DiCOVA-Net: Diagnosing COVID-19 using Acoustics based on Deep Residual Network for the DiCOVA Challenge 2021

Jiangeng Chang[1], *Shaoze Cui*[1,2], *Mengling Feng*[1]

[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[2]Institute of Information and Decision Technology, Dalian University of Technology, Dalian, China
changjiangeng1@gmail.com, csz2016@mail.dlut.edu.cn

Abstract

COVID-19 has been proved to impair the normal function of the respiratory system which makes acoustic detection to be a selective method to detect the virus. In this paper, we propose a deep residual network-based method for the identification of COVID-19 infected people using cough sound data. Since there are far more healthy people than patients, this is a classification problem under imbalanced data. To improve the model's ability to recognize minority class (infected people), we introduce data enhancement and cost-sensitive methods into our model. Besides, considering the particularity of this task, we use fine-tuning technology to adjust the pre-training ResNet50. Further, to improve the generalization ability of the model, we use ensemble learning to integrate the results of the base classifiers which are generated by using different random seeds to form the final prediction result. To evaluate the performance of our proposed method, we have conducted experiments based on the DiCOVA challenge dataset. The results show that our method has achieved 85.43% in AUC. Comparing with other participated methods, our method is competitive.

Index Terms: COVID-19 diagnosis, imbalanced data, data augmentation, focal loss, transfer learning, random seeds

1. Introduction

Since the COVID-19 outbreak in 2020, countries around the world have been shrouded in the haze of the epidemic. Until now, people in many countries were at risk of infection. In the process of combating the epidemic, countries are actively working to avoid large-scale infection in their countries. Countries such as China, the United States and the United Kingdom have all developed vaccines against COVID-19 and called on people to get vaccinated. But the vaccine's protection period is limited, and many people are reluctant to get vaccinated. Therefore, effective testing is still needed to detect COVID-19 infected persons.

In the early days of the outbreak, the primary tests for COVID-19 were X-rays and computed tomography (CT). When medical images of the lungs show symptoms such as shadows, the patient is more likely to be infected with COVID-19 [1]. At the beginning of the epidemic, this detection method achieved relatively good results, and played an important role in the further spread of the epidemic. With the development of research on COVID-19, swab test is considered to be a reliable method for detecting COVID-19 infection. At present, most countries use throat swab or nose swab test. In addition, some studies have shown that saliva may also be an effective test method [2].

In addition to the testing methods for infected persons mentioned above, other testing methods are also being studied. Numerous studies have shown that COVID-19 mainly damages lung tissue, and many patients end up dying of respiratory failure [3]. Therefore, it is clear that COVID-19 has a significant effect on the respiratory system, which may lead to changes in the acoustic characteristics of the infected person. With this in mind, in this paper, we will explore a non-invasive method for detecting COVID-19 infected persons, which uses cough voice data to construct a classification model.

In recent years, machine learning and deep learning have made remarkable achievements in many fields such as image processing [4], speech recognition [5] and text analysis [6]. In the wake of the COVID-19 outbreak, researchers are also trying to apply artificial intelligence technology to the fight against the epidemic. For example, Tuli et al. [7] designed a model based on cloud computing and machine learning to predict the growth and trend of COVID-19 pandemic. Yan et al. [8] developed an XGBoost machine learning-based model that could predict the mortality rates of patients more than 10 days. Ardakani et al. [9] applied the deep learning technique to manage COVID-19 in routine clinical practice using CT images. Inspired by these studies, this paper discusses the recognition of COVID-19 infected person from the perspective of speech recognition.

In order to enhance the recognition ability of the model, we improved the model mainly from data augmentation, transfer learning, cost-sensitive learning, ensemble learning. To overcome the limitation in the diversity of the model, we leverage a novel randomness method to further improve the model's diversity and the ensemble results outperform the previous methods.

The remaining of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents the experiment and results. In the last section, this study is summarized.

2. Related Work

In essence, the task of identifying COVID-19 infected people using acoustic data in this paper falls under the category of speech recognition. In the early stage of speech recognition research, scholars used the acoustic model developed based on Gaussian Mixture Models (GMMs), and achieved certain recognition effect. However, with the proposal of deep learning method and the deepening of related researches, scholars found that the model based on deep neural networks (DNNs) is better than GMMs [10]. Different from the traditional GMMS-based model, the data format processed by deep learning is mostly image data. Therefore, it is necessary to convert acoustic waveform data into spectrogram data [11].

In the practice of medical field, the problem of imbalanced data is common. Through analysis, we found that the recognition of COVID-19 infected people also has the problem of imbalanced data. The existence of imbalanced data will cause problems in the training of the model, that is, the model will pay more attention to the recognition of majority samples, which is not the desired result of medical staff or managers. To deal with the imbalance, researchers have come up with a variety of approaches. For instance, Jiang et al. [12] used the data augmentation technique to add noise to the data, and the experimental results showed that the performance of the speaker recognition model could be improved by adding noise. Moreover, in addition to processing the data to deal with the problem of imbalanced data, the loss function of the model can also be modified in a cost-sensitive approach to achieve the purpose of modifying the model. For example, Lin et al. [13] proposed a new loss function called Focal Loss in the dense object detection task, which was developed on the basis of the Cross-Entropy (CE) loss function. Experiments in the paper verify that Focal Loss helps to improve the detection ability of minority samples when it comes to imbalanced data.

In the field of image processing and recognition, there are many famous deep learning models such as AlexNet [14], VGG [15] and ResNet [16]. Among these various deep learning models, ResNet is now the most widely used and the most prominent one. In this paper, we focus mainly on the ResNet50 network. With the development of deep learning technology, scholars have found that it is very inconvenient to retrain the whole deep neural network every time they encounter a new task, which requires a lot of time and computing resources. Therefore, the researchers came up with a technique called “pre-training” [17]. An already trained deep learning model can be loaded through pre-training, and then the network weight can be partially adjusted (fine-tuning) for specific tasks in the domain, so as to effectively solve new problems. The ”pre-train then tune” paradigm is also the core idea of transfer learning [18].

In addition to deep learning, another very popular area of research in machine learning is ensemble learning. The purpose of ensemble learning is to enhance the generalization ability and prediction accuracy of the entire model by combining the outputs of multiple base learners. Some studies have found that combining the outputs of multiple deep neural networks through ensemble learning method is helpful to improve the accuracy of prediction [19, 20].

However, the previous only focuses on the prior knowledge to increase the model’s diversity by changing the training data space or changing the model’s structure. In this paper, we focus on the importance of randomness on the model’s diversity. Randomness is ubiquitous in the real world. Though randomness is uncontrollable, they also increase the object’s diversity and sometimes can give better results than the experiments with a controllable setting [21, 22].

In this paper, we propose a model based on deep residual network for COVID-19 infection recognition, which integrates the technologies of imbalanced data processing, transfer learning and ensemble learning.

3. The Proposed DiCOVA-Net Method

In this section, we describe the proposed DiCOVA-Net method as shown in Figure 1. First of all, the input acoustic data needs to be transformed into spectrogram data. After that, the data is divided into multiple data subsets by means of cross validation, and Gaussian noise is added to the minority samples to form new minority samples on each subset. The new data set obtained through data augmentation is fed into the ResNet50 pretraining model for fine-tuning, where focal loss is used as the loss function. Finally, ensemble learning methods are used to combine the outputs of the various models which are generated by different random seeds when forming the final prediction results.

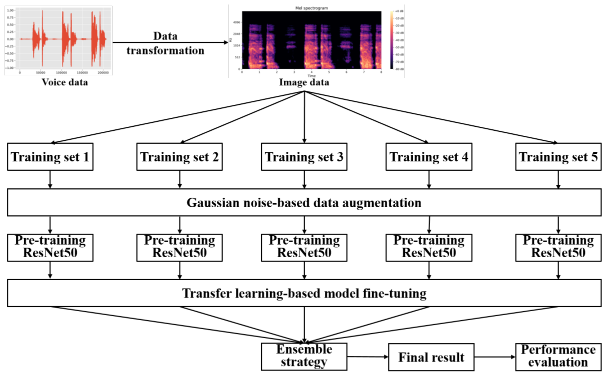

Figure 1: *The proposed DiCOVA-Net method.*

3.1. Gussian Noise-based Data Augmentation

In the field of image processing and recognition, data augmentation is a very popular technology, which is used to expand data sets. In particular, data augmentation can not only increase the diversity and number of training data, but also reduce the domain mismatch between the enrolled and test set data [12]. The data augmentation methods commonly used include flip, rotation, scale, crop, translation, etc., while what we adopt is to add Gaussian noise. Gaussian noise has a mean of zero and a standard deviation of one and can be generated as needed using a pseudorandom number generator. In this paper, Gaussian noise was added to the data of minority samples (namely COVID-19 infected persons) to generate some new synthetic minority samples, which is very beneficial for DNNs and helps to reduce the incidence of overfitting, because DNNs have a large number of parameters.

3.2. Deep Residual Network with Transfer Learning

Residual neural network [16] is a classical method which achieve the best performance in the ILSVRC 2015 classification task. Residual neural network is a new convolutional neural network structure. With the success of the VGG network, the depth of the neural network has attracted more and more attention, but it is accompanied by the problem of gradient disappearance that is difficult to solve. The residual neural network uses the residual block to solve this problem very well. The residual block can be calculated as Eq.(1):

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

where y_l represents the label of sample l, and $h(x_l)$ is the prediction that neural network output at the step l. The $F(x_l, W_l)$ is the residual result of the step l.

In the residual neural network, there are two hypotheses: (1) $h(x)$ is the direct mapping. (2) $F(x, W)$ is the direct mapping. Then, the loss ϵ gradient between the beginning layer and layer l can calculated as Eq.(2):

$$\frac{\partial \epsilon}{x_0} = \frac{\partial \epsilon}{x_l} + \alpha \tag{2}$$

where α can be presented as Eg.(3):

$$\alpha = \frac{\partial \epsilon}{x_l} \frac{\partial F}{x_0} \tag{3}$$

where F can be presented as Eg.(4):

$$F = \sum_{i=0}^{l} f(x_i, W_i) \tag{4}$$

Throughout the training process, α cannot always be -1, which means that the problem of gradient vanishing will not occur in the residual network

Transfer learning [23] is innovated to reduce the human annotation cost by employing the model which is trained on annotated source domain data to predict on the unannotated target domain data. And transfer learning is proved to be efficient in the few data classification [24]. We follow this insight, firstly we pre-train the ResNet50 on the ImageNet [25] which has 12 subtrees with 5247 synsets and 3.2 million images in total. Then we fine-tune the pre-training neural network structure on the DiCOVA dataset. With the help of pre-training, we find the model will reach the best point in only few epoches.

3.3. Focal Loss

In the model of machine learning and deep learning, the loss function is used to measure the degree of difference between the predicted value and the real value. If the model prediction is not accurate, the value of the loss function will be larger. In general, the better the loss function is designed, the better the performance of the model will be. The loss function plays the role of "supervisor" in DNNs, which guides the model training to move towards the direction of reducing the loss function, so as to find the network parameter combination that minimizes the loss function. At present, cross entropy (CE) is often used as the loss function when DNNs are trained, which is calculated as Eq.(5). Compared with the classification error rate and the mean square error (MSE), CE can better capture the differences between different models. Moreover, CE has the property of convex function, so it can find the global optimal value when taking the derivative.

$$L_{CE} = -\sum_{i=1}^{m} y_i \cdot log(p_i) \tag{5}$$

where y_i represents the label of sample i, and p_i is the probability that sample i is predicted to be a positive class.

Cross entropy can achieve good results when the number of samples in each class is not very different, but it is no longer effective when the data is unbalanced. Aiming at the problem of unbalanced data, Lin et al. [13] proposed a new loss function called focal loss which is calculated as Eq.(6).

$$L_{FL} = -\sum_{i=1}^{m} \alpha_i (1 - p_i)^{\gamma} log(p_i) \tag{6}$$

where γ ($\gamma \geq 0$) is the focusing parameter which is used to adjust the weight of difficult samples and easy samples, and the $(1-p_i)^{\gamma}$ is called the modulating factor. In addition, α_i is used to adjust the weight of positive and negative samples.

From Eq.(6), we can get two important properties of the focal loss: (1) When the value of $p_i \to 1$, $(1 - p_i)^{\gamma} \to 0$, indicating that the prediction of the model is accurate and the sample is an easy sample, so the contribution of the sample to the loss is small. (2) The focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. When $\gamma = 0$, focal loss equivalent to CE, and as γ increases, the weight of easy samples will be further reduced. Therefore, using Eq.(6) as the loss function of the model in this paper, on the one hand, the problem of imbalanced data can be considered, and on the other hand, the difficulty of sample discrimination can be taken into account.

3.4. Ensemble Learning

An ensemble framework aggregates the predictions of multiple base models to get better prediction. Formally, suppose that we have K base models with predictions $\boldsymbol{y}_{t+1} = [y_{t+1}^1, \cdots, y_{t+1}^K]$, ensemble is an aggregation function $y_{t+1} = g(\boldsymbol{y}_{t+1}; \boldsymbol{\Phi})$, which has various implementations such as *voting*, *averaging*, and *stacking* [26]. We used the averaging method to ensemble different the same model trained on different distributed feature space.

3.5. Randomness

In this work, we proposed a novel and robust method to achieve high performance in the imbalanced dataset. Randomness is ubiquitous in deep learning which adds uncertainty to deep learning. However, the researchers have explored randomness in the weight initial for deep learning models which helps to generate realized images without training [21]. And other researchers further explored the randomness in layers chosen and models chosen for the ensemble methods [22]. This method achieves state-of-art results on different domain datasets. However, all the previous methods have not tried the randomness in the imbalance dataset. With our knowledge, we can find that neural network pays more attention to the data entered at the beginning. And the model cannot learn the full data with the limit of memory, the input data will be divide into different batches. So the different data input will make a range in the same model's performance. But for the large data, the range will very be small. However, through the experiment, we find the different data input in the few imbalance data will make a huge range of the model's performance. So we used randomness to try different data input to search for the best data input order. With the infinite search space is expensive, we suggested trying four or five different random seeds and ensemble results together. The novel method can be summarized in $\boldsymbol{f}_{t+1} = [f_{t+1}^1, \cdots, f_{t+1}^K]$, the $[f_{t+1}^1 = W(x^r), r \in random(500, 2000)$. The $[f_{t+1}^1, \cdots, f_{t+1}^K]$ are all based on the same model but with different data input order. The experimental result shows our method improves the model's accuracy and robustness.

4. Experiments and Results

4.1. Datasets

We evaluated our proposed method in the DiCOVA 2021 challenge [27]. This challenge provides a dataset of sound recordings collected from COVID-19 infected and non-COVID-19 individuals for a two-class classification. This data set contains a total of 1040 samples, including 965 non-COVID-19 samples and 75 COVID-19 infected samples, with an imbalanced ratio of 13:1, which proves that this is an extremely imbalanced dataset. The average duration of recordings across subjects is 4.72 (standard error (*S.E.*) $\pm$ 0.07) sec. In addition, the organizers of this challenge adopt a blind test set containing 233 samples, so the results obtained in this paper on the test set are given by the organizers of the challenge.

Table 1: *Main parameters*

Parameter name	Settings
Sampling Rate	2048
Audio length	4 sec
FFT window	2048
Frame shift	512
Epoch	20
Batch size	16
Learning rate	0.0002
Focal loss α	0.25
Focal loss γ	2
Random seeds (ensemble)	1001,500,1500,2000

4.2. Experimental Settings

To be objective, the challenge organizers have already split the dataset into train set and validation set. The train set contains 822 data which includes 772 non-COVID-19 and 50 COVID-19. The validation set contains 218 data which includes 193 non-COVID-19 and 25 COVID-19. The organizers also generate five-fold data set which helps participants to gain a more generalized and diverse model.

We use the librosa[1] to transform the audio data into mel-spectrogram. In the guassian augmentation progress, we use the skimage[2] and set the random seed between 0 and 5. We list other parameters in the Table 1.

In the ResNet50, we use the *focal loss* as the loss function and use the *adam* as the optimization method. In model selection, we use the area under the curve (AUC) as an evaluation indicator in each folds.

In the randomness experiment, we only change the random seeds at the beginning of the pipeline.

4.3. Results and Discussions

For this challenge, the organizers present the performance of three baseline methods: Random Forest (RF), Multi-layer Perceptron (MLP), and Logistic Regression (LR). These three methods could not receive audio data, so the organizers use the mel-frequency cepstral coefficients (MFCC) and the delta and delta-delta coefficients methods to extract features. The performance of the three baseline methods is shown in Table 2 [27]. According to the results in Table 2, LR has the worst performance, because compared with RF and MLP, LR has the worst nonlinear fitting ability. In addition, the difference between RF's performance on the validation set and the test set is larger than that of MLP, so MLP is more robust.

Table 2: *Baseline performance*

Methods	Validation AUC	Test AUC
RF	70.63	67.59
MLP	68.81	69.91
LR	66.97	61.97

The experiment result of proposed method is shown in Table 3. We compare the performance of fine-tuning model "Original" with two different loss functions and two different data augmentation methods. The two different loss functions respectively "cross entropy" (CE) loss and "focal loss" (FL). The two different data augmentation methods are "simple duplication" (Dul) and "Gaussian noise" (Gua). The "Ensemble" is the performance by integrating four models with different random seeds.

[1] https://github.com/librosa/librosa

[2] https://github.com/scikit-image/scikit-image

Table 3: *Different methods comparison results*

Methods	Validation AUC	Test AUC
CE_Original	69.77	75.27
CE_Dul	71.15	72.92
CE_Gua	71.57	75.59
FL_Original	74.17	69.78
FL_Dul	69.92	75.10
FL_Gua	73.58	83.59
Ensemble	**76.29**	**85.43**

In Table 3, the performance of the fine-tuning model based on ResNet50 is better than that of the traditional machine learning methods such as RF, MLP and LR. In terms of the loss function, CE performs better than FL on the test set, and overfitting occurs when FL is used for training on the original training set. However, after using "Dul" or "Gua" for data augmentation, FL performed better than CE on the test set. This proves that a combination of multiple imbalanced data processing methods will achieve better results. In addition, it can also be seen that using Gaussian noise to process imbalanced data is better than directly expanding minority samples. Finally, using randomness to train multiple models for integration can further improve the prediction performance of the entire model.

5. Conclusions

In order to use acoustics data to identify people infected with COVID-19 more accurately, we propose a deep learning method that incorporates multiple image processing techniques. First, we transform the acoustics data into spectrogram data, which can better suit the deep learning model. After that, Gaussian noise-based data augmentation and focal loss are introduced to solve the problem of imbalanced data. Based on the pre-training model of ResNet50, we combine the fine-tuning technology in transfer learning to adjust the weight of the deep neural network to make it more suitable for the identification of COVID-19 infected persons. In addition, in order to make the model we designed more robust, we use ensemble learning to build multiple deep learning models. When training these models, we adopt an advanced data extraction method with randomness and uncertainty to build sample subsets. Finally, the experimental results show that the proposed method can effectively identify the infected with COVID-19 and is superior to other state-of-the-art methods.

6. Acknowledgements

We are very grateful to the organizers of the DiCOVA 2021 Challenge for their efforts in providing the participants with data and a platform for the competition. And, this research is supported by the China Scholarship Council (No. 202006060162).

7. References

[1] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing," *Radiology*, vol. 296, no. 2, pp. E41–E45, 2020.

[2] A. K. J. Teo, Y. Choudhury, I. B. Tan, C. Y. Cher, S. H. Chew, Z. Y. Wan, L. T. E. Cheng, L. L. E. Oon, M. H. Tan, K. S. Chan and L. Y. Hsu, "Saliva is more sensitive than nasopharyngeal or nasal swabs for diagnosis of asymptomatic and mild COVID-19 infection," *Scientific Reports*, vol. 11, no. 1, pp. 1–8, 2021.

[3] D. L. Grieco, F. Bongiovanni, L. Chen, L. S. Menga, S. L. Cutuli, G. Pintaudi, S. Carelli, T. Michi, F. Torrini, G. Lombardi, G. M. Anzellotti, G. D. Pascale, A. Urbani, M. G. Bocci, E. S. Tanzarella, G. Bello, A. M. Dell'anna, S. M. Maggiore, L. Brochard and M. Antonelli, "Respiratory physiology of COVID-19-induced respiratory failure compared to ARDS of other etiologies," *Critical Care*, vol. 24, no. 1, pp. 1–11, 2020.

[4] A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network," *Pattern Recognition*, vol. 81, pp. 200–212, 2018.

[5] Y. Wu, H. Mao, and Z. Yi, "Audio classification using attention-augmented convolutional neural network," *Knowledge-Based Systems*, vol. 161, pp. 90–100, 2018.

[6] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, 2017.

[7] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing," *Internet of Things*, p. 100222, 2020.

[8] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu and Y. Yuan, "An interpretable mortality prediction model for COVID-19 patients," *Nature Machine Intelligence*, vol. 2, 2020.

[9] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks," *Computers in Biology and Medicine*, vol. 121, p. 103795, 2020.

[10] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8599–8603.

[11] J. Sharma, O. Granmo, and M. Goodwin, "Environment Sound Classification using Multiple Feature Channels and Attention based Deep Convolutional Neural Network," in *INTERPEECH 2020*, 2020, pp. 1186–1190.

[12] T. Jiang, M. Zhao, L. Li, and Q. Hong, "The XMUSPEECH system for short-duration speaker verification challenge 2020," in *INTERSPEECH 2020*, 2020, pp. 736–740.

[13] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] D. Hendrycks, L. Kimin, and M. Mazeika, "Using Pre-Training Can Improve Model Robustness and Uncertainty," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 2712–2721.

[18] Y. Guo, H. Shi, A. Kumary, K. Grauman, T. Rosing, and R. Feris, "Spottune: Transfer learning through adaptive fine-tuning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4805–4814.

[19] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 4970–4979.

[20] X. Yang, Z. Zeng, T. Sing G., L. Wang, V. Chandrasekar, and S. C. H. HOI, "Deep learning for practical image recognition: Case study on kaggle competitions," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 923–931.

[21] K. He, Y. Wang, J. Hopcroft, "A Powerful Generative Model Using Random Weights for the Deep Image Representation", *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 631-639.

[22] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, L. E. Barnes, "RMDL: Random Multimodel Deep Learning for Classification", *Proceedings of the 2nd International Conference on Information System and Data Mining* , 2018, pp. 19-28.

[23] SJ. Pan, Q. Yang, "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering* , vol. 22, no. 10, pp. 1345-1359, 2009.

[24] Q. Sun, Y. Liu, T. Chua, B. Schiele, "Meta-Transfer Learning for Few-Shot Learning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 2019, pp. 1345-1359.

[25] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, "ImageNet: A large-scale hierarchical image database", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248-255.

[26] Z. Zhou, "Ensemble methods: foundations and algorithms", *Chapman and Hall/CRC press*, 2012.

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