Pattern Recognition 127 (2022) 108656

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath

Vincenzo Dentamaro[a,*], Paolo Giglio[a], Donato Impedovo[a], Luigi Moretti[b], Giuseppe Pirlo[a]

[a] Università degli studi di Bari "Aldo Moro", Department of Computer Science, via Orabona 4, Bari, 70125, Italy
[b] Università degli studi di Bari "Aldo Moro", Medical School, Bari, Italy

ARTICLE INFO

Article history:
Received 31 March 2021
Revised 10 March 2022
Accepted 14 March 2022
Available online 15 March 2022

Keywords:
Audio classification
Spectrograms
Attention mechanism
Covid
Pre-screening
Convolutional neural network

ABSTRACT

This study presents the Auditory Cortex ResNet (AUCO ResNet), it is a biologically inspired deep neural network especially designed for sound classification and more specifically for Covid-19 recognition from audio tracks of coughs and breaths. Differently from other approaches, it can be trained end-to-end thus optimizing (with gradient descent) all the modules of the learning algorithm: mel-like filter design, feature extraction, feature selection, dimensionality reduction and prediction. This neural network includes three attention mechanisms namely the squeeze and excitation mechanism, the convolutional block attention module, and the novel sinusoidal learnable attention. The attention mechanism is able to merge relevant information from activation maps at various levels of the network. The net takes as input raw audio files and it is able to fine tune also the features extraction phase. In fact, a Mel-like filter is designed during the training, thus adapting filter banks on important frequencies. AUCO ResNet has proved to provide state of art results on many datasets. Firstly, it has been tested on many datasets containing Covid-19 cough and breath. This choice is related to the fact that that cough and breath are language independent, allowing for cross dataset tests with generalization aims. These tests demonstrate that the approach can be adopted as a low cost, fast and remote Covid-19 pre-screening tool. The net has also been tested on the famous UrbanSound 8K dataset, achieving state of the art accuracy without any data preprocessing or data augmentation technique.

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (Covid-19) is the etiologic agent of coronavirus disease 2019 (COVID-19). It has rapidly spread worldwide. On the 30th of January 2020 WHO's General-Director declared the novel coronavirus outbreak a public health emergency of international concern (PHEIC), that is WHO's highest level of alarm [1]. Considering the daily and all-embracing impact of COVID-19, it is auspicial and essential the development and implementation of screening tools, in order to guarantee a reliable, rapid, economical, scalable, and highly repeatable approach.

It is important to make a clear distinction between screening and diagnostic tools. Screening concerns a likely presence of the disease, while diagnosis unequivocally indicates the presence or absence of the disease. Currently, the most frequently used solutions to diagnose the presence of the disease are the Polymerase Chain Reaction (PCR) swab test, used to detect genetic material from a specific organism, such as a virus or fragments of it, and the serological tests and rapid antigen or antibody tests [2]. From a pattern recognition perspective, a valuable technique for the diagnosis of Covid-19 is based on Chest X-Ray or CT scans image analysis [3], in fact 99% of accuracy have been already achieved [4]. The problem is that the source X-rays and CT scans are typically performed when the patient is already in a critical phase of the disease. As a result, the person has already experienced advanced symptomatology such as severe breathing difficulties. For early Covid-19 detection, inexpensive screening tools are important to understand if there is a certain probability of having contracted the virus and, consequently, deepen the analysis with diagnostic tools such as the PCR Test. Indeed the basic idea is to inspect early symptoms. Various Chat-bots and tele health systems have been introduced. Chat-bots ask questions aimed at detecting the early symptoms most frequently related to the presence of Covid-19 and at suggesting whether to contact the appropriate medical personnel, do PCR swab testing, or not to worry. Microsoft created "Clara" in collaboration with Center for Disease Control CDC. In this direction, Apple updated "Siri" as well as Amazon updated "Alexa" chatbots. Symptoma is of particular interest as it is able to differentiate 20000 diseases including Covid-19 [5]. Various risk screening

* Corresponding author.
E-mail address: vincenzo.dentamaro@uniba.it (V. Dentamaro).

https://doi.org/10.1016/j.patcog.2022.108656
0031-3203/

platforms have been developed with the aim of screening Covid-19 presence, providing patients with a series of questions to answer, and then evaluating the outcome by classifying it as “at risk” or “without risk” [6]. However, all these approaches do not take a vital biometric signal as input.

This paper presents a novel deep neural network architecture for audio classification (based on breath and cough) which can be engineered for smartphones in a distributed scenario. The solution may be applied as a scalable, non-invasive, prompt and almost zero-cost screening tool for COVID-19 infection. According to the output provided by the system, COVID-19 potential infected subjects would be encouraged to limit their social activities until they undergo laboratory tests, such as the routinely used Covid-19 virus-specific reverse transcriptase polymerase chain reaction (RT–PCR) test [7].

Previous works have put in evidence that the analysis of breath [8,9], cough [9–12] and voice [13] can be used as a precious source of information for pattern recognition algorithms, which are able to extract a great amount of patient data (as gender and provenance) and to determinate with high accuracy if the patient suffers from some diseases and disorders. The most part of these diseases are respiratory ones (such as asthma, pneumonia, pertussis, Chronical Obstructive Pulmonary Disease or COPD, tuberculosis and, last but not least, COVID-19). In fact, COVID-19 does not lead to a “typical” Acute Respiratory Distress Syndrome. More generally, diseases that cause different pathomorphological alterations in the respiratory system are reflected by different sounds of cough, breath, and voice.

Taking into consideration the high percentage of asymptomatic yet contagious COVID-19 patients, researchers have argued that forced-cough (i.e. voluntary cough [11]) keeps the same biomarker potential of the spontaneous one, as data have wildly proved [10–12,14–19]. In fact, half of the asymptomatic cases present CT abnormalities [3]. According to a wide set of previous studies, this work inspects cough and breath, which are universal, language independent, and allow cross dataset tests. In this case, the classification problem has been considered as a binary one. Many previous works are based on features extraction (according to different techniques) and on the use of some classifiers to get the final decision [8,11,20]. The main limitation of these works is based on the impossibility to perform an end-to-end training. In this article, the term “end-to-end training” is referred as the possibility of training a complex system as a whole with gradient descent. End-to-end training is a desirable property because it allows to optimize all the modules of the learning algorithm in order to minimize an error. Thus a single neural network framework is capable of performing feature extraction, feature selection, dimensionality reduction and prediction by optimizing each single module. Other works such as [9,16,19] used end-to-end training and deep neural networks, however these works lack the optimization of the spectrogram’s filter generation as an internal process. Moreover, no attention mechanisms are used to extract relevant patterns.

Another relevant issue deals with the lack of proper cross-dataset tests. These tests are of paramount importance to stress generalization capabilities of the models while reducing biases [21]. All the reviewed works use no more than two datasets merged together but no work has tested the proposed system with the cross-dataset modality, which conversely can be referred to the most part of real cases [9,17]. In order to overcome aforementioned problems the Auditory Cortex ResNet architecture is presented in this work providing the following innovations:

1. The Auditory Cortex ResNet, briefly AUCO ResNet, is proposed and tested. It is a deep neural network architecture especially designed for audio classification trained end-to-end. It is inspired by the architectural organization of rat’s auditory cortex, containing also innovations 2 and 3. The network outperforms the state-of-the-art accuracies on a reference audio benchmark dataset without any kind of preprocessing, imbalanced data handling and, most importantly, any kind of data augmentation.
2. A trainable Mel-like spectrogram layer can finetune the Mel-like-Spectrogram to capture relevant time frequency information. This is achieved by summing trainable Mel-banks filters and then performing the dot product with the short-time Fourier transform (STFT) spectrogram.
3. A novel sinusoidal learnable attention mechanism that can be considered as a technique to weight local and global feature descriptors focusing on high frequency details. This novel attention mechanism is used to create a multi-stages concept learning, i.e. the fact that a complex concept representation is built by merging meaningful information from various intermediate levels.
4. State of the art cross-dataset testing and related accuracies. In this scenario, AUCO ResNet has been trained on the large Coswara [14] dataset and tested on the Cambridge [8] dataset.

The work is organized as follows: Section 2 sketches the state-of-the-art in Covid-19 screening techniques using audio and deep neural network architectures. Section 3 presents the Auditory Cortex ResNet model and all its internal components. Section 4 describes the various experiments on different Covid-19 screening tasks as well as the cross-dataset tests. The section also describes the tests on the UrbanSound 8K dataset. Results and their discussion are presented in Section 5. Conclusions and future work are presented in Section 6.

2. State of the art review

In general, already available solutions can be categorized as shallow learning -based, deep learning -based and hybrid ones. Currently, the most widespread are pure deep learning approaches and hybrid. Hybrid approaches are an orchestration of both deep learning and shallow learning techniques. Hybrid solutions are mainly composed of a feature engineering phase and a shallow learning classifier to provide the final decision. In turn, pure deep learning solutions can adopt non-recurring deep learning models and recurrent deep neural networks. In the first case, Convolutional Neural Network are used to exploit spatial information from spectrogram images obtained from audio samples. In the second case, recurrent neural networks exploit the sequential nature of the audio samples. The work by Brown et al. [8] provide an example of hybrid solution. In this case an Android App and a web tool have been used to collect audio of breaths and coughs from healthy and Covid-19 patients. Feature engineering was performed on data to extract a set of features (i.e. sound period, the acoustic tempo, the root mean squared energy, the roll-off energy, the spectral centroid, the Zero Crossing Rate, the Mel Frequency Cepstral Coefficients MFCCs and its derivatives). Additionally, a pre-trained deep learning VGGish model was used to extract 128 dimensional embedding features. It is important to underline that the VGGish network was pre trained on YouTube-8M sound datasets. Particularly, the feature vector created was trained using the shallow learning technique SVM (Support Vector Machine) with radial basis function kernel. It is worth noting that an inter-patient separation scheme was adopted with aim of balancing the number of Covid-19 positive and healthy controls (HC) thus avoiding bias on the learning process. The Area Under the ROC Curve (ROC) resulted in a value of around 0.8 on task 1. The solution proposed by Laguarta et al. [11] is another example of hybrid solution. Authors employed MFCCs and ResNet deep learning architectures along with four additional biomarkers related to muscle degradation, vocal cords,

sentiments and lung & respiratory tract [11]. These were used as a pre-training phase of the deep learning algorithm: each biomarker output was the input of an in-parallel trained 2D ResNet. The last densely connected layer was concatenated with the others to obtain the final classification. In this case, an accuracy of 98.5% was reported on the task of detecting Covid-19 from cough audios, representing the highest accuracy ever obtained on a dataset designed with the inter-patient separation scheme [8]. A similar approach is proposed by Alsabek et al. [20], even if a limited dataset of only 80 people (60 HC and 20 positive to Covid-19) was employed. The dataset includes audio of coughs, breaths and voice recordings. A Long-Short term Memory (LSTM) Recurrent Neural Network (RNN) was adopted upon the following features: Spectral Centroid, Spectral Roll-Off, Zero-Crossing Rate, MFCCs and its derivatives. In this case no inter-patient separation scheme was designed. The audio soundtracks were merged and then the 70/30 train-test ratio was adopted. The AUC achieved around 97.4%, 98.8% and 84.4% when classifying coughs, breaths and voices respectively.

Sharma et al. [14] introduced in 2020 an interesting and publicly available dataset named Coswara which includes a huge number of coughs, breaths and various phonemes along with the respective metadata. The authors also presented a CNN architecture used for training and inference on the MFCCs features as in [9] reporting a F1-score on Covid-19 detection of 0.8952. In this case no inter-patient separation scheme was adopted, however in our work (this article) the same architecture has been reproduced in order to perform a comparison with the adoption of the interpatient separation scheme. Another hybrid solution tested on the Coswara dataset takes as input the Mel-frequency Cepstral Coefficients for a modified VGG16 architecture [16]. Authors report F1-score of 0.6959 without inter-patient separation scheme. Similarly to these works, it was implemented also the ensemble of methods in [19], where a COVID-19 Identification ResNet (CIdeR) is used (based on ResNets, a variant of the CNN which uses residual blocks) over a decibel representation (the log transformation) of a frequency-time domain spectrogram, authors used the inter-patient separation scheme with a AUC of 0.827 on Covid-19 detection. The previously mentioned datasets in [8] and [14] were used to train a specific system of detection that is then tested on nearly 390 patients from Asia and Latin America [17]. The authors employed an ensemble of 3 different networks: a 1D network, a 2D network and a LSTM network. The first one uses the MFCCs as extracted features. The second one resizes (64×64) the Mel-spectrogram images. The third uses two extra features related to the state of fever/myalgia and the respiratory conditions. The overall AUC is reported to be 0.77. Results already reported have been confirmed by a similar [15]. On the other hand, standard MFCC features and a wide set of classifiers, along with multiple experimental settings have been tested, reporting SVM as the best performing classifier (accuracy of 70%) in the binary classification task [22]. Many works have adopted data augmentation to overcome the problem of data scarcity [23,24] reaching a mean accuracy of 73% [23,24] and of 75% [25] depending on the classification model.

Concerning pure non-recurrent deep learning solutions, a 1D CNN especially designed for audio classification on low end devices has been proposed [26]. An interesting element of this work is the data processing procedure based on the idea to weight samples of the audio tracks. The resulting accuracy is 70.5% on UrbanSound8k. A dilated CNN-based Environmental Sound Classification Tasks D-CNN-ESC system and LeakyReLU activation function were adopted with the aim to evaluate ESC particularly on the UrbanSound8k dataset [27]. Authors of the study presented in [28] aim to classify inter-floor noise according to noise sources by using a convolutional neural network model, in [29] a similar task is performed on recognizing audio events in urban environments adopting a deep learning architectural scheme (Convolutional Neural Networks, CNNs), which has been trained to distinguish between different audio context classes.

Concerning pure non-recurrent deep neural networks with spectrogram generation, the work in [16] generated a Mel-spectrogram from the audio and a small Convolutional Neural Network (CNN) was used for classification aims achieving an accuracy of 70.58%. Typical Covid-19 symptoms such has wet and dry coughs, croup, pertussis and bronchitis coughs have been considered when acquiring undetermined coughs. A similar approach has been adopted on the Pfizer dataset of “Sick Sound” [30], in this case audio signals were converted into images of Spectrograms by using a Short Time Fourier Transform and the resulting images were fed into a Xception deep neural network [31]. A final accuracy of 75% was achieved with no inter-patient separation scheme: the result is consistent with those already found in literature (e.g. [16]).

The work proposed in [10] also belongs to the category of pure non-recurrent deep learning solutions with spectrogram generation is: a pre-training phase of a ResNet-18 CNN architecture is performed on Log-Mel spectrogram images computed on the audio tracks of the Free Sound database and of the Flusense database [32]. The pre-trained network is augmented by adding an adaptive average pooling layer in both the time and frequency domains. The output is then passed through 3 densely connected layers with the last one layer having 2 neurons and a softmax activation function. In this case the inter-patient separation scheme was used by keeping balanced the proportions of Covid-19 positive versus HC on both training and test sets. Multiple tests were performed with and without label smoothing techniques: the AUC values were respectively 0.68 and 0.65. The dataset used by the study in [10] consisted of 3,118 cough sounds from 1,039 subjects (376 of which positive to Covid-19).

Transformers [33] with the Attention mechanism were used by Pinkas et al. in [34], where Transformer Embeddings were injected into a GRU-RNN neural network to perform classification. In this case the F1-score in cross-validation modality was 0.74. Another important aspect in the work by Pinkas [34] is the introduction of a dataset which included recordings of phonemes /ah /e /z, coughing and counting from 50 to 80. The dataset includes 29 Covid-19 positive patients and 59 HC. Authors also report that the /z phoneme can provide higher performance then those obtained with the cough signal. This result is of particular interest; however, it cannot be generalized given the reduced and unbalanced dataset.

Many of the previously mentioned works share a ground assumption related to the excellent capability of CNNs to perform classification tasks over images (e.g. spectrograms). Similar assumptions are at the basis of the use of recurrent neural networks over time series (e.g. raw audio). Moreover, based on the reviewed works, different authors have tested their approaches on different datasets with different testing modalities, so it is difficult to derive some consistent conclusion from them. Sometimes the inter-patient scheme wasn’t adopted, thus using different audio of the same user in training and testing. Moreover, due to the widespread diffusion of the Covid-19 pandemic, approaches should be benchmarked on a wide set of datasets. This work overcame these limitations by proposing a new approach and benchmarking it against the most performing techniques (re-implemented) on the two most used big datasets on which a quality check has been performed by the original authors [8,14], thus providing inter country validation.

3. The Auditory Cortex ResNet Deep Neural Network model

Auditory Cortex ResNet, briefly AUCO ResNet or AUCO, here proposed, is a nature/biologically inspired deep neural network that takes as input a raw audio and outputs the respective class. This

Table 1
The structure of the six layers in the AC. The V layer is the thickest and is of great interest for its projections to telencephalic and mesencephalic targets.

Layer	Histology [37]	Specific layer thickness [37]	% of the total thickness (1100 μm) [37]
I	Few neurons	140 μm	13%
II	Many small and densely packed polymorphic cells	125 μm	11%
III	Less packed pyramidal and non-pyramidal large neurons, organized in small columnar chains	190 μm	17%
IV	Small stellate cells, spherical or oblate	105 μm	10%
V	Low density of large pyramidal neurons with large intercellular space, named corticocollicular neurons, and commissural cells that are smaller and more heterogeneous	270 μm	26%
VI	Closely packed and flattened pyramidal and non-pyramidal cells, oriented parallel to the underling white matter	245 μm	22%

is achieved without preprocessing, data augmentation or manual spectrogram generation. The model includes elements also present in the biological auditory cortex of mammals (rats). The main points are briefly reassumed as follows:

- It is composed by six main blocks.
- It can evolve sound perception depending on the classification problem.
- It has many attentions levels as a combination of features learned from the main blocks.
- The number of neurons within each stage has similar proportions of neurons found in rats and similar functionalities with respect to the type of neurons and their connections.

The result is a deep neural network that learns features in the frequency domain. It generates learned finetuned spectrograms without loss of information due to compression, thus preserving the complex components of the short time Fourier transform. These spectrograms are then fed at lower levels which make use of architectures that simulate non-pyramidal (basket) and pyramidal cells found in mice auditory cortex. Different levels of attention mechanisms are used to model the different type of auditory attentions of mammals [35].

3.1. Brief description of biological rat auditory cortex

Auditory processing takes place in the auditory cortex (AC). Even if each mammalian species has its own frequency range sensibility, there are some common characteristics, such as the presence of two fields (primary auditory and anterior auditory field) with a regular tonotopy. The general histological structure of the cortical depth, that consists of six layers parallel to the pial surface (Table 1), is mostly the same in the primary auditory, visual and somatosensory cortices of the mice. [36]

It is important to specify that two kinds of AC neurons are briefly described here: Pyramidal neuronal cells (PCs) and Basket-like cell (BCs). PCs are found in different areas of the brain; they are basic excitatory-type elements of the mammalian nervous system. In the AC, PCs allow to perform pitch recognition and sound attention [38]. Instead, BCs are interneurons that are common in layers II-IV of rat somatosensory cortex.

From a functional point of view, the cerebral cortex presents a vertical columns organization in which the neurons have similar functions. Even in this case it seems that the AC has peculiar features compared to the other neocortex areas. A study by Tischbirek C. H. et al. aiming to map the activity of mouse primary auditory cortex neurons, identified "functional microcolumns bringing together large-scale tonotopy and locally heterogeneous frequency responses throughout all AC layers." The study concludes that this spatial organization, linked to its response patterns, "may reflect learning-dependent wiring rules in the AC" [39].

These evidences are at the base of choices for the architecture as described in the following.

3.2. The Auditory Cortex ResNet model

Similarly to learnable sound perception of mammals, Mel-like spectrograms are learned and Convolutional Neural Networks (CNNs) are used to learn discriminant features. Both Mel-like spectrograms and CNN kernels are learned using the backpropagation algorithm, which regulates the change in weights of both Mel-like filterbanks and CNNs kernels with respect to the target function. This is similar to the response-based learning process typical of mammal's brain [40]. In fact, the capability of CNNs to perform hierarchical feature extraction from raw signals emulates the multilayer learning process present in the Neocortex of mammals.

From a pattern recognition perspective, the AUCO ResNet is a deep neural network built on top of SE-ResNet [41] architecture with ResNet bottleneck layers [42]. Bottleneck layers are used in place of basic ResNet blocks for allowing deeper networks while saving computational time. They use a 1×1 convolution for reducing channels of the input followed by a 3×3 convolution and finally another 1×1 convolution for repristinating the number of channels. The ResNet [42] architecture was selected as backbone due to its residuals generalization. Its generalization bound is equivalent to other state of the art neural networks such as ResNeXT and DenseNet [43].

In AUCO ResNet three types of attention mechanisms are used for propagating relevant channels and time wise frequency information: the convolutional block attention module (cbam-block) [44], squeeze and excitation network (se-block attention) [41] and sinusoidal learnable attention. The last mechanism is a combination of learnable attention from [45] and a dense layer with a sinusoidal activation function properly initialized as shown in [46]. Fig. 1 shows the legenda of symbols used for the description of each piece of the network.

Hence, AUCO ResNet is composed by six macro layers:

1. The Spectrogram convolutional layer with channel and spatial attention. This layer can be compared to the first layer of the mammal's auditory cortex which performs a complex (spectrogram like) representation of the audio [35].
2. Plain set of two 1×1 2D convolutional layers (VGG-ish style layers) ending with a Squeeze and Excitation attentional layer. This layer contains a small number of neurons and plain convolutions (without residual) and can be compared to the second layer of the rat auditory cortex (which contains a small number of basket cells neurons [37]).

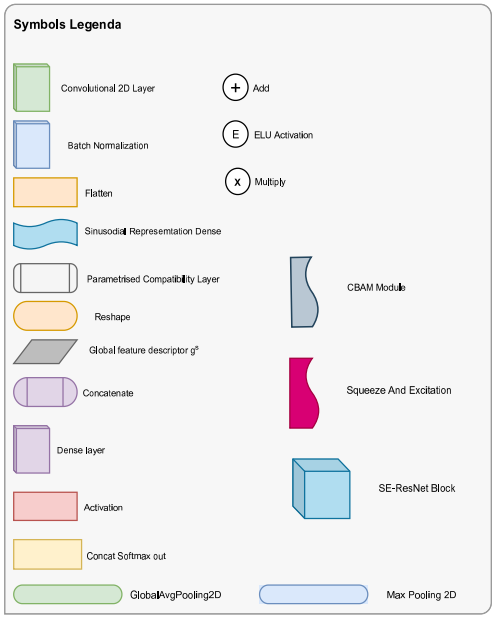

Fig. 1. Legenda of symbols used in the architectural description.

3. The first SE-ResNet block (depicted in Fig. 2) ending with a Max Pooling layer, this architecture would resemble the pyramidal neuronal cells in the third layer of rat's auditory cortex.
4. The second SE-ResNet block ending with a Max Pooling layer whose number of filters is doubled with respect to the first SE-ResNet block.
5. The third SE-ResNet block ending with a Max Pooling layer whose number of filters is doubled with respect to the second SE-ResNet block.
6. The fourth SE-ResNet block ending with a Max Pooling whose number of filters is equal with respect to the third SE-ResNet block. It is followed by a convolutional layer with 64 filters, a batch normalization layer, a max pooling layer and a CBAM block. The output of the CBAM is the general descriptor of the Sinusoidal learnable attention layer. At the end of this macro layer, the classification is carried out by concatenating the learned representation performed by this sinusoidal learnable attention layer, squashing them into a dense layer which uses a "softmax" activation function to perform the final classification. The sixth macro layer acts like a macro pyramidal neuronal cell whose dendrites capture important information from middle to lower levels and its axon ends with a concatenation of the learned attentions as feature embeddings to be fed to the final dense classification layer.

The overall architecture is depicted in Fig 3. The first and the second macro layers are never repeated. The third macro layer is repeated 8 times, the fourth macro layer is repeated 3 times, the fifth macro layer is repeated 7 times and the last macro layer is repeated 2 times plus the final sinusoidal learnable attention mechanism. This organization mimics the equilibrium based on the number of neurons and the physical thickness of each layer of the rat primary auditory cortex (par. 3.1) in Table 1.

Although the design of the AUCO ResNet was guided by the organization of the biological auditory cortex of rats as explained in Section 3.1, it is also interesting to report that many different configurations have been tested by modifying the number of layers in each macroblock, as well as the positions where to cut the network to feed the sinusoidal learnable attention layer. The reported AUCO is the most performing one.

3.3. Trained Mel-like Spectrogram layer

The discrete Fourier transform is one of the classic algorithms that allows conversion from the time domain to the frequency domain. The Short-time Fourier Transform (STFT), refers to an application of the DFT in which the signal is cut into short windows within which the signal can be considered quasi-stationary so that the transform can be computed. The STFT function is composed by two variables, ω which represents a frequency and τ, contains information about the window size. Eq. (1) represents the STFT in continuous domain.

$$\hat{X}(\tau,\omega) = \int^{x}(t)w(t-\tau)e^{-i\omega t}dt \tag{1}$$

The spectrogram of an audio sample is the module of the STFT as shown in Eq. (2):

$$\text{spectr}(x) = |X(\tau,\omega)|^2 \tag{2}$$

Then the Mel filter bank is typically applied having the aim of simulating the non-linear human perception of sounds [15]. This filter provides a better resolution at low frequencies and a smaller one at high frequencies in a logarithmic scale. There are several types of filters, among them the triangular filters are widely used. They capture the energy at each frequency band and approximate the spectrum shape. In this work the original Mel filter-bank's amplitude is evolved to maximize the accuracy on each task. More specifically:

- Let c be the number of classes in of the classification task, in the specific case of Covid-19 recognition 2 classes are considered (positive and negative);
- Let m be the triangular Mel-filterbank;
- Let v be a variable (to be optimized by the backpropagation step) initialized with the transposed of m: m^{T}.

The trainable Mel-like filterbank tm is defined in Eq. (3):

$$tm = \sum_{1}^{c} v_c \tag{3}$$

The sum is performed after that the backpropagation operation changed the weights of each trainable Mel-like filterbank. In this way the designed filter has a set of bands focused on frequencies that are discriminative of one class with respect to the other. The final sum creates a filter for the frequencies information that are not much relevant, by merging only bank's amplitudes that are necessary to preserve important frequency information for that specific classification task. In practice, it leverages the property of end-to-end training of optimizing each single part of the network toward the goal. In this specific case, the intuition is that, by performing the sum of trainable banks before the dot product, the backpropagation algorithm (thus, the gradient descent optimization) optimizes each single v_c so that its sum would merge only filter banks relevant to the task at hand. Thus, the filtered spectrogram is defined as the dot product among the STFT spectrogram and the tm filterbank as in Eq. (4).

$$melspectrogram(x) = spectr(x), tm^{T} \tag{4}$$

The filtered spectrogram is then converted to decibel and returned as a layer of the AUCO ResNet. Images of the trainable filtered spectrograms and standard Mel spectrograms are presented in Appendix A showing partially visible differences able to play an important role for the final classification accuracy, but also for structuring the evolution of the sound perception.

3.4. The Convolutional Block Attention Module (CBAM)

CBAM is composed by a channel attention module and a spatial attention module as depicted in Fig. 4.

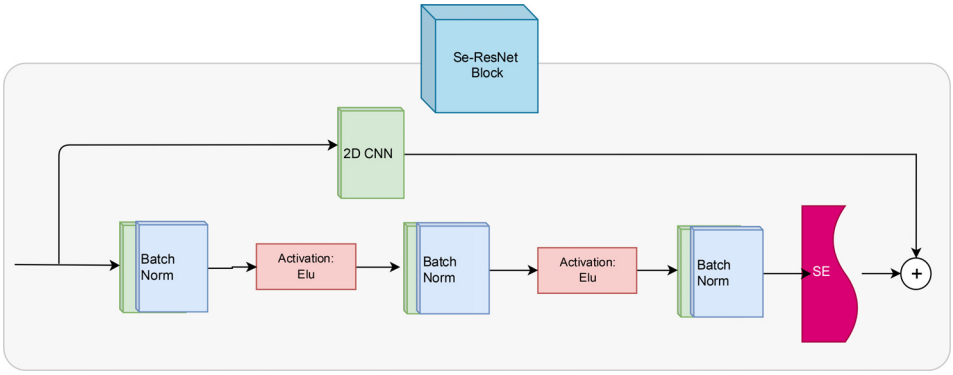

Fig. 2. Se-ResNet block.

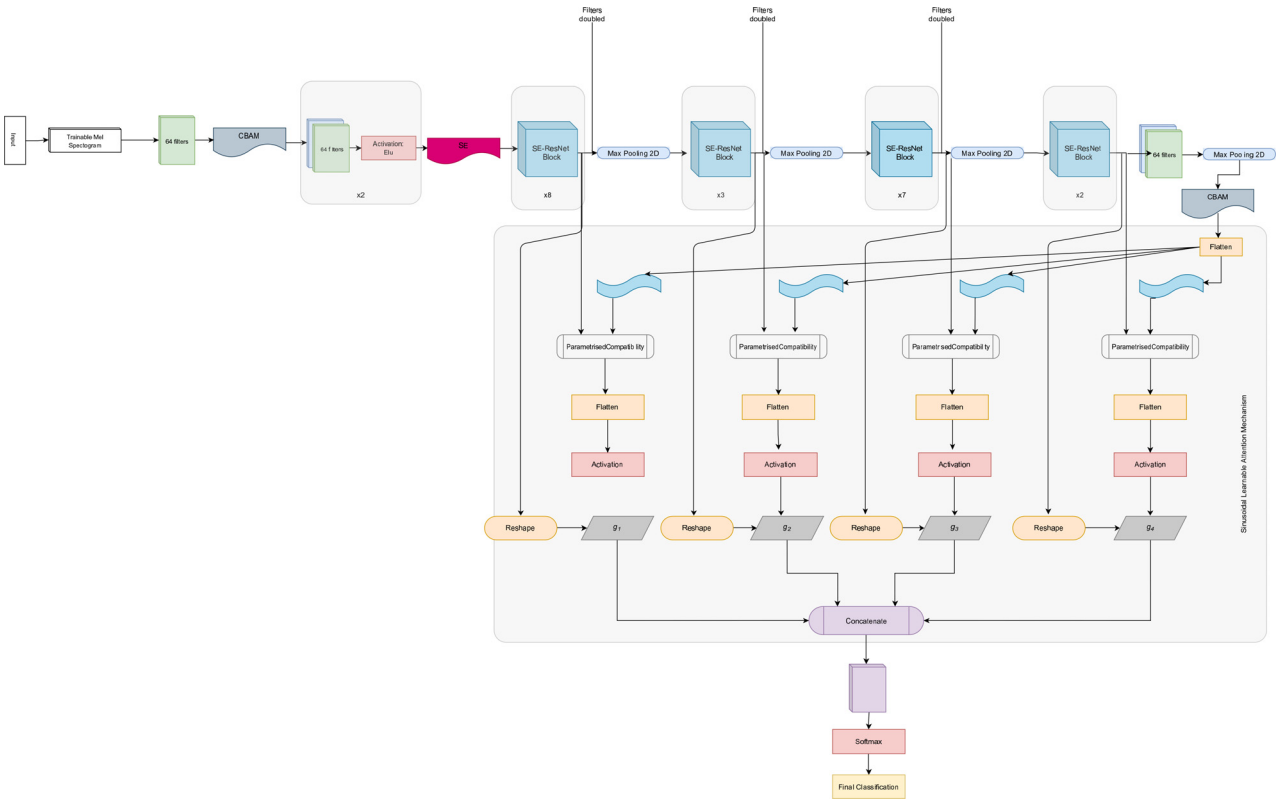

Fig. 3. The Auditory Cortex ResNet architecture.

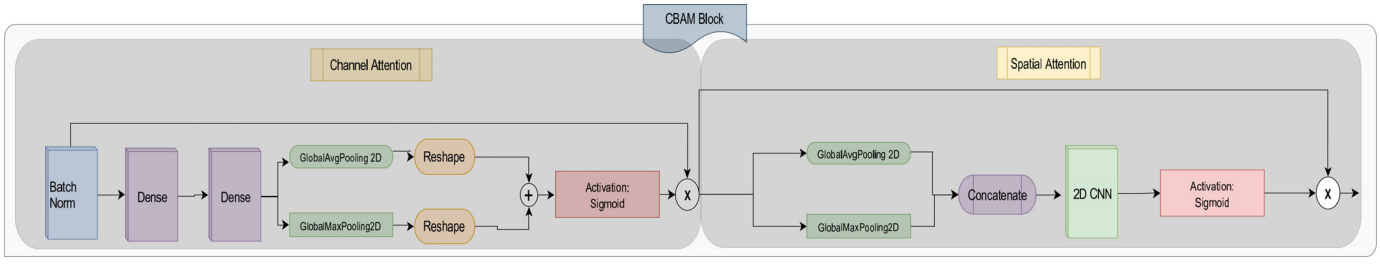

Fig. 4. The CBAM block attention module.

In this specific case, the CBAM generates attention maps by firstly performing the channel attention. This attention is a kind of kernel-wise attention. It uses one dense layer with ELU activation function followed by a linear dense layer. Usually, the first dense layer has a number of neurons smaller than the number of kernels, in the case of AUCO the ratio is $\frac{1}{2}$. Successively, its output is fed into two pooling layers, a max and an average pooling layer; results are then summed and fed to a sigmoid activation layer. The Hadamard product among the output of this sigmoid layer and the original input creates this type of attention: it can be seen as a learnable weighting factor for the original input.

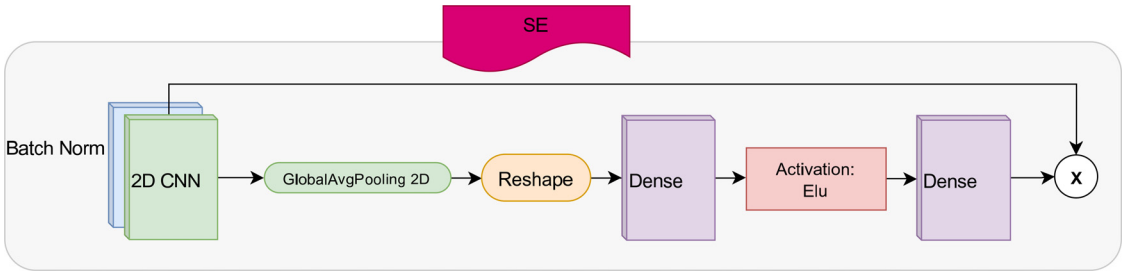

Fig. 5. The Squeeze and Excitation block diagram.

In a second stage, the CBAM block performs spatial attention, which in this specific case is referred to the temporal and the frequency domain. It concatenates the output of both an average pooling layer and a max pooling layer built on the same input and feeds this output to a 1×1 2D convolutional layer with sigmoidal activation function. The Hadamard product of the output of this convolution with the original input generates spatial attention. The authors in [44] showed that the exploitation of the different pooling layers is twofold: first they allow to weight discriminant kernel maps and increase the representational power of previously highly weighted feature-maps. The weights are determined through the sigmoidal activation function.

3.5. The Squeeze and Excitation Network Attention Module (SE-Block)

The SE-Block was originally proposed in [47] as a plug and play block (in Fig. 5), to be added before the CNN block. Here it is used for automatically selecting "important" feature-maps (channel wise).

The importance weight increases for feature maps that contribute positively to the classification task and decreases (tends to zero) for feature maps that do not contribute. It is composed by two operations, namely squeeze and excitation. The squeeze operation is composed by a global 2D average pooling layer that generates feature maps descriptors channel wise. This average pooling operation is meant to squash the dimensions of the tensor to Channel Size 1×1 dimensions. It is important to get global descriptors of each channel. The excitation block starts with a dense layer used for weighting each channel. This dense layer has a number of neurons inferior to the input feature map size, this number is an hyperparameter of the model. In this work the ratio is $\frac{1}{8}$ and the ELU (Exponential Linear Unit) activation function is used in place of the ReLU (Rectified Linear Unit) in its dense layer. The next layer is another dense layer with the number of neurons equal to the feature map size (in order to return to the original channel dimension) which concludes the bottleneck operation and uses a sigmoidal activation function. The sigmoidal activation function is more appropriate to weight more channels, as it returns continuous differentiable values between 0 and 1 with respect to other functions such as softmax, which would set strong weights to few channels only. The Hadamard product among the repristinated shape of this dense layer output and the original feature map is computed and returned as weighted feature maps.

3.6. Sinusoidal Learnable Attention

Learnable attentions can be defined as a set of techniques that allow the continuous learning of increased weights for some relevant features with respect to irrelevant ones for a specific task. This is of paramount importance because the weights of those features are learnt over time. The intuition is that the gradient descend algorithm incorporated within the backpropagation procedure would selectively increase only the weights that decrease the error. In addition to the previously explained attentions, AUCO ResNet uses also a novel type of attention inspired by the work of authors in [45] which uses sinusoidal [46] layers for carrying out the weighting among local and global feature descriptors. It is expected that weighting local feature descriptors by superimposing a nonlinear sinusoidal representation of the global descriptor would increase the capability of the network to filter concepts at different stages. This various stages concept learning, i.e. the fact that a complex concept representation is built by merging meaningful information from various intermediate levels, is present in neuroscience community with the term "grandmother neuron". The "grandmother neuron" is a theory where a hypothetical pyramidal-like neuronal cells code for high level concepts and is able to represent objects (e.g. faces). [48]

Since global feature maps live in the time frequency domain, it was shown that sinusoidal representations and, more in general, differentiable periodic functions, are capable of capturing fine grained, high frequency information especially when spectrograms and Fourier transforms operations are involved [46]. Indeed, authors in [46] defined the problem of signal restoration using a sinusoidal periodic activation function properly initialized. In the specific case of AUCO ResNet, the sinusoidal representational dense layer learns a function that maps frequency points inside a feature map to temporal location. The work in [46] adopted the sine function as the preferred non-linearity and reported that ReLU based architectures for continuous signals (such as sound, video and highly detailed images, but also chaotical systems) lack the capacity to represent with enough power the underlying signal: i.e. ReLU do not effectively represent the derivatives of a target signal.

The proposed learnable sinusoidal attention mechanism is composed by a compatibility function [45] that returns a compatibility score between local feature maps extracted at several intermediate layers. In the case of AUCO ResNet, these local feature maps are located at the end of each SE-ResNet *macro* layer (layers 3,4,5 and 6 of Fig. 3), plus the global feature map. This compatibility score is a learned weight matrix that weights more the patches whose local features are part of the dominant class.

Formally, let $L^s = \{l_1^s, l_2^s, \ldots, l_n^s\}$ be the set of feature vector for a given CNN layer, being l_i^s the i-th feature vector of the n dimensional spatial location of feature vector L^s and with $s \in \{1, 2, \ldots, S\}$ where S is the max number of intermediate convolutional layers used to merge high level concepts, in this case 4.

Let us consider $\Phi(g)$, where g is the global feature descriptor extracted from the last CBAM block just after the last max pooling layer and fed into the sinusoidal dense representation Φ as shown in Fig. 3.

The compatibility function can be defined as:

$$C = u, l_i^s + \Phi(g_i), i = 1, 2, \ldots, n \tag{5}$$

In Eq. (5) the first operation is to add the local features $l_i^s$with the sinusoidal global feature $\Phi(g)$. Later the dot product (the angular parentheses) with a learned vector u is performed for inferring an initial attention mechanism which intuitively can be interpreted

as a learning weighting factor for common patterns among instances of the same class. This attention is derived from the alignment model of attentions used in [49] which led to transformers in natural language processing tasks [28].

The set of compatibility scores $C = u, l_i^s + \Phi(g_i) = \{c_1^s, c_1^s, \ldots, c_n^s\}$ is then normalized with a softmax activation function. Let define $A^s = \{a_1^s, a_2^s, \ldots, a_n^s\}$ as the set of softmax normalized compatibility scores as shown in Eq. (6):

$$a_i^s = \frac{e^{c_i^s}}{\sum_j^n e^{c_j^s}} \tag{6}$$

The global feature descriptor vector g^s is the final attention mechanism for a particular layer s that takes the weighted combination of the original local features l^s. This is computed as in Eq. (7).

$$g^s = \sum_{i=1}^{n} a_i^s \cdot l_i^s \tag{7}$$

All the g^s feature descriptor tensors are concatenated and fed to a fully connected layer ending with a softmax activation function for carrying out the final classification.

In order to make up the learnable sinusoidal attention, inside the AUCO ResNet the following inputs are used:

- l_1^s is the layer prior to the Max Pooling layer in the first SE-ResNet block
- l_2^s is the layer prior to the Max Pooling layer in the second SE-ResNet block
- l_3^s is the layer prior to the Max Pooling layer in the third SE-ResNet block
- l_4^s is the layer prior to the Max Pooling layer in the fourth SE-ResNet block
- $\Phi_1(g)$ is the flatten global feature map after the last CBAM block fed inside a sinusoidal dense representation layer with 64 neurons.
- $\Phi_2(g)$ is the flatten global feature map after the last CBAM block fed inside a sinusoidal dense representation layer with 128 neurons.
- $\Phi_2(g)$ is the flatten global feature map after the last CBAM block fed inside a sinusoidal dense representation layer with 256 neurons.
- $\Phi_2(g)$ is the flatten global feature map after the last CBAM block fed inside a sinusoidal dense representation layer with 512 neurons.

The complete architecture in Fig. 3 shows the use of each attention mechanism and its integration within the overall network, concluding the description of the Auditory Cortex ResNet architecture.

4. Experimental Setup

In order to extract the relevant patterns that are useful to discern Covid-19 positive subjects from healthy ones, it is necessary to apply an *inter*-patient separation scheme (i.e. different people are used for training and test). This separation scheme is more suitable for medical purposes because with an intra-patient separation scheme the independently and identically distributed (i.i.d.) assumption between instances would not be achievable. This separation scheme with a balanced number of samples for each class is used [8], therefore the training set has been balanced with respect to the number of positive and negative patients, and the test set contains a balanced number of different subject's audio of both classes. Metrics adopted for evaluations are the average accuracies and the area under the receiver operator characteristics (AUC).

4.1. Cambridge Tasks

Preliminary tests have been performed on the Cambridge dataset [8]. Data were collected via an Android app and a web app. This dataset contains audio from 7000 users, 235 of which are Covid-19 positive.

Many people did not report their location, but given the available data, source countries are Greece, United Kingdom, Italy, Germany, Spain, France, Iran, United States, Bangladesh, India and France. Age ranged from 0 to 79 with the majority of people being 20-50 years old. The average age of positives and negatives was not specified. The 68% of subjects were male, 31% female, 0.7% preferred to not answer and 0.3% answered "other". Authors found as common symptoms a wet and dry cough, as well as a lack of sense of smell and a chest tightness. Authors also discarded all audio recordings acquired by people who reported to be healthy and having some symptoms. Three tasks can be performed on this dataset:

1. To distinguish Covid-19 PCR-Test positive users from HC, without symptoms, and with a clean medical history.
2. To distinguish Covid-19 PCR-Test positive users reporting Cough as a symptom from HC with a clean medical history.
3. To distinguish users using breath sounds only who resulted positive on Covid-19 PCR-Test and reported cough as a symptom, from HC with asthma and cough symptoms.

The first task is performed on cough and breaths of 62 Covid-19 positive users and 220 HCs without symptoms.

For the second task cough sounds of 23 Covid-19 positives subjects and 29 HCs having cough as symptom are available.

For the third task, the positive class is composed by breath sound of 23 Covid-19 positive subjects with cough and the negative class by breath sounds of 18 healthy subjects with cough that also reported to have asthma.

All audio files were sampled at 22050 Hz mono and only 10 secs of every audio were utilized. Audio longer than 10 secs were truncated, those shorter than 10 secs were zero padded. In an initial phase it was investigated if audio segmentation (e.g. segment audio in coughs and breaths) was of any help. Preliminary results suggested that results were better without performing any kind of data segmentation, in line with the state of the art [8].

In order to get a balanced number of subjects in both covid positive and negative classes in both train and test sets, it has been decided to perform random under sampling on a per subjects' basis as in [8]. In order to get reliable results, this random sampling has been performed 10 times in a 10-fold cross validation fashion as shown in Fig. 6. In average for Task 1, 448 Covid-19 positive samples and 508 healthy samples were used for training as well as 44 Covid-19 positive samples and 56 healthy samples have been used for testing. For Task 2, in average, 266 Covid-19 positive samples and 92 healthy samples were used for training as well as 84 Covid-19 positive samples and 14 healthy samples have been used for testing. For Task 3, in average, 236 Covid-19 positive samples and 50 healthy samples were used for training as well as 56 Covid-19 positive samples and 34 healthy samples have been used for testing. It is important to state that for both Cambridge [8] and Coswara [14] datasets, as shown in Cambridge tasks 2 and 3, Covid positive samples are more numerous because covid sufferers created, on average, more audio samples with respect to control subjects. In contrast, healthy people created fewer samples. The balance was on the number of people and not on samples as in [8].

4.2. Cross Datasets Tests

Because of the limited amount of people in the various Cambridge tasks, it has been decided to perform cross dataset testing

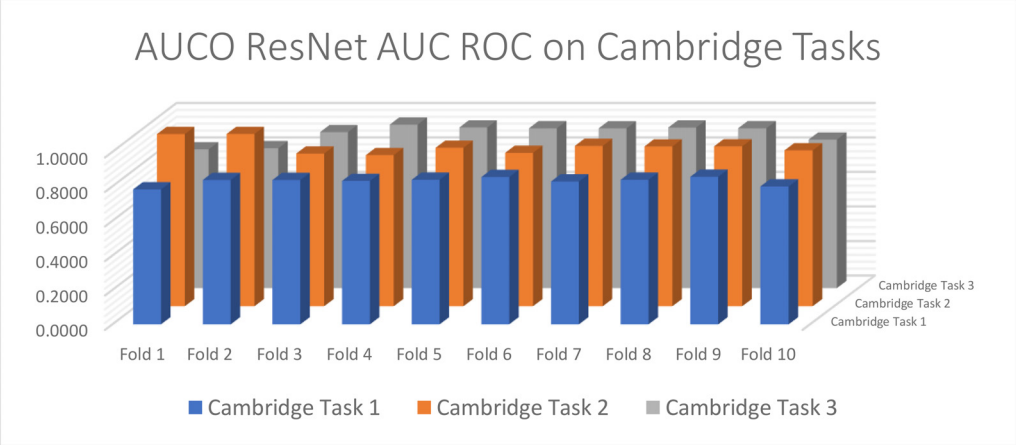

Fig. 6. Area under the ROC curve obtained by training AUCO ResNet for each fold of the various Cambridge Tasks as defined in [8].

and, thus, to ensure the generalization power of the proposed approach. At time of writing, this is the biggest cross dataset test on Covid-19 cough sounds.

The Coswara dataset [14] includes audio recordings of 1167 HCs and 100 Covid-19 positive users. Each participant provided 9 audio files. The 90% of subjects are from India mainland and 10% outside India. The majority of people had an age between 20 and 40. The 87% were male and 13% female. The following tests have been performed:

1. *The cough test*: in addition to the cough samples within the Cambridge dataset in [15] also the Coswara dataset was used. Coswara dataset [14] has been used for training while the Cambridge dataset [8], containing only controlled cough recordings (practically the data used in Cambridge Task 2), was used for testing. In order to keep balanced the number of subjects in both the training and testing set, a class-balanced random number of subjects were selected from Coswara dataset and used for training. A class-balanced random number of subjects were selected from Cambridge dataset for testing. For the validation phase the 25% of the training set was used. The reported accuracy metrics are per patient on the test set. All audio files were sampled at 22050 Hz mono and 10 seconds length, cut if sound was longer, zero padded otherwise. For this task, 292 Covid-19 positive and 284 healthy audio samples were used for training and 434 Covid-19 positive and 108 healthy audio samples were used for testing.
2. *The breath test*: the Coswara dataset was used for training by randomly selecting a class-balance amount of subjects. In this case only breaths audio recordings have been considered. The validation size is set to 25% of training data. For test the Cambridge dataset containing only breaths sound was used. Even for the testing set, class-balance subjects were randomly chosen. Again, all audio files were sampled at 22050 Hz mono and 10 seconds length, cut if sound was longer, zero padded otherwise. For this task, 356 Covid-19 positive and 352 healthy audio samples were used for training and 1358 Covid-19 positive and 282 healthy audio samples were used for testing.

The Cambridge dataset has been used in the test phase as it contains less data then Coswara. On the other hand the latter has been used for training. Moreover, in order to cope with the limited amount of data available for training, transfer learning has been adopted. Transfer learning refers to a technique that makes use of a pre-trained neural network (a neural network whose weights are trained on a bigger dataset). Then this pretrained network is used for fine-tuning (refining weights) of only some top layers, by training them on the new smaller dataset while keeping other bottom layers weights frozen. The former layers learn high level representation of the underlying sound patterns, while top layers need to be specialized on the new smaller dataset. In this specific case, weights of the AUCO ResNet trained on UrbanSound 8K were used to transfer knowledge. UrbanSound 8K was used as suggested in [22]. No data augmentation was performed, and accuracies were averaged over the 10 folds. AUCO ResNet was trained end-to-end.

The AUCO ResNet has a total number of 400 layers, the 250 bottom layers were frozen (apart from the trainable mel-like spectrogram layer) and the remaining 150 top layers were allowed to be trained. These top 150 layers were initialized with the trained weights instead of random weights initialization.

4.3. The Auditory Cortex ResNet setup

The same AUCO ResNet architecture has been used for all the tests reported here. The adapted spectrogram was generated with 2048 bins of the FFT, 150 trainable filters, an overlapping window length of 140ms and a hop size of 344. The first convolution having as input the mel-like spectrogram, performed subsampling (5×5 for avoiding managing big tensors) with 3×3 strides. RMSProp [50] optimizer was used for training the neural network, while the default activation function for all non-sinusoidal layers (apart from the softmax and sigmoidal layers) was the exponential linear unit (elu). For all tests, no preprocessing was performed in order to make results highly reproducible and reliable. The output size of the generated Mel-spectrogram layer is $962 \times 150 \times 1$.

In order to compare the results with those from other works, the following deep neural network architectures have been re-implemented and tested: ResNet 50 [42], DenseNet 201 [51] and Inception ResNet V2 [52]. In order to allow these networks to work on sounds, mel spectrograms images have been generated using the same hyperparameters previously described for the AUCO ResNet setup. All networks were trained for 150 epochs. Moreover, 4 different works (Brown et al. [8], Imran et al. [9], Bansal et al. [16] and Coppock et al. [19]) have been re-implemented and tested in the exact same conditions.

4.4. Shallow Learning setup

In order to compare the approach proposed here with shallow learning techniques (i.e. everything that is not deep learning), it was decided to use the following set of features already adopted in a set of related works [8,16,20]

Table 2
Models' comparison on Cambridge Task 1.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
AUCO ResNet	**0.8039**	**0.8028**	**0.8039**	**0.7998**	**0.8308**
DenseNet 201	0.7703	0.7872	0.7703	0.7315	0.6515
Inception ResNet V2	0.6672	0.6797	0.6672	0.6605	0.7106
ResNet 50	0.6969	0.6639	0.6969	0.6514	0.6571
Shallow SVM	0.52	0.51	0.51	0.51	0.5130
Shallow Random Forest	0.48	0.48	0.48	0.47	0.4785
Shallow KNN	0.47	0.47	0.47	0.47	0.4696
Work in Brown et al. [8]	NA	0.72	0.69	NA	0.80
Work in Bansal et al. [16] (re-imp)	0.7212	0.7124	0.7212	0.7074	0.6846
Work in Imran et al. [9] (re-imp)	0.7088	0.7135	0.7088	0.6797	0.7834
Work in Coppock et al. [19] (re-imp)	0.5946	0.5387	0.5946	0.5353	0.5645

Table 3
Models' comparison on Cambridge Task 2.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
AUCO ResNet	**0.9256**	**0.8881**	**0.9256**	**0.8980**	**0.9257**
DenseNet 201	0.8801	0.8508	0.8801	0.8498	0.6150
Inception ResNet V2	0.8916	0.8632	0.8916	0.8690	0.6786
ResNet 50	0.8596	0.8037	0.8596	0.8251	0.6118
Shallow SVM	0.78	0.84	0.78	0.80	0.625
Shallow Random Forest	0.76	0.88	0.76	0.80	0.7410
Shallow KNN	0.71	0.83	0.71	0.76	0.5892
Work in Brown et al. [8]	NA	0.70	0.90	NA	0.87
Work in Bansal et al. [16] (re-imp)	0.8789	0.8590	0.8789	0.8419	0.6559
Work in Imran et al. [9] (re-imp)	0.8498	0.8193	0.8469	0.8998	0.8816
Work in Coppock et al. [19] (re-imp)	0.7223	0.7680	0.7223	0.6936	0.6303

- Chromagram from power spectrogram with same parameters used in AUCO ResNet setup in Section 4.3 [53]
- Spectral centroid [54]. Parameters are the same as AUCO ResNet setup in Section 4.3
- Spectral bandwidth [54] with same parameters of AUCO as reported in Section 4.3
- Spectral roll-off [55] with same parameters of AUCO as reported in Section 4.3
- Zero-crossing rate [54] with same parameters of AUCO and reported in Section 4.3

Features are stacked horizontally to create one single instance row for each sound to be classified.

For classification purposes, the Random Forest classifier with 100 trees as base learners and pre-pruning depth set to 5, the linear Support Vector Machine and the weighted K-nearest neighbors with Euclidean distance and K = 7 were used. Features were standardized using z-score normalization.

5. Results and discussion

5.1. Camrbidge Tasks results

Fig. 6 shows the various AUC scores obtained by the AUCO Resnet proposed here for each fold of the 10-Fold cross validation performed on each Cambridge [8] task. As it is possible to observe, the AUC is generally high for all folds and all tasks.

Comparisons with other different techniques are reported in Tables 2, 3 and 4 which are referred, respectively, to the three Cambridge tasks as described in paragraph 4.1. As it can be easily observed, Auditory Cortex ResNet proposed here outperforms all the other approaches. Moreover, it performs with reasonable accuracies even when the amount of data is limited and the network is trained end-to-end (without transfer learning). The reason of this robustness is due to the use of the three different attention mechanisms and the use of the learnable Mel-like spectrograms. Detailed ablation is in Section 5.3. It has been proved that switching off the sinusoidal learnable attention causes a decrease of 1.3% AUC on Cambridge Task 1, thus achieving an AUC equal to that in the work [8]. This behavior is justified by the fact that the sinusoidal learnable attention can capture fine grained information within the respective feature map at different levels. As a consequence, it was able to merge only the relevant information. In addition, the trainable mel-like spectrograms contributed to the achieved results by filtering only the Mel-bands that extrapolated meaningful 2D representation of the audio.

5.2. Cross Datasets Tests

Results related to the cross-dataset tests are reported in Tables 5 and 6 respectively for cough and breath. AUCO ResNet always outperforms all the other approaches in terms of accuracy and AUC scores, however the best generalization capability is reached only when transfer learning is performed. This result suggests that, when trained with transfer learning, the network at former layers learns the internal representations of filter maps (kernels) that are used to extract relevant patterns from Mel-like spectrograms. If large amount of data is not present, AUCO ResNet can still deliver results that are better than those from the standard Deep Neural Network (DNN) architectures and even higher than those from pre-trained standard DNNs. This is probably due to the presence of the attention mechanisms, especially the sinusoidal learnable attention one, where sinusoidal representation layers can represent the information with low amount of data as input. Additionally, an initial blurred representation of log-frequencies patterns or a careful automatic selection of Mel amplitudes could have played major roles. However, more investigations are needed.

Considering tests on breath sounds (Table 6), results are lower than those obtained on cough. This is probably due to the fact that breath sounds were disturbed by the noise generated by the air blowing on the microphone.

In all tests, shallow learning techniques' performance is worse when compared with DNNs and in particular AUCO ResNet. One single exception is in Camrbidge task 2 in Table 3, where the shallow learning technique using Random Forests reported a higher

Table 4
Models' comparison on Cambridge Task 3.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
AUCO ResNet	0.8630	0.7198	0.8230	**0.8439**	**0.8972**
DenseNet 201	0.8548	0.7771	0.8548	0.8008	0.7478
Inception ResNet V2	0.8212	0.7946	0.8212	0.7979	0.8181
ResNet 50	0.8760	0.8541	0.8760	0.8425	0.8489
Shallow SVM	0.66	0.84	0.55	0.65	0.4523
Shallow Random Forest	0.71	0.83	0.71	0.77	0.3857
Shallow KNN	0.55	0.84	0.55	0.65	0.4523
Work in Brown et al. [8]	NA	0.61	0.81	NA	0.88
Work in Bansal et al. [16] (re-imp)	0.9013	0.9049	0.9013	0.8805	0.7575
Work in Imran et al. [9] (re-imp)	0.8462	0.7806	0.8462	0.8055	0.8709
Work in Coppock et al. [19] (re-imp)	**0.8648**	**0.8212**	**0.8648**	0.8298	0.8776

Table 5
Models comparison cross dataset cough tests.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
AUCO ResNet non transfer	0.7412	0.6732	0.7412	0.6892	0.6220
AUCO ResNet transfer	**0.7688**	**0.7326**	**0.7688**	**0.7098**	**0.8186**
DenseNet 201 non transfer	0.5180	0.5968	0.5180	0.3949	0.5518
DenseNet 201 transfer	0.6315	0.6126	0.6315	0.6027	0.6013
ResNet 50 non transfer	0.5037	0.7506	0.5037	0.3403	0.5634
ResNet 50 transfer	0.5834	0.6956	0.5834	0.4698	0.5821
Inception ResNet V2 non transfer	0.5505	0.5742	0.5505	0.5114	0.5576
Inception ResNet V2 transfer	0.5977	0.5994	0.5977	0.5885	0.6012
Shallow SVM	0.42	0.35	0.42	0.34	0.4234
Shallow Random Forest	0.51	0.53	0.51	0.41	0.5102
Shallow KNN	0.41	0.36	0.41	0.35	0.4132
Work in Brown et al. [8]	0.66	0.66	0.67	0.66	0.6221
Work in Bansal et al. [16] (re-imp)	0.62	0.67	0.62	0.64	0.5814
Work in Imran et al. [9] (re-imp)	0.72	0.59	0.72	0.63	0.641
Work in Coppock et al. [19] (re-implemented)	0.72	0.66	0.72	0.67	0.5253

Table 6
Models' comparison cross dataset breath tests.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
AUCO ResNet non transfer	0.59	0.62	0.59	0.55	0.6436
AUCO ResNet Transfer	**0.6822**	**0.7031**	**0.6816**	**0.6724**	**0.7051**
DenseNet 201 non transfer	0.5403	0.5776	0.5403	0.4774	0.5811
DenseNet 201 transfer	0.6236	0.6521	0.6235	0.6531	0.6612
ResNet50 non transfer	0.5967	0.6107	0.5967	0.5836	0.6269
ResNet50 transfer	0.6644	0.6824	0.6644	0.6623	0.6971
Inception ResNet V2 non transfer	0.5666	0.5694	0.5666	0.5622	0.5330
Inception ResNet V2 transfer	0.6421	0.6247	0.6241	0.6316	0.6542
Shallow SVM	0.51	0.51	0.51	0.50	0.5080
Shallow Random Forest	0.52	0.52	0.52	0.52	0.5161
Shallow KNN	0.52	0.52	0.52	0.51	0.5161
Work in Brown et al. [8]	0.55	0.55	0.55	0.55	0.5466
Work in Bansal et al. [16] (re-imp)	0.61	0.62	0.61	0.60	0.5829
Work in Imran et al. [9] (re-imp)	0.62	0.63	0.63	0.62	0.6691
Work in Coppock et al. [19] (re-imp)	0.59	0.61	0.59	0.58	0.6136

AUC with respect to other state of art DNNs, but lower if compared with AUCO ResNet.

5.3. UrbanSound 8K Test

Given that good results were obtained by AUCO-ResNet on Covid-19 related tasks, tests have been also performed on the wide UrbanSound 8K dataset [22]. Results obtained on the different folds are reported in Fig. 7.

Tests have been also performed considering data augmentation. For this purpose a time stretching technique has been adopted with the aim of reducing the speed of audio samples and background noise [22]. This data augmentation was performed on the fly on each training set of the 10-folds of UrbanSound 8K. Results are reported in Table 7. It can be observed that AUCO ResNet without data augmentation outperforms all the other solutions that do not use data augmentation.

On the other hand, the approach does not perform well in the case of data augmentation. State of art accuracy is achieved by Shin et al. in [28], where both data augmentation and transfer learning were adopted. Confusion matrices for AUCO-ResNet are reported in Fig. 8. It is possible to observe that data augmentation

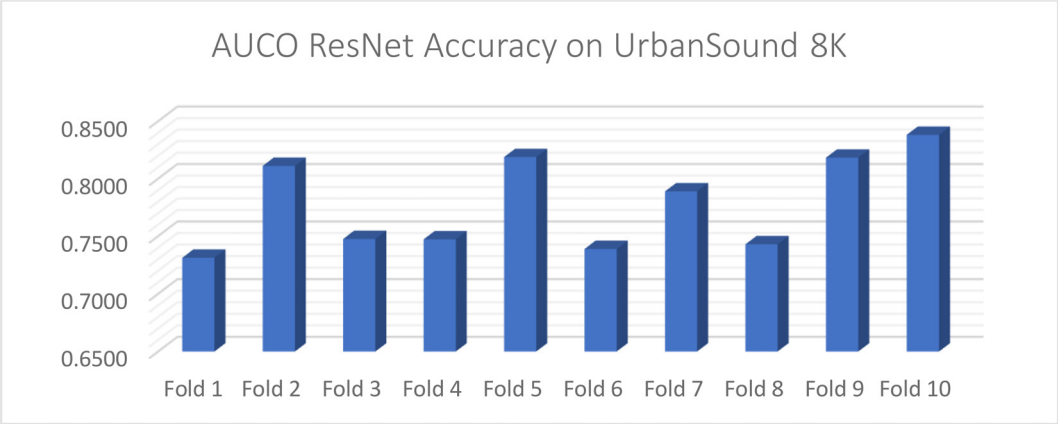

Fig. 7. Accuracies of AUCO ResNet on original (unaltered) folds of UrbanSoud 8K dataset.

Table 7
Urbansound 8K results compared with other state of the art works. NA means not acknowledged. NO means not present and YES means the presence of the pre-training or data augmentation technique.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC	Pre Trained	Data Augmentation
AUCO ResNet	0.7783	0.7851	0.7783	0.7709	0.9677	NO	NO
Chong et al. [27]	0.751	NA	NA	NA	NA	NO	NO
Salamon et al. [25]	0.75	NA	NA	NA	NA	NO	NO
Giannakopoulos et al. [29]	0.731	NA	NA	NA	NA	NO	NO
Salamon et al. [23]	0.73	NA	NA	NA	NA	NO	NO
Piczac et al. [24]	0.73	NA	NA	NA	NA	NO	NO
Jin et al. [26]	0.705	NA	NA	NA	NA	NO	NO
Salamon et al. [22]	0.70	NA	NA	NA	NA	NO	NO
Shin et al. [28]	**0.8514**	NA	NA	NA	NA	YES	YES
Shin et al. [28] without transfer learning	0.7632	NA	NA	NA	NA	NO	YES
Zhang et al. [56]	0.819	NA	NA	NA	NA	NO	YES
AUCO ReseNet with data augmentation	0.7222	0.7329	0.7222	0.7146	0.9413	NO	YES

	1	2	3	4	5	6	7	8	9	10
1	35	0	0	14	10	39	0	2	0	0
2	0	35	0	1	0	0	0	0	0	0
3	1	0	81	7	0	4	0	0	3	4
4	0	3	6	84	1	3	0	0	2	1
5	7	1	2	4	59	10	2	3	3	9
6	7	0	0	0	28	60	0	0	0	1
7	0	0	0	1	0	0	33	1	0	0
8	5	0	0	0	50	7	0	58	0	0
9	0	0	10	2	3	1	0	0	69	1
10	2	0	3	0	0	4	0	2	0	89

	1	2	3	4	5	6	7	8	9	10
1	49	0	2	22	10	6	0	10	1	0
2	0	36	0	0	0	0	0	0	0	0
3	0	0	87	3	0	1	0	0	1	8
4	0	1	5	85	1	1	3	0	3	1
5	18	0	5	0	43	1	2	8	13	10
6	4	0	0	0	31	60	0	1	0	0
7	0	3	0	0	1	1	30	0	0	0
8	0	0	0	0	45	12	0	63	0	0
9	1	0	9	2	2	0	0	0	72	0
10	4	0	5	0	0	1	0	5	3	82

Fig. 8. Results on first fold of UrbanSound 8K. (a) left confusion matrix of AUCO ResNet trained on augmented training set. (b) right confusion matrix of AUCO ResNet trained on non-augmented training set. Names of columns and rows are: 1 = Air Conditioned, 2 = Car Horn, 3 = Child, 4 = Dog Barking, 5 = Drilling, 6 = Engine, 7 = Gun Shot, 8 = Jackhammer, 9 = Siren, 10 = Street Music.

has created some confusions in recognizing the right patterns for some classes such as engine sound (confused with air conditioner) and drilling class (confused with jackhammer). It is in the authors' opinion that data augmentation increases the chance to have features maps sharing the same weights among the different classes so that the first layer of the net (trainable Mel-like filter) specializes itself more on synthetic data than on the real ones. Moreover, data augmentation reduces reproducibility and is questionable in the medical field, as a consequence it has not been used in experiments related to Covid-19 [57].

5.3.1. Ablation Studies

Ablation studies are presented in Table 8. The intuition behind the end-to-end training is the capability of the system to optimize every single module of the entire learning system toward a global goal (decreasing the error). As previously discussed, the inclusion of the trainable mel-spectrogram layer within the end-to-end training process is aimed to optimize the learned spectrogram representation with the goal of decreasing the error. But this process does not happen in isolation, Table 8 shows that only the synergic integration of trainable Mel-spectrogram layer, CBAM, SE-Block and sinusoidal learnable attention allows to achieve the best results in terms of average accuracy, precision, recall, F1-score and AUC ROC on UrbanSound 8K dataset trained end-to-end. Specifically, as reported in the works [41] and [44], it is observed that adding Squeeze and Excitation block into a ResNet architecture improved the AUC ROC as well as adding the CBAM block, although the latter with a little increase in AUC ROC. Results in Table 8 also confirm the importance of adding the learnable attention as noted in [45] achieving same importance of the Squeeze and Excitation

Table 8
Ablation studies performed on UrbanSound 8K dataset. Results show average 10-fold scores achieved with original (vanilla) AUCO ResNet and the modified AUCO ResNet where, one at time, various properties have been disabled. NT means non trainable. SLA means sinusoidal learnable attention.

Model	Accuracy	Precision	Recall	F1 score	AUC ROC
Vanilla AUCO ResNet	0.7783	0.7851	0.7783	0.7709	0.9677
AUCO with NT Mel	0.7620	0.7706	0.7629	0.7542	0.9499
AUCO without CBAM-Block	0.7779	0.7843	0.7779	0.7708	0.9531
AUCO without SE-Block	0.7715	0.7849	0.7715	0.7673	0.9538
AUCO without SLA	0.7687	0.7768	0.7687	0.7625	0.9538
AUCO without CBAM, SE-Block, SLA and NT Mel	0.7617	0.7649	0.7617	0.7518	0.9494

Table 9
Deep Neural Networks complexity in terms of number of trainable weights.

Model	Trainable Weights
AUCO ResNet	22128290
DenseNet 201	18090498
ResNet 50	23517186
Inception ResNet V2	54278690
Work in Bansal et al. [16] (re-imp)	44953883
Work in Imran et al. [9] (re-imp)	65506
Work in Coppock et al. [19] (re-imp)	15276096

block in [41]. These results are confirmed by the respective works in [41,43] and [44]. It is also important to underline that the results obtained are not a mere sum of the improvements from each technique, thus showing that the various attention techniques compete with each other in a sensitive way and also the limits of the overall model and datasets.

5.3.2. Limitations of AUCO ResNet

AUCO ResNet is a novel deep neural network that can be valuable for audio classification tasks, but this comes with a cost as it is not without some limits. It is not clear why data augmentation decreases the accuracy of the model. In fact, results on UrbanSound 8K dataset are lower when compared to techniques such as [28] that make use of data augmentation. Our intuition is that AUCO ResNet is very sensitive to the input audio quality. It is also important the complexity of training the network: from Table 9 it is possible to observe that its complexity in terms of number of trainable parameters is in the average. It is by far less than an Inception ResNet V2 or Bansal et al. in [16], but it is higher than a DenseNet with 201 layers as well as other deep neural networks designed specifically for Covid-19 recognition from audio such as Imran et al. [9] and Coppock et al. [19]. The number of trainable parameters in the AUCO ResNet is governed by the presence of Squeeze and Excitation blocks which add dense layers for channel attention and recalibration, but their presence is of paramount importance as shown in ablation studies and results in Table 8. Implementation of vanilla pre-trained AUCO ResNet models in edge AI devices is difficult but not impossible: in paper [58] authors were capable of running ResNet-50 with 23.6 million trainable parameters and Inception V3 with 21.8 million trainable parameters on an edge AI device, and AUCO ResNet complexity stands just in the middle of those two, but it requires a modern edge AI device. Another limitation is that it cannot process large audio inputs of minutes and hours because the tensors within the network would fill the available GPU RAM, differently from models that are intrinsically recurrent.

6. Conclusions and future remarks

In this work a biologically inspired deep neural network, namely the Auditory Cortex ResNet, for audio classification has been presented. AUCO ResNet takes as input raw audio and works without any preprocessing, data augmentation or manual spectrogram generation.

AUCO ResNet has been tested on several Covid-19 sound datasets including cough and breath recordings and adopting an inter-patient schema. Moreover, cross dataset tests (training/test) have been performed. This neural network has been compared to a wide set of state of art approaches and it has reported an AUC of 0.8308, 0.9257 and 0.8972 respectively on Cambridge Task1, Task2 and Task3 as defined in [8]. Improvements (compared to the other approaches) range from 3% to 20% of AUC. It is also interesting to observe that, given the independence of cough and breath from the language, cross-dataset tests have been possible. In this case 0.8186 and 0.7051 of AUC, respectively on cough and breath have been achieved thus demonstrating that the system proposed here can be a viable Covid-19 pre-screening solution.

The approach has been also tested on the wide UrbanSound 8K dataset obtaining state of the art performances without data augmentation. On the other hand, it has been shown that the approach reduces its performance in case of data augmentation. AUCO ResNet is a complex network with 400 layers and number of trainable weights just slightly below a common ResNet-50, it exploits some parallel computation with the sinusoidal learnable attention mechanism, but the presence of several CNN layers limits its parallelization. In addition, it cannot process long audio sequences, as it would a recurring neural network: the size of the tensors, and consequently the GPU ram used, would get saturated.

Future works efforts will be addressed to reduce the network complexity, to investigate different configurations of the different layers as well as to stress the generalization capabilities of the network on other audio-related health tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. A comparison of trainable and non-trainable mel spectrogram layer

In this appendix there are plots of trainable and non-trainable Mel-spectrogram and their differences. The Mel-spectrogram is a spectrogram generated by converting sound frequencies in the Mel scale. The key insight about using a trainable Mel-spectrogram layer is to seek within the end-to-end training framework definition. Thanks to the end-to-end approach, the training of a weight matrix and its dot product with the original Mel-filterbank, produce a spectrogram that is optimized for the global goal of reducing the error by enhancing only frequencies that are important. As it is possible to observe from Figs. 9 and 10, the two spectrograms are visually very similar, but Fig. 11 shows their difference by using the known structural similarity algorithm. It seems that in certain places of the spectrogram (e.g. areas localized toward the low-

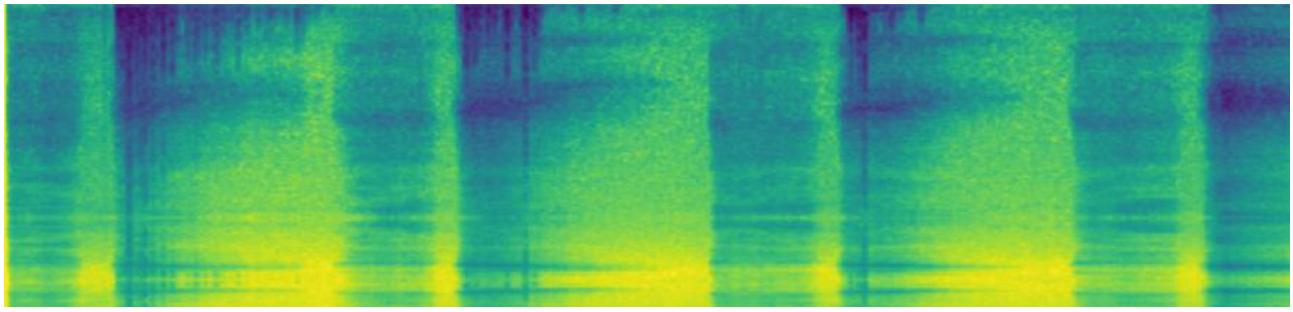

Fig. 9. Non trained Mel-spectrogram for first covid-19 positive audio of the entire dataset.

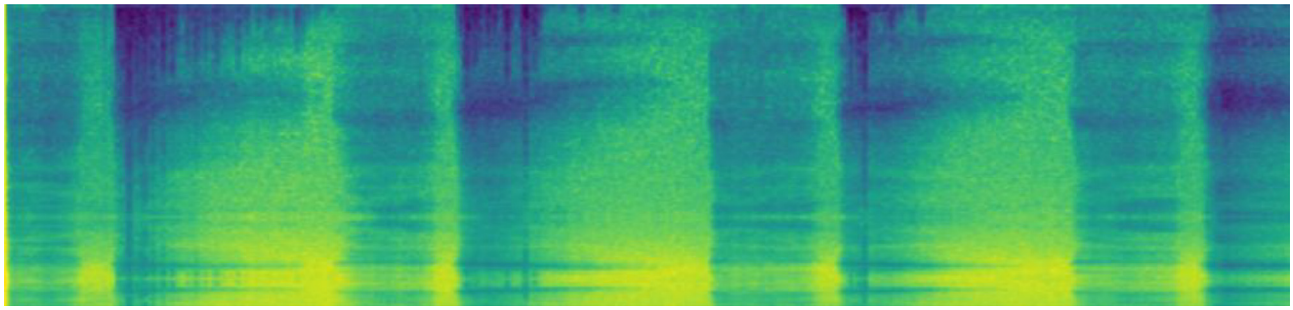

Fig. 10. Trained Mel-like spectrogram for first covid-19 positive audio of the entire dataset.

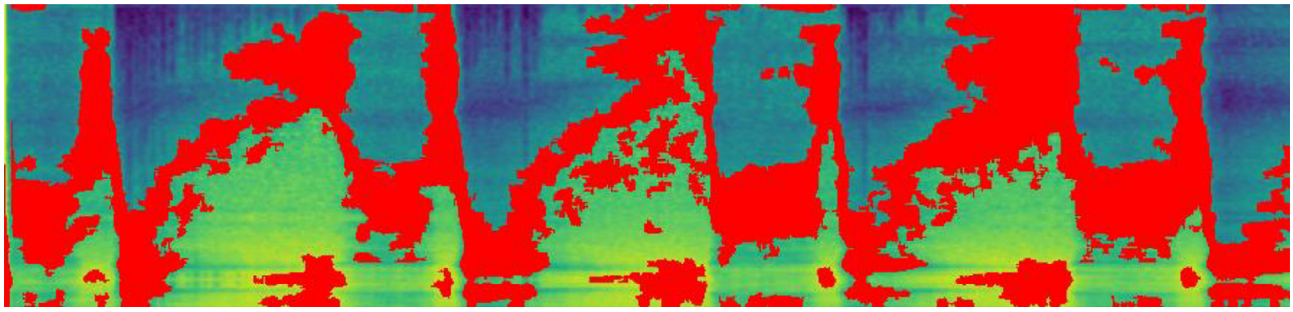

Fig. 11. differences (marked in red) from trained and non-trained Mel-spectrograms using structural similarity algorithm.

band frequencies), the magnitude of some frequencies is higher in non-trainable Mel-spectrogram and somewhat less pronounced on the trained Mel-spectrogram, thus confirming authors' hypotheses of performing additive selection on Mel-filterbanks for filtering out less important information.

References

[1] C. Sohrabi, et al., World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), International journal of surgery (London, England) 76 (2020) 71–76, doi:10.1016/J.IJSU.2020.02.034.

[2] L. Falzone, G. Gattuso, A. Tsatsakis, D.A. Spandidos, M. Libra, Current and innovative methods for the diagnosis of COVID-19 infection (Review), International Journal of Molecular Medicine 47 (6) (2021) 1–23, doi:10.3892/IJMM.2021.4933/HTML.

[3] J. Li, et al., Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19, Pattern recognition 114 (2021), doi:10.1016/J.PATCOG.2021.107848.

[4] Z. Wang, et al., Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognition 110 (2021) 107613, doi:10.1016/J.PATCOG.2020.107613.

[5] A.R. Watson, R. Wah, R. Thamman, The Value of Remote Monitoring for the COVID-19 Pandemic, Telemedicine journal and e-health : the official journal of the American Telemedicine Association 26 (9) (2020) 1110–1112, doi:10.1089/TMJ.2020.0134.

[6] J. Portnoy, M. Waller, T. Elliott, Telemedicine in the Era of COVID-19, The journal of allergy and clinical immunology. In practice 8 (5) (2020) 1489–1491, doi:10.1016/J.JAIP.2020.03.008.

[7] M.J. Mina, R. Parker, D.B. Larremore, Rethinking Covid-19 Test Sensitivity — A Strategy for Containment, New England Journal of Medicine 383 (22) (2020) e120, doi:10.1056/NEJMP2025631/SUPPL_FILE/NEJMP2025631_DISCLOSURES.PDF.

[8] C. Brown, et al., Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 3474–3484, doi:10.1145/3394486.3412865.

[9] A. Imran, et al., AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, Informatics in medicine unlocked 20 (2020), doi:10.1016/J.IMU.2020.100378.

[10] P. Bagad, et al., Cough against COVID: Evidence of COVID-19 signature in cough sounds (2020).

[11] J. Laguarta, F. Hueto, B. Subirana, COVID-19 Artificial Intelligence diagnosis using only cough recordings, IEEE Open J Eng Med Biol 1 (2020) 275–281.

[12] A. Pal, M. Sankarasubbu, Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing (2020).

[13] M. al Ismail, S. Deshmukh, R. Singh, Detection of COVID-19 through the analysis of vocal fold oscillations, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2021-June, 2020, pp. 1035–1039, doi:10.1109/ICASSP39728.2021.9414201.

[14] N. Sharma, et al., Coswara–A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis (2020).

[15] R. Dunne, T. Morris, S. Harper, High accuracy classification of COVID-19 coughs using Mel-frequency cepstral coefficients and a Convolutional Neural Network with a use case for smart home devices (2020).

[16] V. Bansal, G. Pahwa, N. Kannan, Cough Classification for COVID-19 based on audio mfcc features using Convolutional Neural Networks (2020).

[17] G. Chaudhari, et al., Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough (2020).

[18] A. Imran, et al., AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app (2020).

[19] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, B. Schuller, End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study, BMJ Innov 7 (2) (2021) 356–362.
[20] M.B. Alsabek, I. Shahin, A. Hassan, Studying the similarity of COVID-19 sounds based on correlation analysis of MFCC (2020).
[21] T. Tommasi, N. Patricia, B. Caputo, T. Tuytelaars, A Deeper Look at Dataset Bias, Advances in Computer Vision and Pattern Recognition (9783319583464) (2017) 37–55, doi:10.1007/978-3-319-58347-1_2.
[22] J. Salamon, C. Jacoby, J.P. Bello, A dataset and taxonomy for urban sound research (2014).
[23] J. Salamon, J.P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal Process. Lett. 24 (3) (2017) 279–283.
[24] K.J. Piczak, Environmental sound classification with convolutional neural networks (2015).
[25] J. Salamon, J.P. Bello, Unsupervised feature learning for urban sound classification (2015).
[26] X. Jin, et al., WSNet: Compact and efficient networks through weight sampling (2017).
[27] D. Chong, Y. Zou, W. Wang, Multi-channel convolutional neural networks with multi-level feature fusion for environmental sound classification, in: MultiMedia Modeling, Springer International Publishing, Cham, 2019, pp. 157–168.
[28] H.-K. Shin, S.H. Park, K.-W. Kim, Inter-floor noise classification using convolutional neural network, PLoS One 15 (12) (2020) e0243758.
[29] T. Giannakopoulos, E. Spyrou, S.J. Perantonis, Recognition of urban sound events using deep context-aware feature extractors and handcrafted features, in: IFIP Advances in Information and Communication Technology, Springer International Publishing, Cham, 2019, pp. 184–195.
[30] C.R. Rodriguez, D. Angeles, R. Chafloque, F. Kaseng, B. Pandey, Deep learning audio spectrograms processing to the early COVID-19 detection (2020).
[31] F. Chollet, Xception: Deep learning with depthwise separable convolutions (2017).
[32] A. Hossain, F. Lover, A.A. Corey, G.A. Reich, N.G. Rahman, FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4, 2020, pp. 1–28.
[33] A. Vaswani, et al., Attention Is All You Need, Advances in Neural Information Processing Systems 2017-December (2017) 5999–6009 Accessed: Nov. 11, 2021. [Online]. Available https://arxiv.org/abs/1706.03762v5 .
[34] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G Bachar, V Aharonson, Covid-19 Detection From Voice, IEEE Open Journal of Engineering in Medicine and Biology 1 (2020) 268–274.
[35] M. Moerel, F. de Martino, K. Uğurbil, E. Yacoub, E. Formisano, Processing complexity increases in superficial layers of human primary auditory cortex, Sci. Rep. 9 (1) (2019) 5502.
[36] I. Stiebler, R. Neulist, I. Fichtel, G. Ehret, The auditory cortex of the house mouse: left-right differences, tonotopic organization and quantitative analysis of frequency representation, J. Comp. Physiol. A 181 (6) (1997) 559–571.
[37] K.D. Games, J.A. Winer, Layer V in rat auditory cortex: projections to the inferior colliculus and contralateral cortex, Hear. Res. 34 (1) (1988) 1–25.
[38] A.K.C. Lee, et al., Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch, Front. Neurosci. 6 (2012) 190.
[39] C.H. Tischbirek, T. Noda, M. Tohmi, A. Birkner, I. Nelken, A. Konnerth, In vivo functional mapping of a cortical column at single-neuron resolution, Cell Rep 27 (5) (2019) 1319–1326.e5.
[40] A. Khan, A. Sohail, U. Zahoora, A.S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, Artif. Intell. Rev. 53 (8) (2020) 5455–5516.
[41] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (8) (2017) 2011–2023, doi:10.1109/TPAMI.2019.2913372.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2016).
[43] F. He, T. Liu, D. Tao, Why ResNet works? Residuals generalize, IEEE Trans. Neural Netw. Learn. Syst. 31 (12) (2020) 5349–5362.
[44] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional Block Attention Module, in: Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 3–19.
[45] S. Jetley, N.A. Lord, N. Lee, P.H.S. Torr, Learn to pay attention (2018).
[46] V. Sitzmann, J.N.P. Martel, A.W. Bergman, D.B. Lindell, G. Wetzstein, Implicit neural representations with periodic activation functions (2020).
[47] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023.
[48] T. Moldwin, I. Segev, Perceptron learning and classification in a modeled cortical pyramidal cell, Front. Comput. Neurosci. 14 (2020) 33.
[49] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate (2014) Accessed: Nov. 09, 2021. [Online]. Available https://arxiv.org/abs/1409.0473v7 .
[50] F. Zou, L. Shen, Z. Jie, W. Zhang, W. Liu, A sufficient condition for convergences of Adam and RMSProp (2019).
[51] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks (2017).
[52] C. Szegedy, S. Ioffe, Vincent Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the impact of residual connections on learning (2016).
[53] C. Harte, M. Sandler, Automatic chord identifcation using a quantised chromagram, Audio Engineering Society Convention, 118, Audio Engineering Society, 2005.
[54] A. Klapuri, M. Davy, Signal processing methods for music transcription, Springer, New York, NY, 2006.
[55] M.N. Stolar, M. Lech, S.J. Stolar, N.B. Allen, Detection of adolescent depression from speech using optimised spectral roll-off parameters, Biomedical Journal 2 (2018).
[56] X. Zhang, Y. Zou, W. Shi, Dilated convolution neural network with LeakyReLU for environmental sound classification (2017).
[57] F. Renard, S. Guedria, N. de Palma, N. Vuillerme, Variability and reproducibility in deep learning for medical image segmentation, Sci. Rep. 10 (1) (2020) 13724.
[58] P.T. Anh, H.T.M. Duc, A Benchmark of Deep Learning Models for Multi-leaf Diseases for Edge Devices (2021) 318–323, doi:10.1109/ATC52653.2021.9598196.

Vincenzo Dentamaro (M'20) received the degree in computer science from the Department of Computer Science, University of Bari, Italy and a Master Of Science in Machine Learning from Gerogia Institute of Technology Atlanta USA. He is currently PhD student at University of Bari "Aldo Moro" with scholarship offered by InnovaPuglia S.p.A. Vincenzo is currently publishing on various pattern recognition journals and conferences. He is also reviewer for Elsevier Pattern Recognition Journal, IEEE Access, MDPI Sensor, MDPI Information and many more. He has previously published about indoor positioning and localization techniques on Microsoft Research Journal and holds two international patents on localization technologies. Previous work experience at Johnson Controls Inc. as Software Engineer, IBM Rome as an intern, CEO and CTO Nextome S.R.L. Seal of Excellence European Commission, 1st prize Busan Metropolitan City (South Korea), IBM's Global Mobile Innovator Tournament Award at the Mobile World Congress, MIT Technology Review award.

Paolo Giglio was born on January 2, 1990 in Bari. After graduating in Physics at the University of Bari he moved to Lecce where he achieved a master degree in Physics (Condensed Matter Physics and Nanotechnology) at the University of Salento, Department of Physics Ennio De Giorgi. Particularly he performed the research thesis "Beam Steering with Phase-Only Spatial Light Modulator for Optogenetic Applications" at the Center for Biomolecular Nanotechnologies (CBN) of the Istituto Italiano di Tecnologia (IIT) in Arnesano (LE). He then moved to Milan to work in the private sector, particularly as a python/javascript developer with AWS technology, designing products for customers like Campari and Pirelli. He is currently a PHD student at the University of Bari (Department of Informatics), with a Thesis on Artificial Intelligence applications in the Smart City context, particularly focalizing on the field of Urban Security.

Donato Impedovo (M'08-SM'17) is associate professor at Department of Computer Science of the University of Bari (IT). His research interests are in the field of signal processing, pattern recognition, machine learning and biometrics. He is co-author of more than 100 articles on these fields in both international journals and conference proceedings. He received the "distinction" award in May 2009 at the International Conference on Computer Recognition Systems (CORES – endorsed by IAPR), and the first prize of the first Nereus-Euroavia Academic competition on GMES in October 2012. Prof. Impedovo is also very involved in research transfer activities as well as in industrial research, he has managed more than 25 projects funded by public institutions as well as by private SMEs. Prof. Impedovo is IEEE Access and IEEE OJ-CS Associate Editor, he serves as reviewer for many international journals including IEEE THMS, IEEE T-SMC, IEEE-TIFS, IEEE-TECT, Pattern Recognition and many others. He serves as reviewer and rapporteur for the EU in H2020 projects evaluation. He was the general co-chair of the International Workshop on Smart Cities and Smart Enterprises (SCSE 2018), International Workshop on Artificial Intelligence with Application in Health (WAIAH 2018), Emergent Aspects in Handwritten Signature Processing (EAHSP 2013) and of the International Workshop on Image-Based Smart City Application (ISCA 2015). He has been member of the scientific committee and program committee of many international conferences in the field of computer science, pattern recognition and signal processing such as the ICPR and ICASSP. He is IAPR member and IEEE senior member.

Luigi Moretti has graduated in medicine and surgery at School of Medicine of Bari, University Aldo Moro of Bari, Italy and has obtained a certification as "Expert3D", in medical image post processing through 3D printing and artificial intelligence, at Sant Joan De Déu Barcelona Hospital, Spain. He is currently registered as a medical doctor at the Association of Doctors and Surgeons of Bari region, Italy. In 2018, thanks to the Pugliesi Innovativi (PIN) grant, he started a company, named FabCraft Srls, based on the use of digital manufacturing technologies; during the summer of the same year, he has attended an IFMSA research exchange in the Immunology Department of the University of Sfax, Tunisia. Moreover, in 2019 he had the chance to present, at Premio Nazionale dell'Innovazione (PNI), the main national event for projects and spin-offs by universities and research institutions in Italy, a project named MyHealthData (MyHD) which aimed to provide a patient-centric universal tool for health data management both in standard and emergency scenarios.

Giuseppe Pirlo (M'92–SM'13) received the degree in computer science (cum laude) from the Department of Computer Science, University of Bari, Italy, in 1986. Since 1986, he was carrying out research in the field of computer science and neuroscience, signal processing, handwriting processing, automatic signature verification, biometrics, pattern recognition and statistical data processing. Since 1991, he was an Assistant Professor with the Department of Computer Science, University of Bari,

where he is currently a Full Professor. He developed several scientific projects and authored over 250 papers on international journals, scientific books and proceedings. Prof. Pirlo is currently an Associate Editor of the IEEE Transactions on Human–Machine Systems. He also serves as a Reviewer for many international journals including the IEEE T-PAMI, the IEEE T-FS, the IEEE T-SMC, the IEEE T-EC, the IEEE T-IP, the IEEE T-IFS, the Pattern Recognition, the IJDAR, and the IPL. He was the general co-chair of the International Workshop on Smart Cities and Smart Enterprises (SCSE2018), of the International Workshop On Artificial Intelligence With Application In Health (WAIAH2017), of the International Workshop on Emerging Aspects in Handwriting Signature Processing, Naples, in 2013, the International Workshop on Image-based Smart City Applications, Genoa, in 2015, and the General Co-Chair of the International Conference on Frontiers in Handwriting Recognition, Bari, in 2012. He was a reviewer in the scientific committee and program committee of many international conferences in the field of computer science, pattern recognition and signal processing, such as the ICPR, the ICDAR, the ICFHR, the IWFHR, the ICIAP, the VECIMS, and the CISMA. He is also the editor of several books. He was an Editor of the Special Issue Handwriting Recognition and Other PR Applications of the Pattern Recognition Journal in 2014 and the Special Issue Handwriting Biometrics of the IET Biometrics Journal in 2014. He was the Guest Editor of the Special Issue of the Je-LKS Journal of the e-Learning and Knowledge Society Steps toward the Digital Agenda: Open Data to Open Knowledge in 2014. He is currently the Guest Co-Editor of the Special Issue of the IEEE Transactions on Human–Machine Systems on Drawing and Handwriting Processing for User-Centered Systems. Prof. Pirlo is a member of the Governing Board of Consorzio Interuniversitario Nazionale per l'Informatica (CINI), a member of the Governing Board of the Società Italiana di e-Learning and the e-learning Committee of the University of Bari. He is currently the Deputy Representative of the University of Bari in the Governing Board of CINI. He is also the Managing Advisor of the University of Bari for the Digital Agenda and Smart Cities. He is the Chair of the Associazione Italiana Calcolo Automatico-Puglia. He is also a member of the Gruppo Italiano Ricercatori Pattern Recognition, the International Association Pattern Recognition, the Stati Generali dell'Innovazione, and the Gruppo Ingegneria Informatica.