HIGH-CONFIDENCE PSEUDO-LABELS FOR DOMAIN ADAPTATION IN COVID-19 DETECTION

PREPRINT

Robert Turnbull
Melbourne Data analytics Platform
The University of Melbourne
Parkville, VIC 3053
robert.turnbull@unimelb.edu.au

Simon Mutch
Melbourne Data analytics Platform
The University of Melbourne
Parkville, VIC 3053
simon.mutch@unimelb.edu.au

March 20, 2024

ABSTRACT

This paper outlines our submission for the 4th COV19D competition as part of the 'Domain adaptation, Explainability, Fairness in AI for Medical Image Analysis' (DEF-AI-MIA) workshop at the Computer Vision and Pattern Recognition Conference (CVPR). The competition consists of two challenges. The first is to train a classifier to detect the presence of COVID-19 from over one thousand CT scans from the COV19-CT-DB database. The second challenge is to perform domain adaptation by taking the dataset from Challenge 1 and adding a small number of scans (some annotated and other not) for a different distribution. We preprocessed the CT scans to segment the lungs, and output volumes with the lungs individually and together. We then trained 3D ResNet and Swin Transformer models on these inputs. We annotated the unlabeled CT scans using an ensemble of these models and chose the high-confidence predictions as pseudo-labels for fine-tuning. This resulted in a best cross-validation mean F1 score of 93.39% for Challenge 1 and a mean F1 score of 92.15 for Challenge 2.

Keywords COVID-19 · CT Scan

1 Introduction

Deep learning models are becoming an increasingly common tool used for medical image analysis. In combination with expert medical professionals, these models can aid in the accurate detection of diseases such as COVID-19 [Kollias et al., 2020a,b]. Here, deep learning models have been shown to provide accurate predictions for the presence of the disease from CT scans alone.

The 4th COV19D competition is being run as part of the 'Domain adaptation, Explainability, Fairness in AI for Medical Image Analysis' (DEF-AI-MIA) workshop [Kollias et al., 2024] at the Computer Vision and Pattern Recognition Conference (CVPR) in 2024. It follows on from previous competitions held as part of the IEEE ICCV 2021 [Kollias et al., 2021], ECCV 2022 [Kollias et al., 2022] and ICASSP 2023 [Kollias et al., 2023a, Arsenos et al., 2023] workshops. In the 2024 competition, two challenges presented to participants. The first is to take over one thousand CT scans from the COV19-CT-DB database [Kollias et al., 2023b, Arsenos et al., 2022], annotated as belonging to patients with or without COVID, and train a classifier. The second challenge is to perform domain adaptation. A smaller dataset with CT scans from a different distribution to Challenge 1 is provided. This also includes almost 500 scans which have not been annotated. The challenge is to use the dataset for challenge 1 and make the best classifications on data from a distribution like the additional dataset.

In our submission, we build on work for previous years [Turnbull, 2023a,b] where we trained 3D ResNet and SwinTransformer models. In our 2023 submission, we segmented the lungs and cropped the CT scans accordingly. Here we experiment with segmenting both lungs and training additional models with the individual lungs as input. We also use pseudo-labels for augmenting the annotated dataset in the domain adaptation challenge.

	COVID	NON-COVID	Total
Training	703	655	1358
Validation	170	156	326
Test	—	—	1,413

Table 1: The Challenge 1 Dataset.

	COVID	NON-COVID	Total
Training	120	120	240
Validation	65	113	178
Unannotated	—	—	494
Test	—	—	4,055

Table 2: The Challenge 2 Dataset.

2 Dataset

The 2024 competition dataset is divided between the two challenges. The Challenge 1 dataset comprises a total of 3,107 scans, with 1,684 used for training and validation (table 1). We divided the training dataset into four partitions which together with the official validation set gives five partitions for cross-validation. The Challenge 2 specific dataset comprises 4,979 scans, including 912 scans to be used in training and validation. Of these, 494 scans are not labeled as to whether or not the subject is infected with COVID-19. We combined both training and validation partitions from the Challenge 2 dataset and then divided this into roughly equal partitions for five-fold cross-validation.

We also included the public STOIC dataset [Revel et al., 2021] which includes 2,000 CT scans labeled as COVID-19 positive or negative. We ignored the severity categories of COVID-19 positive scans.

3 Methods

3.1 Preprocessing

As in Turnbull [2023b], we first preprocessed the CT scans to segment the lungs and to crop the volumes to the lungs. We then further crop each lung individually. Our aim is not to perfectly segment the lungs from all surrounding tissue, but instead to find the maximum crop that guarantees each lung will be fully contained and contamination from the opposite lung is minimized. This is achieved by taking each slice in turn, applying a binary threshold using Otsu's method, identifying all contours in the resulting image, removing contours with enclosed areas below a 500 pixels^2 and which are clearly not associated with lungs (e.g. span the entire width of the slice), and finally taking the two largest contours which overlap by less than 20% of their horizontal axis extent. Once we have identified the lungs in each slice, we determine their axis-aligned bounding boxes. The left (/right) lung is cropped to be from the left (/right) side of the volume, to the left (/right) edge of the largest bounding box surrounding the right (/left) lung. A volume containing each lung is stored and these are able to be used as input to the model.

The cropped volumes are interpolated to a single size. The cropped volumes of both lungs are interpolated to $256 \times 256 \times 176$. The individual lungs are interpolated to a size of $320 \times 160 \times 224$.

3.2 Models

We use two neural network architectures. The first is a 3D ResNet [He et al., 2016, Tran et al., 2018] with adaptations discussed in Turnbull [2023b]. The second is a 3D Swin Transformer [Liu et al., 2021]. We used the 'Tiny' size of the Swin Transformer to allow for larger CT scan volumes on the GPU. Both neural network models were pretrained on the Kinetics 400 video classification dataset Kay et al. [2017].

3.3 Training Procedure

The models were trained for 30 epochs with a batch size of 2 using cross-entropy loss with the Adam optimizer [Kingma and Ba, 2014]. Each volume was included in the training and validation datasets twice with the second one

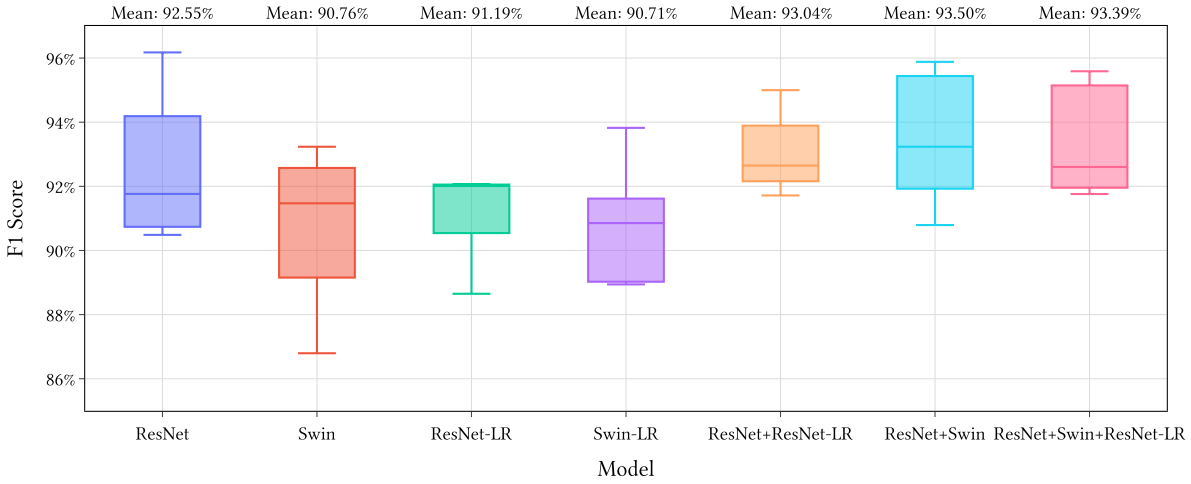

Figure 1: The cross-validation results for challenge 1. Models joined with a '+' are ensembles with prediction probabilities averaged.

reflected through the sagittal plane. The brightness and contrast for each scan was randomly adjusted during the training according to the scheme discussed in Turnbull [2023b].

3.4 Pseudo-Labels

For Challenge 2, we train an ensemble of models on the annotated scans and then make predictions on the unannotated scans. These predictions can be used as pseudo-labels [Lee et al., 2013]. To mitigate against training with too many scans with incorrect pseudo-labels, we only include predictions with higher confidence, meaning that only include predictions with a probability of the 0.7 or greater. These scans with their pseudo-labels are then included in the training dataset for fine-tuning the models for an additional 10 epochs.

4 Results

4.1 Challenge 1

Three models were trained for Challenge 1: a ResNet model, a Swin Transformer model and a ResNet model trained on the individual left and right lungs (ResNet-LR). The best performing model was the ResNet which achieved a mean F1 score of 92.55% across the five cross-validation partitions (fig. 1. Averaging the ResNet and the Swin Transformer results gave the highest F1 score overall at 93.5%, although including the ResNet-LR model results in the ensemble produced a slightly lower F1 score of 93.4% but with a smaller variance. The five competition submissions for Challenge 1 are:

1. ResNet
2. Swin Tranformer
3. Ensemble of ResNet and ResNet-LR
4. Ensemble of ResNet and Swin Transformer
5. Ensemble of ResNet, Swin Transformer and ResNet-LR

All submissions average results across models trained on the five cross-validation partitions.

4.2 Challenge 2

A ResNet and a Swin Transformer were trained to predict the pseudo-labels. An ensemble of both achieved an F1 score of 91.2% (fig. 2). If we filter the validation datasets for only high-confidence scans with a probability of being with or

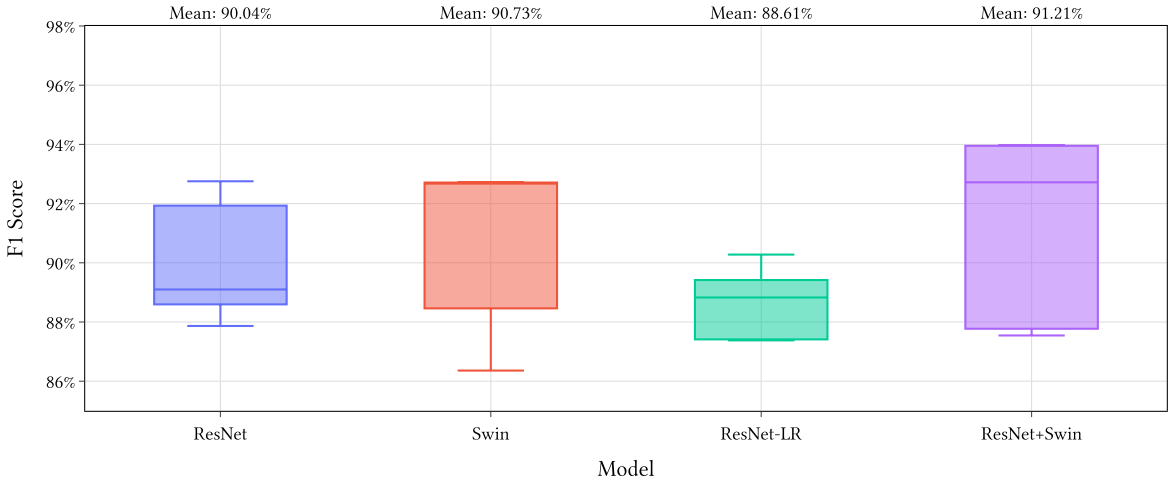

Figure 2: The cross-validation results for challenge 2 before adding in pseudo-labels.

without COVID-19 above 0.7, then the F1 score increases to 95.8%. Using this ensemble, predictions were made on the 494 unannotated scans for Challenge 2. Of these, 414 predictions were above the threshold of 0.7 and these were assigned as pseudo-labels. This improved the F1 score for the Swin Transformer to 91.22% but the result for the ResNet decreased a small amount (fig. 3). The ResNet-LR model improved from 88.6% to 89.85%. An ensemble of both Swin Transformer models (with and without pseudo-labels) together with the ResNet-LR trained with pseudo-labels achieved the highest F1 score of 92.15%.

The five competition submissions for Challenge 2 are:

1. ResNet
2. Swin Tranformer
3. Swin Tranformer with pseudo-labels
4. Ensemble of Swin Transformer and Swin Transformer with pseudo-labels
5. Ensemble of Swin Transformer and Swin Transformer with pseudo-labels and the ResNet model with individual lungs and pseudo-labels.

5 Conclusion

The approach used in this paper achieved high validation F1 scores for both challenges. The best result for Challenge 1 was an ensemble of the ResNet and Swin Tranformer models with an average F1 score of 93.5%. The best single model for Challenge 2 was the Swin Transformer at an F1 score of 90.73%. This improved to 91.22% when pseudo-labels with high-confidence were added to the training set. An ensemble achieved even better results with an F1 score of 92.15%. These results for the domain adaptation challenge show that high accuracy can be obtained for classification of a new distribution of CT scans with a relatively small number of annotated examples.

6 Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus. We acknowledge the help of Evelyn Mannix.

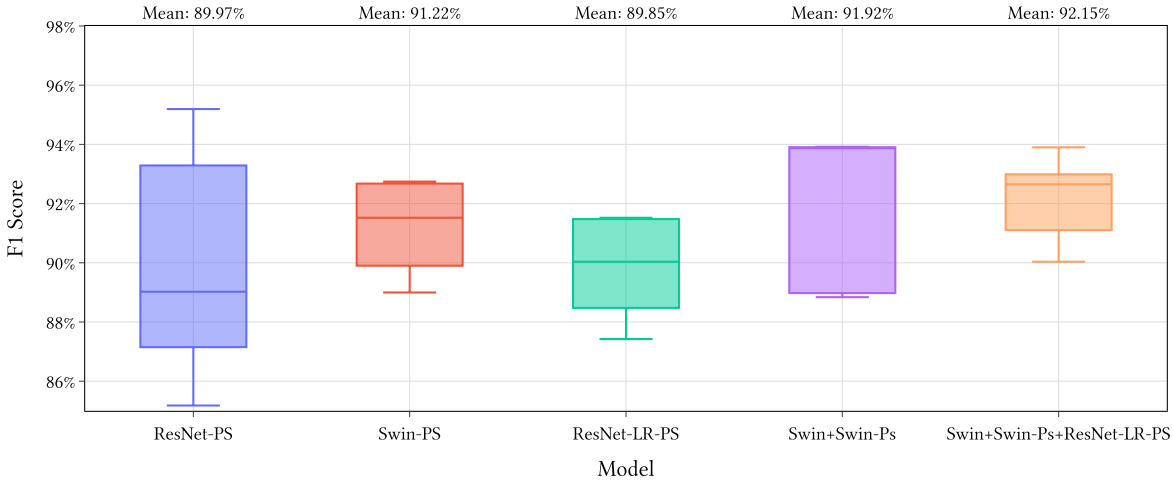

Figure 3: The cross-validation results for challenge 2 after adding in pseudo-labels.

References

Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020a.

Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, page 251–267, 2020b.

Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Domain adaptation, explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. *arXiv preprint arXiv:2403.02192*, 2024.

Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 537–544, 2021.

Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *European Conference on Computer Vision*, page 677–690. Springer, 2022.

Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-enabled analysis of 3-d ct scans for diagnosis of covid-19 & its severity. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023a.

Anastasios Arsenos, Andjoli Davidhi, Dimitrios Kollias, Panos Prassopoulos, and Stefanos Kollias. Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023.

Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023b.

Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, page 1–5. IEEE, 2022.

Robert Turnbull. Using a 3D ResNet for Detecting the Presence and Severity of COVID-19 from CT Scans. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, number 7, pages 663–676, Cham, 2023a. Springer Nature. ISBN 978-3-031-25082-8. doi:10.1007/978-3-031-25082-8_45.

Robert Turnbull. Lung segmentation enhances covid-19 detection. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023b. doi:10.1109/ICASSPW59220.2023.10193492.

Marie-Pierre Revel, Samia Boussouar, Constance de Margerie-Mellon, Inès Saab, Thibaut Lapotre, Dominique Mompoint, Guillaume Chassagnon, Audrey Milon, Mathieu Lederlin, Souhail Bennani, Sébastien Molière, Marie-Pierre Debray, Florian Bompard, Severine Dangeard, Chahinez Hani, Mickaël Ohana, Sébastien Bommart, Carole Jalaber, Mostafa El Hajjam, Isabelle Petit, Laure Fournier, Antoine Khalil, Pierre-Yves Brillet, Marie-France Bellin, Alban Redheuil, Laurence Rocher, Valérie Bousson, Pascal Rousset, Jules Grégory, Jean-François Deux, Elisabeth Dion, Dominique Valeyre, Raphael Porcher, Léa Jilet, and Hendy Abdoul. Study of thoracic ct in covid-19: The stoic project. *Radiology*, 301(1):E361–E370, 2021. doi:10.1148/radiol.2021210384. URL `https://doi.org/10.1148/radiol.2021210384`. PMID: 34184935.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. doi:10.1109/CVPR.2018.00675.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. URL `https://arxiv.org/abs/2106.13230`.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL `https://arxiv.org/abs/1705.06950`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.