

This work is on a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing [scholarworks-group@umbc.edu](mailto:scholarworks-group@umbc.edu) and telling us what having access to this work means to you and why it's important to you. Thank you.

# Neural Networks for Pulmonary Disease Diagnosis using Auditory and Demographic Information

Morteza Hosseini  
University of Maryland  
Baltimore County

Haoran Ren  
University of Maryland  
Baltimore County

Hasib-Al Rashid  
University of Maryland  
Baltimore County

Arnab Neelim Mazumder  
University of Maryland  
Baltimore County

Bharat Prakash  
University of Maryland  
Baltimore County

Tinoosh Mohsenin  
University of Maryland  
Baltimore County

## ABSTRACT

Pulmonary diseases impact millions of lives globally and annually. The recent outbreak of the pandemic of the COVID-19, a novel pulmonary infection, has more than ever brought the attention of the research community to the machine-aided diagnosis of respiratory problems. This paper is thus an effort to exploit machine learning for classification of respiratory problems and proposes a framework that employs as much correlated information (auditory and demographic information in this work) as a dataset provides to increase the sensitivity and specificity of a diagnosing system. First, we use deep convolutional neural networks (DCNNs) to process and classify a publicly released pulmonary auditory dataset, and then we take advantage of the existing demographic information within the dataset and show that the accuracy of the pulmonary classification increases by 5% when trained on the auditory information in conjunction with the demographic information. Since the demographic data can be extracted using computer vision, we suggest using another parallel DCNN to estimate the demographic information of the subject under test visioned by the processing computer. Lastly, as a proposition to bring the healthcare system to users' fingertips, we measure deployment characteristics of the auditory DCNN model onto processing components of an NVIDIA TX2 development board.

## CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Neural networks*.

## KEYWORDS

respiratory sounds dataset, demographic feature extraction, deep convolutional neural networks, embedded devices, early-stage diagnosis, public health

## 1 INTRODUCTION

In 2016, pulmonary diseases were among the top 10 causes of death: ranked 1 for low-income and ranked 5 for high-income countries [1]. Recently, with the outbreak of the COVID-19 as a novel pulmonary infection, a tremendous amount of attention has been directed to control the pandemic crisis about which extreme measures are taken by countries to diagnose the infected patients. Measures such as extensive testing and early-stage diagnosis help to locate and contain the infection, and are reportedly the most effective preventive actions to control the contagion in a pandemic.

Pulmonary problems encompass a wide range of chronic and infectious diseases, and because of the common organ, lung, that they affect, they develop respiratory symptoms whose auditory signals recorded from various medical devices are among the first to be scrutinized by a medical expert. As an example, COVID-19 develops symptoms such as dry cough, fever, fatigue, dyspnea, and shortness of breath that vary in severity at different stages of the disease progression, and correlate with certain ethnicity, gender, and age groups differently [10]. More than 70% of the confirmed COVID-19 patients have reported fever in tandem with a dry cough [23]. Meanwhile, clinical case records indicate that the young population is less likely to develop COVID-19 relevant symptoms, contrary to the elderly that is the most vulnerable group [9].

Traditionally, when someone feels symptoms, they either call a doctor or have themselves seen/scrutinized by medical experts at walk-in clinics, where extensive use of vital signs, visual and auditory information are applied to make diagnostic decisions. Such practice during a pandemic or in remote locations is unsuitable/impractical as a result of the limited capacity of existing facilities and human resources at health centers, and, ironically, can expedite spreading the infection. On the other hand, calls are made by governments/organizations during the pandemic for people to stay at home that, by itself, has caused a state of confusion and has made another barrier. Thus, early-stage and clinic independent machine assistance is critical for the initial diagnosis of the disease and/or for evaluating/assessing its severity.

Our goal in this research is to allow machine learning algorithms running on general computing processors (e.g., those in cell-phones and tablets) to assess patients similar to what doctors do at triage and telemedicine, using passively recorded audio and/or video and self-declared information, to bring proactive healthcare to users' fingertips and to estimate the urgency/necessity of whether they need to attend clinics and have themselves further examined with the use of more specialized test-kits or facilities. Our vision is to

Permission to make digital or hard copies of all or part 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 components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*epiDAMIK 2020, Aug 24, 2020, San Diego, CA*

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-xxxx-XXXX-X... \$15.00

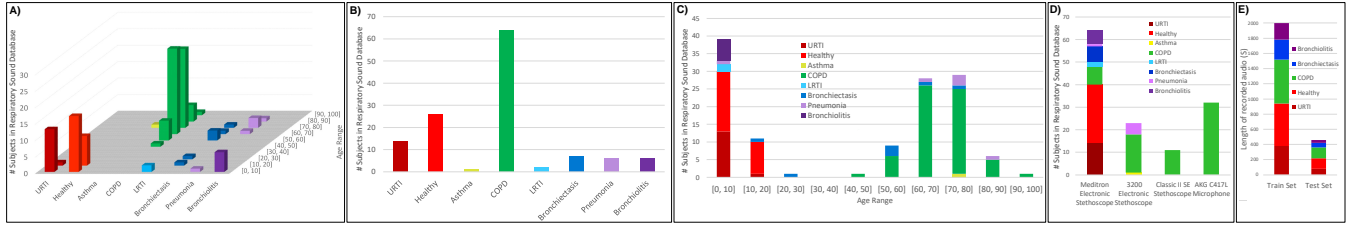
<https://doi.org/xx.xxxx/xxxxxxxxxx.xxxxxx>

provide a detection framework that can provide early detection for anyone and anywhere. We develop our work on a publicly released respiratory sound database that includes both auditory and demographic information recorded from 126 subjects covering 7 pulmonary diseases including healthy condition, and with two sets of annotations. More specifically, since the respiratory sound dataset includes the two types of information per patient, we examine how the lack/existence of the demographic information impacts the total accuracy of the model. For further compilation, we develop another deep neural network that estimates demographic information including the age and the gender of the captured images and to correlate them with the auditory signals recorded from the subject under test in order to assess a higher sensitivity and specificity rate of diagnosis. The main contributions of this work include:

- ## 2 RELATED WORK

learning (ML) tools on a limited unpublished dataset can distinguish solely between coughs from COVID-19 patients and those who are healthy or with upper-respiratory coughs with high accuracy of 96.8%. [?] introduces End to End convolutional neural networks for cough and dyspnea detection. Authors in [3] used both DCNNs and recurrent neural networks (RNNs) to classify cough sound that they collected using chest-mounted sensors. Authors in [13] used deep learning to detect sleep apnea. Classification of heart sound into normal and abnormal classes was conducted in [19] using DCNNs. Authors in [4] and [15] used DCNNs and RNNs to classify lung sounds respectively. Most of these works report high levels of accuracy on unpublished datasets that are accessible by the research community. The 2017 International Conference on Biomedical Health Informatics (ICBHI) [17] issued a benchmark dataset of respiratory sound to facilitate researching on respiratory sound classification. Since then, researchers proposed various algorithms [5, 11, 14, 16] using different deep learning techniques to classify respiratory cycle anomalies such as the precise locations of wheezes and crackles within the cycle of each respiratory sound recording. In [12] authors showed innovativeness by proposing a digital stethoscope to provide an immediate diagnosis of respiratory diseases. They developed a modified bi-ResNet architecture using STFT and wavelet feature extraction. Log quantized deep CNN-RNN based model for respiratory sound classification was proposed in [2] for memory limited wearable devices.

The framework, depicted in Fig. 1, leverages audio/video to extract necessary and medically relevant information and combines the extracted features with other inserted/self-declared patient data. The audio processing incorporates an ML approach such as a DCNN that extracts symptomatic features like crackles and wheezes of lung sounds from a given window of recorded sound of a subject under test. At the video processing path, the captured RGB images are



**Figure 2: Statistics of the respiratory sound database that contains auditory samples from 126 patients, A) a 2D histogram of 7 pulmonary classes with respect to 10 age groups, B) Break-down of the pulmonary classes in the dataset, C) Break-down of each pulmonary class with respect to age groups, D) Break-down of each pulmonary class with respect to the four recording medical devices, E) Our selection of 52 and 11 individual subjects for train and test datasets respectively that cover 5 pulmonary classes recorded with Meditron Electronic Stethoscope.**

given to a ResNet-34 DCNN to process and estimate the user’s other demographic and symptomatic features such as age and gender. The extracted audio/video features along with the other relevant inserted data are concatenated towards the final layers and with the addition of a few more neural network layers or an ensemble of classifiers in the last layer, a probability vector of diagnosis for the user under test is reported. Both audio/video data can extend the scope of the clinical-reported symptoms to more diverse features that may be invisible to a human’s perception. For example, when listened by a trained machine, the features extracted from the sound of a patient’s cough can include more useful features beyond terms like “dry” or “productive” that are commonly reported in clinical case records.

### 3.1 Datasets

**3.1.1 Respiratory Sound Database.** For the auditory dataset, we used a public respiratory sound database [18], which includes 920 recordings acquired from 126 participants annotated with 8 types of respiratory conditions including URTI, Healthy, Asthma, COPD, LRTI, Bronchiectasis, Pneumonia, and Bronchiolitis. The recordings were collected using four types of medical equipment including AKG C417L Microphone, 3M Littmann Classic II SE Stethoscope, 3M Littmann 3200 Electronic Stethoscope, and Welch Allyn Meditron Master Elite Electronic Stethoscope. The duration of each recording range from 10 to 90 seconds mostly dominated with 20s samples. Fig. 2-A, B, and C plot the distribution of the subjects with respect to their diagnosed disease and the age groups they impact, and Fig. 2-D shows the contribution of each of the 4 medical devices for recordings from participants. Among the four recording devices, the Meditron Electronic Stethoscope is the only device that encompasses the 8 pulmonary conditions except for the Asthma, and is used for 63 out of the 126 participants. The recordings from the other 3 devices are majorly taken from COPD-diagnosed participants. By eliminating Asthma, Pneumonia, and LRTI that have little or no samples within the Meditron recordings, we extracted a random subset encompassing all 63 participants and split it into a semi-balanced train and a test set of 52 and 11 participants that include 5 types of pulmonary classes. Fig. 2-E shows a plot of the selected train/test dataset based on the total duration of each class. The database is meanwhile provided with demographic information of the 126 participants and another annotation that marks

**Table 1: Selected semi-balanced Dataset out of the respiratory sound database. Meditron recordings from 61 patients that include 5 pulmonary classes were selected and the 20s sounds were chopped into overlapping frames of 5s. The total dataset of frames includes 1968 samples that were split into mutually-subject-exclusive between train (81%) and test (19%) subsets.**

Selected Dataset		URTI	Healthy	COPD	Bronchiect.	Bronchiol.	Total
Train Set	# Subject	12	24	6	5	5	52
	Duration (S)	380	560	580	260	220	2000
	# Respiratory Cycles	207	257	406	88	141	1099
	# Augmented Frames	304	448	464	208	176	1600
Test Set	# Subject	2	4	2	2	1	9
	Duration (S)	80	140	140	60	40	460
	# Respiratory Cycles	26	48	119	17	16	226
	# Augmented Frames	64	112	112	48	32	368

begin/end of respiratory cycles and the precise locations of events of crackles and wheezes per recording. Based on the second annotation, we counted the total number of respiratory cycles to estimate the slowest and average respiratory cycles within the dataset and to decide on a window size to cut the recordings into smaller frames. Table 1 summarized the number of subjects, duration of recordings, and the number of respiratory cycles per pulmonary class within both train and test datasets.

**3.1.2 Face Images Database.** UTKFace dataset [22] is a large-scale dataset consisting of over 20,000 face images with annotations of age (ranging from 0 to 116 years old) gender, and ethnicity. In [8], the UTKFace dataset is trained on a ResNet-34 [6], and we reproduced the results of training over ResNet-34, and report the accuracy it gives for precise age as well as for the age group of a random split of 20% test data.

### 3.2 Data Pre-processing and Augmentation

For data augmentation of the respiratory sound database, every recorded audio sample is cut into frames with a duration of 5s and with a stride of 1s, which means every two adjacent frames overlap a duration of 4s, and every 20s recorded sample results in 16 5s frames. Therefore the total 2000 seconds of the training dataset generates 1600 frames, and the total 460s testing data generates 368 frames of 5s samples. The choice of the 5s frames is inferred empirically by experiencing frames ranging from 1s to 10s.

**Table 2: The classification accuracy and the model complexity of a ResNet-34 DCNN that extracts demographic information on the UTKface test dataset.**

DCNN characteristics			Test accuracy			
Model	#params	FLOPs	Age±0	Age±5	Age±10	Gender
ResNet-34	21M	3.6B	19.6%	65.5%	87.1%	90.3%

**Table 3: Respiratory sound classification accuracy and model complexity with and without taking the demographic information into account.**

DCNN characteristics			Sensitivity					Accuracy
Model	#params	FLOPs	URTI	Healthy	COPD	Bronchiec.	Bronchiol.	Test
EnvNet-like on Sound w/o Demographic Info	320k	0.194B	21%	68%	96%	88%	4%	78%
EnvNet-like on Sound with Demographic Info	320k	0.194B	16%	72%	100%	88%	15%	83%

For the data augmentation of the UTKface images, we use common image augmentation techniques such as flipping, shifting and resizing the images within the dataset.

## 4 EXPERIMENTAL SETUP

We used a ResNet-34 DCNN [6] for the UTKface RGB images of size 200×200, and an EnvNet-like [21] DCNN for the respiratory sound frames of size 1×220500. For the EnvNet-like DCNN, the input from the audio recordings is a one-dimensional vector where the size depends on the window selected for the framework. To best utilize the one-dimensional input, we use two one dimensional convolution layers to extract relevant features with a follow up of non-overlapping max-pooling operation to downsample the feature map. The subsequent layers include two-dimensional convolutional layers with max-pooling layers in between for efficient classification of the diseases. Finally, the fully connected layers summarize the required feature information and feed it to the extended model to generate a generalized output that classifies 5 types of pulmonary conditions as in our extracted dataset.

### 4.1 Demographic Classification

#### 4.1.1 ResNet-34 for UTKFace.

The classification accuracy of age and gender estimation of ResNet-34 is reported in Table 2. Although the DCNN model does not precisely classify the age within the test dataset, it is able to classify the gender and estimate the age groups when the range of the groups expands. This is in correspondence to combining the auditory data with the age group, as conducted and reported in the next subsection where we combine the auditory information with the age group of the subjects, rather than the precise age of each participant.

### 4.2 Auditory Classification

#### 4.2.1 EnvNet for Respiratory Sound and Demographic Information.

We first conduct a set of experiments to explore the best DCNN configuration based on the EnvNet DCNN that achieves the highest accuracy. Then, we combined the audio dataset with the age groups they are recorded from as depicted in Fig. 1. Table 3 compares the

**Table 4: Deploying the EnvNet model to commercial off-the-shelf devices including a dual-core Denver CPU, a quad-core ARM A57 CPU, and a combination of ARM CPU + Pascal GPU from the NVIDIA TX2 board.**

Configuration	CPU Freq. (MHz)	GPU Freq. (MHz)	Power (mW)	Latency (S)	Performance (GFLOP/S)	Energy (J)	Energy Eff (GFLOP/W)
Denver CPU	345	-	881	10.0	0.019	8.81	0.021
	2035	-	3170	0.9	0.215	2.85	0.068
ARM A57 CPU	345	-	1168	3.7	0.052	4.32	0.045
	2035	-	4425	0.6	0.322	2.66	0.073
TX2 CPU+GPU	2035	1300.5	9106	0.1	1.935	0.91	0.210

two sets of experiments, indicating that the COPD and healthy conditions are diagnosed with higher accuracy and resulting in a total test accuracy increase by 5% when the demographic information is taken into account.

## 5 COMMERCIAL OFF-THE-SHELF DEVICE DEPLOYMENT

The framework is intended to be flexibly deployable for general-purpose devices where the developed ML models trained on the framework can be deployed onto processing machines that may range from front-end edge devices to back-end computer servers. Trading off between the computation complexity and the classification accuracy, trained ML models can be deployed to edge devices (e.g. a cell-phone, tablet) to process data locally if the information privacy is a concern, or otherwise to the cloud servers that can process data with more elaborate up-to-date models that yield higher quality metrics.

All of the DCNN models are attributed to at least two hardware-level characteristics: the model size and the number of computer operations per inference, both of which are upper-bounded by the platform resources that they deploy to, or by the inference deadline. When putting all the components of the framework together, both the hardware resource constraints and the diagnosis latency should meet the application goals. Having set the batch-size equal to 1, the trained models obtained from the previous Section are deployed on two mobile CPUs including Denver (dual-core) and ARM-Cortex A57 (quad-core) as well as an embedded CPU+GPU implementation with different frequency settings. All of the settings were performed on the TX2 development board that provides precise on-board power measurement. Table 4 summarizes the implementation, indicating that, provided a 5s frame of recording to the memory, the least power dissipating implementation (Denver with a low frequency) takes 10 seconds to classify one frame whereas the most energy-efficient implementation (CPU+GPU) dissipates approximately 10× more power to classify the same frame within 0.1 seconds.

## 6 COMPARISON

The most related work to ours that has developed a DCNN on the same respiratory sound database is the work in [20] that reports an overall accuracy of 97%. The main difference between the two works is that our model uses additional information in tandem with the audio data and proposes a framework that suggests combining as much existing correlated information within the dataset as possible

**Table 5: Comparison to the Related Work**

Related Work	#Augmented Audio Samples within the Dataset								Test Accuracy
	URTI	Healthy	COPD	Bronchiec.	Bronchiol.	Asthma	LRTI	Pneum.	
[20]	403	455	10,205	377	-	13	26	481	97%
This Work	370	560	567	256	208	-	-	-	83%

to rectify and increase the diagnosis accuracy. The other difference is that our selected dataset is semi-balanced among 5 classes of respiratory sounds recorded from one unique medical device that has been indistinguishably utilized for 61 subjects diagnosed with 7 out of 8 classes within the database, whereas the dataset selection in [20] is excessively dominated with COPD recordings, a major portion of which, as depicted in Fig. 2-D, are recorded by two medical devices that have been used to merely sample from COPD-diagnosed participants. Table 5 provides a comparison and a summary of the total number of augmented samples per class within the two works.

## 7 CONCLUSION

In an attempt to exploit machine learning algorithms to classify respiratory problems, we proposed a framework that employs as much correlated information as a dataset provides and showed that with combining both auditory and demographic information for a selection of reasonably balanced dataset out of a publicly released respiratory sound database the diagnosis accuracy of the trained deep convolutional neural networks (DCNNs) increases by 5%. Since the demographic data can be extracted and estimated using computer vision, we suggest using another DCNN that works in parallel to the auditory signal processing DCNN to estimate the demographic information of the subject under test. Lastly, we deploy our DCNN models on a dual-core Denver CPU, a quad-core ARM Cortex A57, and a heterogeneous implementation of CPU+GPU from the NVIDIA TX2 development board to measure hardware characteristics when deploying the model to an embedded device.

## REFERENCES

- [1] 2018 (accessed June, 2020). The Top 10 Causes of Death. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>
- [2] Jyotibha Acharya and Arindam Basu. 2020. Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning. *IEEE transactions on biomedical circuits and systems* 14, 3 (2020), 535–544.
- [3] Justice Amoh and Kofi Odame. 2016. Deep neural networks for identifying cough sounds. *IEEE transactions on biomedical circuits and systems* 10, 5 (2016), 1003–1011.
- [4] Murat Aykanat, Özkan Kılıç, Bahar Kurt, and Sevgi Saryal. 2017. Classification of lung sounds using convolutional neural networks. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 65.
- [5] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj. 2020. Convolutional neural networks based efficient approach for classification of lung diseases. *Health Information Science and Systems* 8, 1 (2020), 4.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [7] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275* (2020).
- [8] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913* (2019).
- [9] Ping-Ing Lee, Ya-Li Hu, Po-Yen Chen, Yhu-Chering Huang, and Po-Ren Hsueh. 2020. Are children less susceptible to COVID-19? *Journal of Microbiology, Immunology, and Infection* (2020).
- [10] Kai Liu, Ying Chen, Ruzheng Lin, and Kunyuan Han. 2020. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection* (2020).
- [11] Renyu Liu, Shengsheng Cai, Kexin Zhang, and Nan Hu. 2019. Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network. In *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. IEEE, 298–303.
- [12] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang. 2019. LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.
- [13] Hiroshi Nakano, Tomokazu Furukawa, and Takeshi Tanigawa. 2019. Tracheal sound analysis using a deep neural network to detect sleep apnea. *Journal of Clinical Sleep Medicine* 15, 8 (2019), 1125–1133.
- [14] Diego Perna. 2018. Convolutional neural networks learning from respiratory data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2109–2113.
- [15] Diego Perna and Andrea Tagarelli. 2019. Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 50–55.
- [16] Lam Pham, Ian McLoughlin, Huy Phan, Minh Tran, Truc Nguyen, and Ramaswamy Palaniappan. 2020. Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases. *arXiv preprint arXiv:2002.03894* (2020).
- [17] BM Rocha, D Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, A Oliveira, C Jácume, A Marques, et al. 2017. A respiratory sound database for the development of automated classification. In *International Conference on Biomedical and Health Informatics*. Springer, 33–37.
- [18] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement* 40, 3 (2019), 035001.
- [19] Heechang Ryu, Jinkyoo Park, and Hayong Shin. 2016. Classification of heart sound recordings using convolution neural network. In *2016 Computing in Cardiology Conference (CinC)*. IEEE, 1153–1156.
- [20] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. 2019. Lung Disease Classification using Deep Convolutional Neural Network. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 732–735.
- [21] Yuji Tokozume and Tatsuya Harada. 2017. Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2721–2725.
- [22] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5810–5818.
- [23] Xianxian Zhao, Bili Zhang, Pan Li, Chaoqun Ma, Jiawei Gu, Pan Hou, Zhifu Guo, Hong Wu, and Yuan Bai. 2020. Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. *MedRxiv* (2020).