Hierarchical Multi-modal Transformer for Automatic Detection of COVID-19

Shuyun Tang
University of California, Berkeley;
Virufy AI Research
Berkeley, CA
shuyuntang@berkeley.edu

Xinying Hu
University of California, Berkeley;
Virufy AI Research
Berkeley, CA
xinying_hu@berkeley.edu

Les Atlas
University of Washington
Seattle, WA
atlas@uw.edu

Amil Khanzada
University of California, Berkeley;
Virufy AI Research
Berkeley, CA
amil@berkeley.edu

Mert Pilanci
Stanford University
Stanford, CA
pilanci@stanford.edu

ABSTRACT

Automated COVID-19 detection based on analysis of cough recordings has been an important field of study, as efficient and accurate methods are necessary to contain the spread of the global pandemic and relieve the burden on medical facilities. While previous works presented lightweight machine learning models [9], these models may sacrifice accuracy and interpretability to integrate into mobile devices. Besides, the question of how to effectively associate indicators from audio signals to other modality inputs (i.e. patient information) is still largely unexplored, as previous works predominantly relied on simply concatenated features to learn. To tackle these issues, this paper proposes a novel Hierarchical Multi-modal Transformer (HMT) that learns more informative multi-modal representations with a cross attention module during the feature fusion procedure. Besides, the block aggregation algorithm for the HMT provides an efficient and improved solution from the Vanilla Vision Transformer for limited COVID-19 benchmark datasets. Extensive experiments show the effectiveness of our proposed model for more accurate COVID-19 detection, which yield state-of-the-art results on two public datasets, Coswara and COUGHVID.

CCS CONCEPTS

• **Networks**; • **Applied computing**; • **Information systems**;

KEYWORDS

Machine Learning, Neural Networks, Signal Processing, Transformer, Artificial Intelligence in Health, Accountable Artificial Intelligence

ACM Reference Format:
Shuyun Tang, Xinying Hu, Les Atlas, Amil Khanzada, and Mert Pilanci. 2022. Hierarchical Multi-modal Transformer for Automatic Detection of COVID-19. In *2022 5th International Conference on Signal Processing and Machine Learning (SPML 2022), August 4–6, 2022, Dalian, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3556384.3556414

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SPML 2022, August 4–6, 2022, Dalian, China
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9691-2/22/08.
https://doi.org/10.1145/3556384.3556414

1 INTRODUCTION

The COVID-19 pandemic has shaken society globally, claiming millions of lives and causing trillions of dollars in damage to GDP. As of May 5th, 2022, the World Health Organization (WHO) reports 513M COVID positive cases and 6.2M deaths [1]. Experts claim that mass testing plays an essential role in social distancing, contact tracing and impeding the spread of the virus [17]. The gold standard for COVID-19 diagnosis is the PCR test. Although the PCR test receives general acceptance, this clinical test is limited to the areas with enough testing equipment, trained clinical staff, and other medical resources. Additionally, during infection spikes, testing centers are often overwhelmed with massive testing needs.

The development of machine learning techniques and digital health shows promise in providing an alternative solution for COVID-19 testing. Researchers have studied COVID-19 prediction from image analysis and audio analysis. Researchers in [23] used CT scanning to classify COVID-19 infected patients from healthy individuals. Minaee et al. predicted COVID-19 from around 5,000 chest X-ray images using deep transfer learning [13] However, X-ray and CT scanning still requires clinical equipment, and as such this image-based machine learning method is difficult to apply to general testing circumstances.

On the other hand, cough sounds have been utilized for many respiratory disease studies and have shown great value in conveying vital information. Korpa et al. [12] applied basic signal processing techniques, inspecting both time-domain waveform and frequency-domain periodograms and showed the difference between normal cough and inflammation cough. Furthermore, cough sound analysis has proved helpful in diagnosing respiratory conditions like pertussis, pneumonia and asthma [2, 16]. Since the COVID-19 pandemic, cough sound classification has been studied for developing affordable and diagnostic AI tools. And many researchers have contributed to building COVID-19 cough audio databases providing the foundation for COVID-19 cough analysis algorithm development.

Previous works [3, 6] have shown that combining three modalities as inputs: raw audio spectrogram, MFCCs, and clinical features could yield the optimal performance. However, these methods are

not effective in capturing the importance of different modalities and their correlation with the final prediction. The underlying reason is that different modalities are neither completely independent nor correlated, posing challenges for filtering out the noise and keeping useful information in each modality during fusion. Previous works only consider straightforward data concatenations to fuse these features from different modalities, which may yield suboptimal performance.

Transformer models [24] have become dominant deep learning architectures that outperform the traditional recurrent neural networks (RNNs) in many natural language processing benchmarks. Vision Transformer (ViT) [5] also shows improved results over the traditional convolutional neural networks (CNNs) in image recognition tasks. Transformer-based architectures such as the Audio Spectrogram Transformer (AST) [7, 8] also shows promising results over CNNs in processing and learning audio spectrograms. However, global self-attention between pixel pairs is computationally expensive and is hard to integrate with mobile devices. In addition, the AST leveraged the pretraining process, which is not practical as the quantity of cough audio files is limited. Data-efficient ViT (DeiT) [22] attempts to address this problem by introducing teacher distillation from a convolutional network. Despite its effectiveness to the limited data, it increases the training complexity and is not possible to leverage other clinical features during the training.

To tackle the above-mentioned challenges, in this paper, we propose Hierarchical Multi-modal Transformer (HMT), an end-to-end deep neural network that detects and classifies COVID-19 based on the cough sounds. It consists of three branches: Multilayer Perceptron (MLP) networks for the 1D clinical and audio features and transformer for the 2D audio spectrograms. Inspired by [15, 21], we propose a **cross-attention module** that fuses the intermediate representations from each branch, effectively capturing the relations between each of them and their importance contributed to the final prediction. Inspired by the work of Pan et al. [25], the transformer branch integrates nested transformer layers with **block aggregation algorithm**, which conducts local self-attention on every non-overlapping image block independently, and then nests them hierarchically. Not only this approach outperforms the traditional convolutional based models and vanilla ViT, but also substantially simplifies previous sophisticated (local) self-attention mechanisms and improves data efficiency.

In summary, our contributions are threefold:

1. We propose a novel transformer based neural network called HMT that can fuse audio signals and clinical features and explores their underlying interactions with a cross-attention module.

2. To the best of our knowledge, this is the first attempt to use hierarchically stacked transformers to classify the audio spectrograms. The block aggregation algorithm significantly improves the efficiency and cross-block communication without adding extra training complexity.

3. We conduct extensive experiments on two public COVID-19 cough audio datasets: Coswara and COUGHVID and achieve new state-of-the-art results. Additionally, ablation studies are conducted to demonstrate the the effectiveness of the proposed modules.

2 DATA AND PREPROCESSING

For this paper, we chose to work with two open-access cough audio-recording datasets for model training. The first is the Coswara dataset [19], a public crowd-sourced database prepared by the Indian Institute of Science (IISc) Bangalore for COVID-19 diagnostic study. The data is collected from healthy individuals and COVID-19 patients via a crowd-sourcing web portal. The dataset contains 2,746 audio samples with metadata collected between April 2020 and February 2022. Audio recordings include breathing, cough, and speech sounds of healthy and COVID-19 patients. Metadata consists of the participant's age, gender, location (country, state/province), current health status (healthy /exposed /positive /recovered) and the presence of comorbidities. We chose the "shallow-cough" recording as the input source among the various sound recordings in the dataset. From the metadata, we decided on "fever muscle pain" and "respiratory condition" as extra clinical information for prediction in a 1D vector of binary numbers. For COVID-19 status labeling, samples labeled with 'positive mild', 'positive moderate', and 'positive asymp' are classified in the positive class (n=681), and all other statuses are classified as negative (n=2,065).

The second open-access dataset we used for training is the COUGHVID dataset [14] prepared by the Swiss Federal Institute of Technology Lausanne (EPFL) for automatic COVID-19 screening research. COUGHVID provides over 25,000 crowd-sourced cough recordings, with 1,155 claiming COVID-19 positive (n=1,155) collected between April 1st, 2020 and December 1st, 2020 via crowd-sourcing in a web portal. Similar to the Coswara dataset, participant location, age, gender, and respiratory conditions are also included as labels. In the COUGHVID dataset, the "dry-cough" recording is chosen as the input audio source. We also included the common metadata "fever muscle pain" and "respiratory condition" as training features. To address quality issues in the crowd-sourced datasets, COUGHVID's cough detection classifier filters out audio files with the "cough-detected" level below 90%. As a result, there are 8868 useful samples in total, including 441 positive samples and 8427 negative samples.

To combine the two datasets of different sizes, we include the entire smaller dataset (Coswara) and a subset of the more extensive dataset (COUGHVID) as our training and evaluation dataset. Since the datasets are highly imbalanced, we include all data samples with COVID-19 positive class and randomly sampled 1,000 samples with COVID-19 negative class from the COUGHVID dataset. After manually balancing the dataset and removing data with missing labels, clinical features, or audio sources, our combined dataset contains 4,102 data samples, including 1,023 positive samples and 3,079 negative samples.

To extract information from the cough audios, we chose two main representations: mel-frequency cepstral coefficient (MFCC) and mel-frequency spectrogram, which are very commonly used for audio processing and speech recognition. MFCC is an acoustic parametric representation derived from the power spectrogram [4]. The audio signal is decomposed into an unevenly sampled time-frequency distribution that mimics the human auditory peripheral system's response to sound by a biologically inspired spectral filter [18]. Additionally, we use the Librosa package to extract the first 39 MFCC features of each frame, from a 2048-window-length Fast

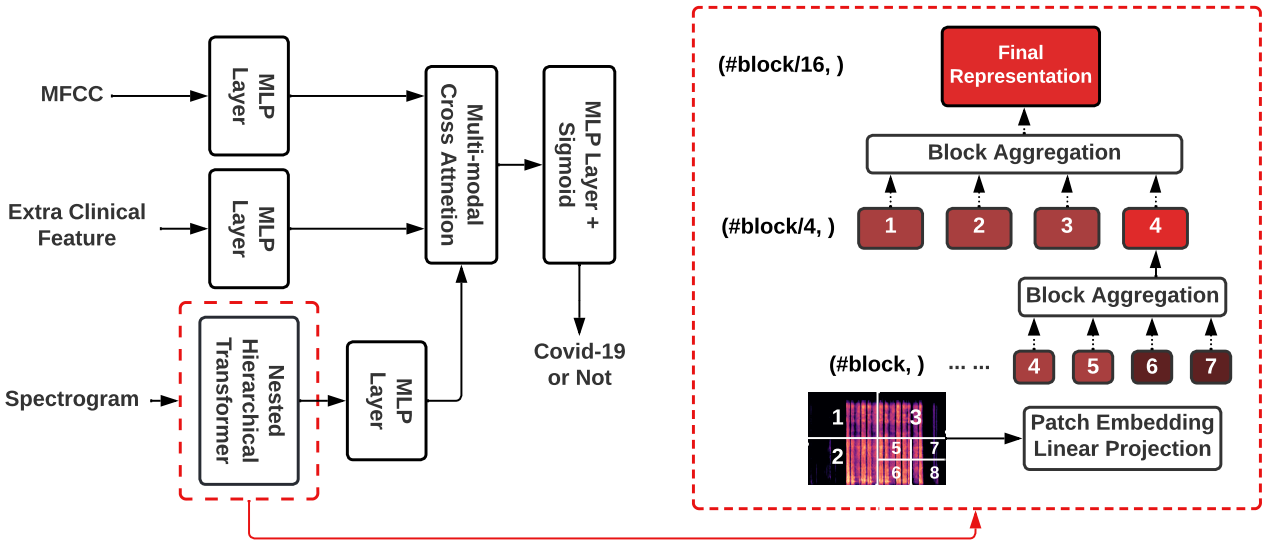

Figure 1: Illustration of Proposed Model. In our setting, the block aggregation in the red dotted box is performed between hierarchies = 3 to achieve cross-block communication on the audio spectrogram plane.

Fourier Transform with hop length of 512 successive frames and a hamming window function. To represent the characteristics of the whole audio signal, we take the average value among time frames. The other feature representation is the mel-spectrogram which visualizes the time-frequency information. For each audio input, a spectrogram image is generated by the Librosa, reshaped into (288, 432, 3), and used as input to the model.

3 METHODS

The proposed HMT is shown on Fig. 1. We have used the MFCC, extra clinical features, and audio spectrogram for each individual cough recording as the input. Note that in our setting, the MFCCs have the shape (#batch size, 39), the clinical features have the shape (#batch size, 39), and spectrograms have the shape (#batch size, 3, 224, 224). The HMT consists of two MLP sub-networks (branches) for the MFCC and extra clinical features, and one Nested Hierarchical Transformer followed by an MLP layer as the sub-network for the spectrograms. The intermediate outputs from these three sub-networks are fused, passed through the Multi-modal Cross Attention module, and fed through a final MLP layer with the Sigmoid function, which outputs the logits for COVID-19 binary classification. Below we elaborate on each component in detail.

3.1 MLP Branches

The first branch for the MFCC inputs is an MLP neural network consisting of two stacked dense (fully-connected) layers of 256 units each. Each layer is followed by a 1D batch normalization layer and a dropout layer.

The second branch for the clinical features is also an MLP neural network of two stacked dense layers of 64 units each. Similar to the first branch, each layer is followed by a 1D batch normalization layer and a dropout layer.

3.2 Nested Hierarchical Transformer

Due to the confidentiality of COVID-19 patient data, cough recording training data is limited and it is not feasible to conduct self-supervised pre-training with large amounts of unlabeled data. Teacher distillation from a large pre-trained convolutional network [22] could address this problem, however this increases the training complexity, making it unfeasible to combine other features (MFCCs, clinical features) during training. Inspired by the NesT [25], we leverage the block aggregation algorithm shown in Algorithm 1 to address the problem and significantly improve data efficiency, making our model easier to train on Coswara and COUGHVID. Specifically, the 2D spectrogram batches are patchified, linearly embedded, and concatenated with trainable position embeddings. Then the block function is applied to partition all the patches based on the hyperparameter **num_hierarchy**.

Inside each block, we stack a number of canonical transformer layers [24], notated as T_i, which are multi-head self-attention (MHSA) layers followed by a feed-forward fully connected network (FFN) with skip connections and layer normalization. Note that all the blocks inside each hierarchy share one set of parameters, which reduces the training complexity. Thereafter, we build a nested hierarchy to aggregate every four spatially connected blocks into one, as is shown in the red dotted box in Fig. 1. Since every block processes information independently via canonical transformer layers, we ensure global (cross-block) communication by applying:

1. The unblock process to obtain the full image plane,
2. Spatial operations to reduce the number of blocks (here 3x3 convolution and 3x3 max pooling are used),

3. Blocking back to obtain the next hierarchical blocks.

At the end of the block aggregation that the iteration reaches the **num_hierarchy**, a global average pooling and an MLP layer of 256 units are applied to the intermediate representations.

3.3 Multi-modal Cross Attention

After obtaining the representations from each branch, we concatenate them and pass them through the multi-modal cross attention module, which computes the cross-modal attention and self-attention. This encodes important information that let the model pays more attention to the useful modalities while filtering the trivial ones. We denote the representations from the MFCC branch, clinical feature branch, and spectrogram branch as M, C, S, respectively.

We compute the query with a learnable weight:

$$q_m = W_{mq}^T M,\ q_c = W_{cq}^T C,\ q_s = W_{sq}^T S \tag{1}$$

The key K_m and value V_m for M are computed below with a new set of learnable weights (K_c, V_c, K_s, V_s are computed using the same way):

$$\begin{aligned} K_m &= concat\{M^T W_{mk},\ C^T W_{ck},\ S^T W_{sk}\} \\ V_m &= concat\{M^T W_{mv},\ C^T W_{cv},\ S^T W_{sv}\} \end{aligned} \tag{2}$$

The cross-modal and self-attention scores are computed by the cosine similarity of the query and keys. Instead of a vanilla attention mechanism that uses dot product pairwise on each element, the cosine similarity provides a more efficient way without losing information. The interaction of different modalities answering the query is given in Equation 3:

$$\begin{aligned} \hat{m} &= softmax(CosineSimilarity(q_m,\ K_m))V_m \\ \hat{c} &= softmax(CosineSimilarity(q_c,\ K_c))V_c \\ \hat{s} &= softmax(CosineSimilarity(q_s,\ K_s))V_s \end{aligned} \tag{3}$$

The $\hat{m}, \hat{c}, \hat{s}$ are then concatenated together, added to the original m, c, s, and passed through a final MLP layer that had learned the ensembled representation. Sigmoid activation function is used to obtain the binary classification logits.

4 EXPERIMENTS AND RESULTS

4.1 Implementation Details

To align with our previous studies [3, 6], the combined Coswara and COUGHVID datasets are used to evaluate the proposed network under 5-fold cross-validation. We use both area under the ROC curve (AUC) and accuracy as metrics. We implement our proposed HMT on the PyTorch framework and PyTorch Vision library for the ViT structures and train with a Tesla 100 GPU on Google Colab Pro. During training, we choose the Adam optimizer [11] and set the hyperparameters as follows: learning rate of 0.0003, 20 epochs with early stopping, dropout rate of 0.4 for all fully connected layers, num_hierarchy of 3, and patch size of 4 for splitting spectrograms.

ALGORITHM 1: Block Aggregation Algorithm

Input: Spectrogram (#batch size, channel, height, width)
Parameter: num_hierarchy = 3, T_i = Transformer layers
Output: Intermediate representation (#block, sequence length, embedding dimension)
Function ***aggregate*** *Input*
  x ← Block(PatchEmbed(Input));
  forall *i* ← *1 to num_hierarchy* **do**
    y ← T_i(x);
    if *i* < *num_hierarchy - 1* **then**
      z ← UnBlock(y);
      z ← ConvMaxPool(z);
      z ← Block(z);
    end
  end
  h ← GlobalAvgPool(x);
  return h;
end

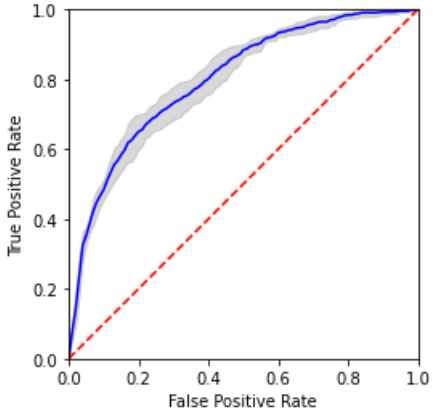

Figure 2: AUC curve for the proposed HMT model. The blue line the 5-fold cross-validation average. The shaded area is the 95% confidence region for the 5-fold cross-validation results.)

4.2 Comparison and Analysis

The experimental results for the proposed HMT and previous approaches are shown in Table 1. We leverage the source code provided by Virufy [3, 6] to re-implement their models in our setting. Each implementation is optimized using grid search to ensure a fair comparison. We find that our HMT performs better than the compared methods, attaining the best overall AUC score and accuracy with approximately +4% improvements. We also compare the vanilla ViT without block aggregation algorithm, which decreases the overall performance by 4%. Finally, we visualize the 5-fold cross-validation AUC results in Figure 2.

Table 1: 5 fold cross-validation results

Model Name	Average AUC	Average Accuracy
Inception ResNet V2 Branch [20]	0.6621	0.6576
DenseNet 121 [10]	0.6871	0.6790
Virufy Multi-branch w. CNN [3]	0.7684	0.7540
Virufy Multi-branch w. ResNet50 encoder [6]	0.7730	0.7701
Virufy Multi-branch w. ViT encoder	0.7782	0.7716
Proposed HMT	**0.8206**	**0.8132**

4.3 Ablation Studies

To measure the impact of each component in the proposed HMT model, we conducted an ablation study using the same datasets with the average AUC reported, as shown in Table 2. With only an MFCC branch that consists of MLP layers as a baseline, the performance was reduced by around 30%. Adding the extra clinical branch that consists of MLP layers slightly increased the performance by about 4%. Adding the spectrogram branch (without the transformer architecture but only the CNNs) led to around 20% performance boost, which is aligned with Virufy's previous model's performance [3] and demonstrates the useful information provided by the spectrograms. Upgrading the CNNs to the Nested Hierarchical Transformer resulted in another performance boost of 3%, which is already better than the state-of-the-art as reported in [6]. Lastly, we adding the multi-modal cross attention module in our proposed HMT model which led to the new state-of-the-art performance.

Table 2: Ablation Studies.

MFCC	Clinical	Spectrogram	NesT	Cross Attn	AUC
✓					0.53
✓	✓				0.57
✓	✓	✓			0.76
✓	✓	✓	✓		0.79
✓	✓	✓	✓	✓	**0.82**

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed HMT (Hierarchical Multi-modal Transformer), a multi-branch Transformer-based network for better performance in COVID-19 cough classification. HMT is equipped with two MLP networks for MFCCs and clinical features, a Nested Hierarchical Transformer for the audio spectrogram, and a novel multi-modal cross attention module that effectively fuses these modality inputs. Our method addresses the challenges of limited cough recording data and under-explored interactions between the audio and other modalities. Extensive experiments on two public datasets show the effectiveness of our approach, proving it as a potential solution to help alleviate the pandemic. Additionally, ablation studies further demonstrate the efficacy of each component of our HMT model.

ACKNOWLEDGMENTS

This work is fully supported by Virufy AI Research Group. Virufy is a nonprofit research organization developing artificial intelligence (AI) technology to screen for COVID-19 from cough patterns, rapidly and at no cost through use of a smartphone. Mert Pilanci is partially supported by an Army Research Office Early Career Award, and the National Science Foundation under grants ECCS-2037304 and DMS-2134248.

REFERENCES

[1] [n.d.]. World Health Organization. Coronavirus (COVID-19) Cases and Deaths. https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths

[2] Yusuf Amrulloh, Udantha Abeyratne, Vinayak Swarnkar, and Rina Triasih. 2015. Cough Sound Analysis for Pneumonia and Asthma Classification in Pediatric Population. In *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*. 127–131. https://doi.org/10.1109/ISMS.2015.41

[3] Gunvant Chaudhari, Xinyi Jiang, Ahmed Fakhry, Asriel Han, Jaclyn Xiao, Sabrina Shen, and Amil Khanzada. 2020. Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough. https://doi.org/10.48550/ARXIV.2011.13320

[4] S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[6] Ahmed Fakhry, Xinyi Jiang, Jaclyn Xiao, Gunvant Chaudhari, Asriel Han, and Amil Khanzada. 2021. Virufy: A Multi-Branch Deep Learning Network for Automated Detection of COVID-19. https://doi.org/10.48550/ARXIV.2103.01806

[7] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*. 571–575. https://doi.org/10.21437/Interspeech.2021-698

[8] Yuan Gong, Yu-An Chung, and James Glass. 2021. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021). https://doi.org/10.1109/TASLP.2021.3120633

[9] Esin Darici Haritaoglu, Nicholas Rasmussen, Daniel C. H. Tan, Jennifer Ranjani J., Jaclyn Xiao, Gunvant Chaudhari, Akanksha Rajput, Praveen Govindan, Christian Canham, Wei Chen, Minami Yamaura, Laura Gomezjurado, Aaron Broukhim, Amil Khanzada, and Mert Pilanci. 2022. Using Deep Learning with Large Aggregated Datasets for COVID-19 Classification from Cough. https://doi.org/10.48550/ARXIV.2201.01669

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. https://doi.org/10.48550/ARXIV.1608.06993

[11] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[12] J. Korpáš, J. Sadloňová, and M. Vrabec. 1996. Analysis of the Cough Sound: an Overview. *Pulmonary Pharmacology* 9, 5-6 (Oct 1996), 261–268. https://doi.org/10.1006/pulp.1996.0034

[13] Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. 2020. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical Image Analysis* 65 (2020), 101794. https://doi.org/10.1016/j.media.2020.101794

[14] Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. https://www.nature.com/articles/s41597-021-00937-4

[15] Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li. 2020. Multi-modal Attention for Speech Emotion Recognition. https://doi.org/10.48550/ARXIV.2009.04107

[16] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. 2016. A Cough-Based Algorithm for Automatic Diagnosis of Pertussis. *PLOS ONE* 11, 9 (09 2016), 1–20. https://doi.org/10.1371/journal.pone.0162128

[17] Philip J. Rosenthal. 2020. The Importance of Diagnostic Testing during a Viral Pandemic: Early Lessons from Novel Coronavirus Disease (COVID-19). *The American Journal of Tropical Medicine and Hygiene* 102, 5 (2020), 915 – 916. https://doi.org/10.4269/ajtmh.20-0216

[18] Elliot Saba. 2018. Techniques for cough sound analysis. http://hdl.handle.net/1773/43034

[19] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy. 2020. Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Proc. Interspeech 2020*. 4811–4815. https://doi.org/10.21437/Interspeech.2020-2768

[20] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. https://doi.org/10.48550/ARXIV.1602.07261

[21] Shuyun Tang, Zhaojie Luo, Guoshun Nan, Yuichiro Yoshikawa, and Ishiguro Hiroshi. 2021. Fusion with Hierarchical Graphs for Mulitmodal Emotion Recognition. https://doi.org/10.48550/ARXIV.2109.07149

[22] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers amp; distillation through attention. https://doi.org/10.48550/ARXIV.2012.12877

[23] Anusua Trivedi, Anthony Ortiz, Jocelyn Desbiens, Caleb Robinson, Marian Blazes, Sunil Gupta, Rahul Dodhia, Pavan Bhatraju, W. Conrad Liles, Aaron Lee, and Juan M. Lavista Ferres. 2020. Effective Deep Learning Approaches for Predicting COVID-19 Outcomes from Chest Computed Tomography Volumes. *medRxiv* (2020). https://doi.org/10.1101/2020.10.15.20213462 arXiv:https://www.medrxiv.org/content/early/2020/10/20/2020.10.15.20213462.full.pdf

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[25] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan O. Arik, and Tomas Pfister. 2021. Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding. https://doi.org/10.48550/ARXIV.2105.12723