works use the whole audio recordings with variable lengths to train their models. We aim to train deep learning-based models using a supervised learning framework end-to-end. To achieve this, we truncate each breathing and coughing sample into frames of a fixed length to ensure a common format for the samples to feed into the model. The truncation is performed without overlapping information between consecutive frames, and we discard the recording tails with insufficient length.

From preliminary experiments not reported in this work, we empirically determined that the optimal length to analyse the coughing samples is 2 seconds, and to analyse the breathing samples, 2.5 seconds. The optimal length determined for the breathing samples guarantees the presence of at least one inhalation or exhalation process. Some coughing frames may include the preparation period before the actual cough, and therefore contain, for instance, a deep inhale.

This work aims to investigate the impact of fusing breathing and coughing information for the detection of COVID-19. To achieve this, we combine each breathing and coughing frame belonging to the same patient into a breathing-coughing frames pair. This strategy leads to an increase of the data available to train and evaluate our models. The distribution of the breathing frames, the coughing frames, and the breathing-coughing frames pairs available for each data partition is detailed in Table I. For each breathing and coughing frame, the first 40 *Mel-Frequency Cepstral Coefficients* (MFCCs) [11] extracted from their short-term power spectrum are used as input to our models.

III. METHODOLOGY

In this section, we first describe the *Convolutional Neural Network* (CNN) [12] implemented to detect COVID-19 from single-type information from a patient, using either a breathing frame or a coughing frame. Additionally, we detail the multi-type neural networks implemented, so both breathing and coughing frames can be modelled together, and discuss two conventional methods to fuse the features learned from the involved data types. For a better exploitation of the salient information from both breathing and coughing frames, in this paper we propose an injection fusion approach to replace conventional data fusion methods.

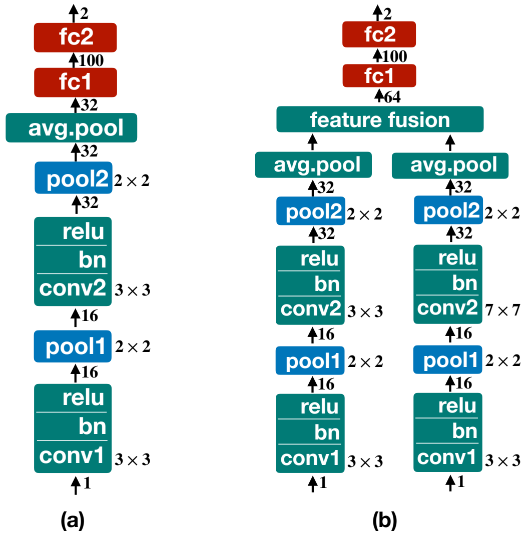

Fig. 1: Block diagram illustrating **(a)** the single-type model and **(b)** the multi-type fusion model implemented. The kernel size of each convolutional and max-pooling layer is given next to each block. The channel change is provided next to each transition arrow between adjacent blocks. **(a)** Single-type model in which either breathing or coughing frames are used as input. **(b)** Multi-type fusion model in which both breathing and coughing frames are simultaneously used as input. The feature fusion mechanisms applied to this network are either direct concatenation or 1×1 convolution.

A. Single-Type Model

The single-type model to process either breathing or coughing frames implements a CNN with two convolutional layers followed by two dense layers (cf. Figure 1(a)). While the goal of the former is to learn salient feature maps from the input data, the latter aims to project these embedded representations into a two-dimensional space. The output can then be used to infer whether patients have COVID-19 or not. The embedding features learned at the output of the convolutional block can be seen as deep-learned representations of the input data type.

Between the convolutional and the dense layers, the network utilises average pooling to squeeze the learned feature maps into embedding feature vectors. Batch normalisation [13] in each convolutional layer is also applied for a stable convergence. All layers in the network use the *Rectified Linear Unit* (ReLU) [14] as the activation function, with the exception of the last dense layer, which applies a Softmax activation function. The use of the Softmax activation function at the output of the network allows the interpretation of its outputs as the confidence scores with which the network classifies the current input sample into each possible class.

B. Multi-Type Fusion Model

To assess the performance of multi-type models exploiting both breathing and coughing frames simultaneously, we explore two widely used information fusion methods based

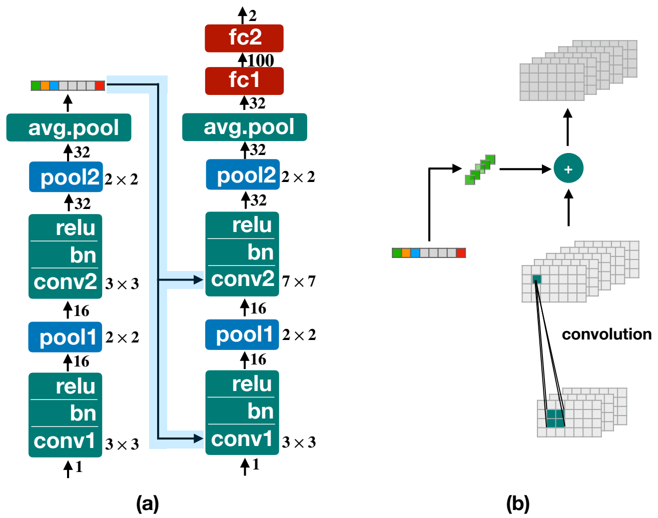

Fig. 2: Block diagram illustrating the multi-type injection fusion model proposed. **(a)** Multi-type injection fusion model, which feeds breathing and coughing samples separately in each branch of the network. **(b)** Injection fusion mechanism, which projects the learned embedding features to match the channel-dimension of the convolutional layer, and adds them to the feature map obtained at the output of the convolution.

on the embedding feature representations learned from the two data types. The multi-type network implemented consists of two branches, one processing the breathing frames and the other, the coughing frames (cf. Figure 1(b)).

The embedding feature vectors learned at the output of the convolutional blocks are fused together, and fed into the dense layers for the detection of COVID-19. Specifically, we consider two conventional information fusion approaches. The first approach concatenates the two embedding feature vectors learned into a single-channel representation. The second approach concatenates the two embedding feature vectors learned into a two-channel representation, and then applies a 1×1 convolution to project the two-channel representation into a single-channel representation.

C. Multi-Type Injection Fusion Model

The model we propose presents a novel fusion schema in a CNN-based architecture composed of two branches (cf. Figure 2(a)). The first branch learns embedding feature representations from the audio frames of one data type. The deep representations learned are then injected to all the convolutional layers of the second branch, which learns the embedding feature representations from the audio frames of the other data type, and performs the actual detection.

The fusion schema proposed is illustrated in Figure 2(b). The embedding feature vector learned from the first branch is projected to match the channel-dimension of the convolutional layer in the second branch, and added to the output of the convolution. This way, the learned features of one data type consider the information of the other data type, resulting in a more thorough multi-type information fusion.

TABLE II: Performance comparison [μ±CI in %] of the models trained on the test set, when considering the audio frames as individual samples. B and C correspond to the single-type models trained using breathing and coughing samples, respectively. B+C corresponds to the multi-type models. The performance of the multi-type models is differentiated in terms of the fusion method they use. B2C/C2B indicate the models that inject the deep representations learned from the breathing audio frames into the convolutional layers responsible for learning the deep representations of the coughing audio frames, and vice versa.

Model	ACC	UAR	UAP	UF1
B	71.1 ± 2.2	70.0 ± 3.5	74.0 ± 2.9	69.2 ± 2.7
C	74.1 ± 2.3	72.7 ± 3.2	73.6 ± 3.3	72.9 ± 2.6
B+C – Concat	74.2 ± 2.1	75.0 ± 3.3	74.7 ± 3.0	74.2 ± 2.5
B+C – Conv	74.4 ± 2.2	75.1 ± 3.2	75.6 ± 3.0	75.1 ± 2.4
B+C – B2C	$\mathbf{83.1 \pm 2.3}$	$\mathbf{83.7 \pm 3.4}$	$\mathbf{83.9 \pm 3.4}$	$\mathbf{83.8 \pm 2.7}$
B+C – C2B	81.4 ± 3.5	81.9 ± 3.3	83.1 ± 2.8	81.9 ± 2.3

IV. Experimental Results

All the models reported in this section are trained using the cross-entropy between the COVID-19 predictions and the ground truth as the loss to optimise. We use Adam as the optimiser with a learning rate of $1e^{-4}$. The models are trained using a batch size of 32 samples. To assess the overall performance of the trained models, we consider the evaluation metrics: *Accuracy* (ACC), *Unweighted Average Recall* (UAR), *Unweighted Average Precision* (UAP), and *Unweighted F1* (UF1) score. We report the evaluation metrics including their 95% *Confidence Interval* (CI) computed from the different executions. To compute the CI, we use 100x bootstraping for test (random selection with replacement), and compute each evaluation metric.

Section IV-A presents the results obtained when considering each audio frame as an independent sample. Section IV-B summarises the results obtained when considering each original audio sample as a whole. For this, all the audio frames segmented from the original audio sample are fed into the model, and the individual predictions are combined to determine the final prediction for the overall sample.

A. Performance comparison based on audio frames

As shown in Table II, multi-type models outperform the single-type ones. The two conventional fusion approaches investigated achieve a similar performance in terms of both accuracy and UAR. Nonetheless, when analysing the UAP and UF1 metrics, the model trained with channel convolution obtains a better result than the model applying direct concatenation. The most notorious results from these experiments are the performances of the multi-type injection fusion models. Their performance surpass all the previous models in terms of all the evaluation metrics considered.

Injecting cough information into the convolutional layers responsible for learning deep representations from the breathing frames achieves more discriminative representations for the detection of COVID-19, in comparison with the conventional information fusion methods. The performance of the injection fusion model is even better when the breathing information is

TABLE III: Performance comparison [$\mu \pm$ CI in %] of the models trained on the test set, when considering each audio sample as a whole. B+C indicate the performance of the multi-type models. The performance of the multi-type models is differentiated in terms of the fusion method they use. B2C/C2B indicate the models that inject the deep representations learned from the breathing audio frames into the convolutional layers responsible for learning the deep representations of the coughing audio frames, and vice versa.

	ACC	**UAR**	**UAP**	**UF1**
B	71.2 ± 2.4	72.1 ± 3.7	73.5 ± 3.6	71.6 ± 3.1
C	73.3 ± 2.3	73.8 ± 3.7	73.1 ± 3.5	72.8 ± 3.0
B+C – Concat	74.6 ± 2.4	73.8 ± 2.9	69.2 ± 3.4	70.2 ± 3.7
B+C – Conv	76.7 ± 2.2	76.9 ± 3.4	71.3 ± 3.3	72.4 ± 2.8
B+C – B2C	$\mathbf{78.4 \pm 2.4}$	$\mathbf{78.3 \pm 3.5}$	77.6 ± 3.6	$\mathbf{78.0 \pm 2.7}$
B+C – C2B	$\mathbf{78.4 \pm 2.0}$	77.6 ± 3.1	$\mathbf{79.7 \pm 2.9}$	$\mathbf{78.0 \pm 2.6}$

injected into the convolutional layers responsible for learning deep representations from the coughing frames. Thus, these results support the hypothesis that injection fusion performs a more thorough coupling between the two data types, and achieves better detection results.

B. Performance comparison based on audio samples

In this section we compare the performance of the multi-type injection fusion models when considering each audio sample as a whole. For those recordings segmented into more than one audio frame, our models produce a sequence of COVID-19 predictions, one for each individual frame. If the majority of the individual predictions is positive or negative, we consider the whole audio sample to correspond to a COVID-19 or to a healthy patient, respectively.

The performance of our proposed models using injection fusion surpass the single-type models, and the multi-type fusion models in terms of all the metrics assessed. However, we can observe that the variances among the results obtained are higher for most of the metrics. The two injection strategies achieve a similar accuracy and UF1 scores. Nonetheless, the injection of breathing information to learn embedded features from the coughing information scores a higher UAR. In terms of UAP, the reverse fusion provides better results.

The performance of our models cannot be fairly compared with the baseline performance reported by the authors of the dataset [10], as different validation strategies and data splits are chosen to assess the performance of the models trained.

V. CONCLUSIONS

In this work, we presented a novel multi-type feature fusion method in a CNN-based architecture. We validated the suggested approach for the first time in the context of COVID-19 detection by fusing breathing and coughing information from the same patient. The results obtained from the models trained indicate that the proposed method outperforms those using either single-type information or conventional approaches to fuse multi-type information. Hence, the injection fusion approach proposed seems to be effective for the task at hand.

Future works will focus on the validation of the proposed approach using other COVID-19 related datasets containing breathing and coughing sounds. The fusion mechanism selected impacts the performance of the overall model. Hence, further investigation could target at the development of new information-fusion mechanisms that can better exploit the complementarity between both data types.

REFERENCES

[1] Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat, “CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images,” *Computer Methods and Programs in Biomedicine*, vol. 196, pp. 105581, 2020.

[2] Stephanie A. Harmon, Thomas H. Sanford, Sheng Xu, and et al., “Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets,” *Nature Communications*, vol. 11, 2020, 7 pages.

[3] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy, “Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis,” in *Proceedings of Interspeech*, Shanghai, China, 2020, pp. 4811–4815, ISCA.

[4] A. Hassan, I. Shahin, and M. B. Alsabek, “COVID-19 Detection System Using Recurrent Neural Networks,” in *Proceedings of the International Conference on Communications, Computing, Cybersecurity, and Informatics*, Sharjah, United Arab Emirates, 2020, IEEE, 5 pages.

[5] M. Cohen-McFarlane, R. Goubran, and F. Knoefel, “Novel Coronavirus Cough Database: NoCoCoDa,” *IEEE Access*, vol. 8, pp. 154087–154094, 2020.

[6] Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Koike, Xiao Li, Zixing Zhang, Yoshiharu Yamamoto, and Björn W. Schuller, “An Early Study on Intelligent Analysis of Speech Under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety,” in *Proceedings of Interspeech*, Shanghai, China, 2020, pp. 4946–4950, ISCA.

[7] Björn W. Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panagiotis Tzirakis, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, Leon J. M. Rothkrantz, Joeri Zwerts, Jelle Treep, and Casper Kaandorp, “The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates,” in *Proceedings of Interspeech*, Brno, Czech Republic, 2021, ISCA, To appear.

[8] 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.

[9] Miad Faezipour and Abdelshakour Abuzneid, “Smartphone-Based Self-Testing of COVID-19 Using Breathing Sounds,” *Telemedicine and e-Health*, vol. 26, no. 10, pp. 1202–1205, 2020.

[10] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo, “Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data,” in *Proceedings of the 26^{th} International Conference on Knowledge Discovery & Data Mining*, Virtual Conference, 2020, pp. 3474–3484, ACM.

[11] Steven Davis and Paul Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proceedings of the 25^{th} International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105, Curran Associates Inc.

[13] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *Proceedings of the 32^{nd} International Conference on Machine Learning*, Lille, France, 2015, pp. 448–456, PMLR.

[14] Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee, “Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units,” in *Proceedings of the 33^{rd} International Conference on Machine Learning*, New York City, NY, USA, 2016, pp. 2217–2225, PMLR.