Assembling Existing Labels from Public Datasets to Diagnose Novel Diseases: COVID-19 in Late 2019

Zengle Zhu[1] **Mintong Kang**[2] **Alan Yuille**[3] **Zongwei Zhou**[3,*]

[1]Tongji University [2]University of Illinois at Urbana Champaign [3]Johns Hopkins University

Code: https://github.com/MrGiovanni/LabelAssemble

Abstract

The success of deep learning relies heavily on the availability of large annotated datasets, but neither sizable data nor annotation is easily accessible for novel diseases. This paper uses the classification of COVID-19 in late 2019 as an example to demonstrate the effectiveness of a novel strategy, named "Label-Assemble". To facilitate the diagnosis of novel diseases, we propose to assemble existing labels from public datasets. Although novel diseases are not in the existing labels, we discover that learning from alternative labels can dramatically improve the diagnosis of the novel disease as these labels can better define the classification boundary of the novel disease. This discovery has the potential to accelerate the development circle of computer-aided diagnosis of novel diseases, in which positive label is hard to collect, yet negative labels are usually available and relatively easier to assemble. Label-Assemble achieves 99.3% accuracy on the COVIDx-CXR2 dataset, which significantly exceeds the previous state of the art (96.3% accuracy) and only uses 3% of the annotated COVID-19 images. We further investigate the implementation of the assembling strategy, showing that assembling pathologically related labels, supplemented by semi-supervised learning, is preferred.

1 Introduction

Despite the grand success of deep learning in a few medical applications [9, 2, 15, 19, 6, 7, 22, 23], its prohibitively high annotation costs raise doubts about the feasibility of applying it to those medical specialties that lack such magnitude of annotation [28, 27, 25, 26]. For example, it is impossible to acquire sufficient annotation or even to gather sizable data for novel diseases and emerging pandemics during the outbreak. Meanwhile, collecting data and labels for a few common diseases is relatively easier, and our research community has already created large, annotated datasets through a collective effort [24, 12, 16, 4, 17, 11, 10, 1, 5, 3]. We ask: *can we exploit these existing, large, annotated datasets*[2] *to facilitate computer-aided diagnosis of novel diseases?*

We examine the effectiveness of assembling existing labels from public datasets. We name it "Label-Assemble" [14] and demonstrate this strategy using COVID-19 (which occurred in late 2019) as an example, and NIH ChestX-ray14[3] [21] as existing datasets to be exploited (released in 2017) [29]. Normally, one would not consider using these extra labels which seem unrelated, but we find that those existing datasets, even if they were not created for the novel diseases *per se*, are helpful for improving the performance of diagnosing novel diseases while substantially reducing annotation cost. Label-Assemble requires four prerequisites: (1) same medical imaging modality (e.g., X-ray), (2)

[*] Corresponding author: Zongwei Zhou (zzhou82@jh.edu)

[2] These datasets often do not contain labels of novel diseases.

[3] NIH ChestX-ray14 provides 112,120 annotated X-rays of Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

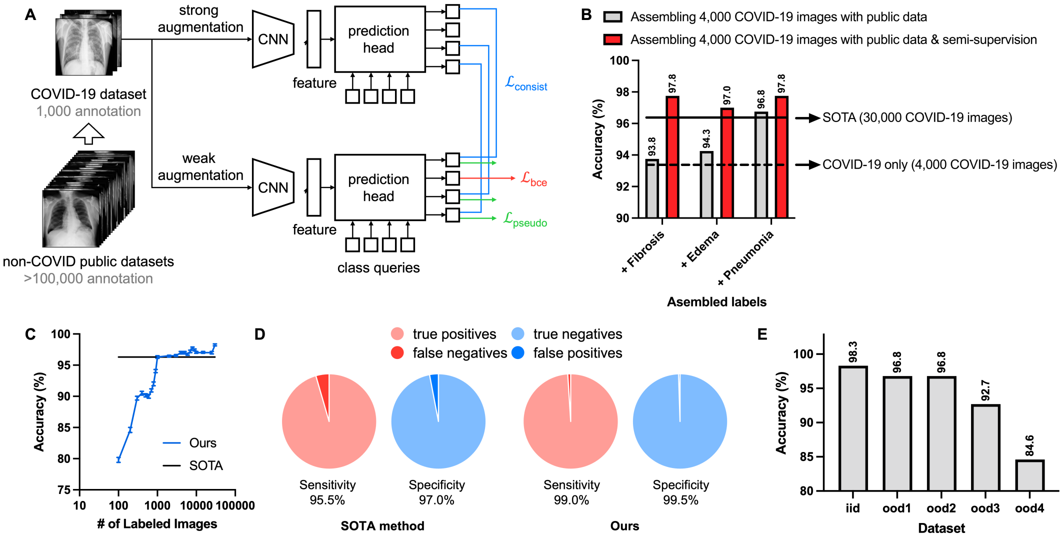

Figure 1: **A.** Semi-supervised Label-Assemble model architecture. **B.** The types of non-COVID labels to be assembled do not matter if semi-supervised learning is applied. **C.** By assembling 14 existing non-COVID labels from ChestXray-14, only 4,000 annotated COVID-19 X-rays can achieve similar performance to the current state of the art, which uses over 30,000 annotated images. **D.** Assembling existing non-COVID labels can help the model classify COVID-19 with higher sensitivity and specificity. **E.** The test result of the four out-of-distribution (ood) COVID-19 datasets. Our model—trained on an assembly of existing datasets and COVIDx—performs robustly on X-ray images from a variety of hospitals.

similar body region (e.g., chest region), (2) reasonably large in data scale (e.g., 1,000 COVID-19 examples), and (4) consistent labeling format (e.g., classification).

Currently, there are more than 100,000 COVID-19 X-rays in publicly available datasets, and the number is increasing [8]. The accuracy of COVID-19 classification has reached 96.3% using current state-of-the-art methods, trained on a dataset with 15,000 annotated X-ray images [20]. Our study has concentrated on the beginning of the outbreak when annotated data was insufficient. By assembling the existing 14 labels of common thorax diseases, with only 3% of the annotated COVID-19 data, our strategy enables the model to produce a comparable performance to the current state of the art (Figure 1A). When training with the entire annotated COVID-19 dataset, we obtain a significant performance improvement over the previous best solution (96.3%→99.3% accuracy). Finally, we demonstrate that the model trained with diverse data generalizes well on four other COVID-19 datasets, which account for a total of 4,000 X-rays collected from various hospitals.

2 Materials & Methods

The COVIDx CXR-2 and ChestX-ray14 datasets are used in our study. COVIDx CXR-2 provides about 15,000 subjects from at least 51 countries. This dataset is split into training (over 30000 images) and testing (400 images) sets. The performance of COVID-19 diagnosis is evaluated on the test set. The entire ChestX-ray14 dataset is used for label assembling, which provides 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined 14 disease image labels. We use DenseNet121 as the classification backbone [13] and sigmoid as the activation function for the final output. This classifier contains 1+N output headers. One of the output headers is to predict whether it is a COVID positive or not, and the other N output headers are used to predict assembled labels. Compared with fully-supervised Label Assemble, semi-supervised Label Assemble can better improve the effect on classification by learning more negative sample features of different categories. The semi-supervised component consists of three loss functions as follows.

BCE Loss: Given the ground truth (y) and the output(a), binary cross entropy loss is used if the label is provided, i.e., $L_{\text{bce}} = -(y \cdot log(a) + (1 - y) \cdot log(1 - a))$.

Pseudo Labels & Consistency Constraints: To unleash the full potential of unannotated labels, we introduce a sharpening operator to generate pseudo-labels, i.e.,

$$\tilde{a} = \begin{cases} a + (1-a)/t, & a > \tau \\ a - a/t, & a \leq \tau \end{cases} \tag{1}$$

where $\tilde{a}$ is the pseudo-label of the answer, t is the sharpen temperature, and τ is the threshold ($\tau = 0.5$ in our experiments). The prediction beyond (below) the threshold τ can be assigned to a higher (lower) score controlled by t. If $t = \infty$, there is no pseudo-labeling; if $t = 1$, the model converts a soft label to a completely hard label (either 1 or 0, equivalent to FixMatch et al. [18]). With the sharpening operator, the loss enables the model to operate self-training on unlabeled data, i.e., $L_{\text{pseudo}} = \|a_w - \tilde{a_w}\|_2^2$, where a_w and $\tilde{a_w}$ denote the answer of weakly augmented images and their sharpened pseudo-labels, respectively. To reduce the domain gap across the heterogeneous data sources, we further employ consistency constraints on weakly augmented (a_w) and strongly augmented (a_s) images. The consistency loss can be formulated as $L_{\text{consist}} = \|a_s - \tilde{a_w}\|_2^2$.

3 Results & Discussion

The current state-of-the-art (SOTA) method for COVID-19 classification [20] holds an accuracy of 96.3%, and the model was trained on 30,000 annotated X-ray images. Figure 1 shows that training on 1,000 annotated COVID-19 X-rays, assembled by the labels in ChestXray14, can achieve similar performance to the SOTA method trained on 15,000 annotated COVID-19 X-rays. The one-tailed independent t-test between our Label-Assemble and SOTA method indicates that there is no statistical difference (p-value=0.6) between the two accuracy scores.

When using 4000 images, assembling Fibrosis labels improves the accuracy of COVID-19 classification from 93.5% (95%CI: 93.1%-93.6%) to 93.75% (95%CI: 93.5%-94.0%). And assembling Edema labels improves accuracy of COVID-19 classification from 93.5% (95%CI: 93.1%-93.6%) to 94.25% (95% CI: 94.0%-94.4%). Moreover, assembling Pneumonia labels improves accuracy of COVID-19 classification from 93.5% (95%CI: 93.1%-93.6%) to 96.75% (95% CI: 96.6%-97.0%). More experimental results are presented in Appendix. Pneumonia contributes more to identifying COVID-19 because it is more pathologically related. Lesion maps of COVID-19 appeared as indistinct patchy ground-glass opacity with little consolidation, while Pneumonia appeared as ground-glass opacity with some consolidation. Clinical features also partially overlap.

When using the semi-supervised component, it not only improves accuracy but also eliminates the effects of category similarity. Assembling fibrosis, edema, and pneumonia labels, we get an accuracy of 97.75% (95% CI: 97.6%-97.9%), 97.0% (95% CI: 96.8%-97.1%), 97.75% (95% CI: 97.5%-97.8%) respectively.

The classification of COVID-19 can greatly benefit from leveraging the labels of 14 common thorax diseases. Especially, assembling labels of pathologically similar diseases (e.g., pneumonia and COVID-19) results in much more performance gain compared with other diseases (e.g., Edema and COVID-19). Label-Assemble is expected to improve the COVID-19 classification model effect. It shows a good prospect in disease diagnosis: we can use common pneumonia to improve the accuracy of rare pneumonia and reduce the cost of labeling.

Finally, assembling existing labels from public datasets generalizes our model well to X-rays acquired from different hospitals and protocols. We directly evaluated the model without further training on four different COVID-19 datasets. These datasets are selected from Kaggle[4], and our model achieves an accuracy of COVID-19 classification by 96.8% (95% CI: 96.8%-96.9%), 96.8% (95% CI: 96.8%-97.0%), 92.7% (95% CI: 92.6%-92.9%), 84.6% (95% CI: 84.5%-84.9%), respectively.

Although our paper focuses on COVID-19, the proposed method and discovery are applicable to many novel diseases, e.g., Silicosis. Note that we still need many labeled data for novel diseases for evaluation. Our work demonstrates that only a few annotated examples are needed, which does not mean the problem of annotation sparsity is solved. In future work, we will investigate how to find a few examples of novel diseases to work the best with existing labels. This supposes we have many unlabeled data.

[4] Evaluated on four out-of-distribution (ood) data sources: dataset1, dataset2, dataset3, and dataset4.

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Zhuotun Zhu, Liangyu Chen, Jie-Neng Chen, and Bowen Li for their constructive suggestions; thank Camille Torrico and Alexa Delaney for improving the writing of this paper.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[3] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[4] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[5] Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

[6] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[7] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[8] Arman Haghanifar, Mahdiyar Molahasani Majdabadi, Younhee Choi, S Deivalakshmi, and Seokbum Ko. Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning. *Multimedia Tools and Applications*, pages 1–31, 2022.

[9] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021.

[10] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

[11] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020.

[12] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[13] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

[14] Mintong Kang, Yongyi Lu, Alan L Yuille, and Zongwei Zhou. Data, assemble: Leveraging multiple datasets with heterogeneous and partial labels. *arXiv preprint arXiv:2109.12265*, 2021.

[15] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[16] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national

institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[17] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[18] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[19] Hongkai Wang, Zongwei Zhou, Yingci Li, Zhonghua Chen, Peiou Lu, Wenzhi Wang, Wanyu Liu, and Lijuan Yu. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images. *EJNMMI research*, 7(1):1–11, 2017.

[20] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020.

[21] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[22] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022.

[23] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. *arXiv preprint arXiv:2210.12843*, 2022.

[24] Yang Yang, Xueyan Mei, Philip Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Katherine Link, Thomas Yang, Chendi Cao, et al. Radimagenet: A large-scale radiologic dataset for enhancing deep learning transfer learning research. 2021.

[25] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021.

[26] Zongwei Zhou, Michael Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health: The Role of AI*. Springer, 2022.

[27] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

[28] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019.

[29] Zengle Zhu, Mintong Kang, Alan Yuille, and Zongwei Zhou. Assembling and exploiting large-scale existing labels of common thorax diseases for improved covid-19 classification using chest radiographs. In *Radiological Society of North America (RSNA)*, 2022.