Science in Information Technology Letters
Vol. 1., No. 2, November 2020, pp. 92-99
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Detection and segmentation of lesion areas in chest ct scans for the prediction of COVID-19

Aram Ter-Sarkisov [a,1,*]

[a] City, University of London, Northampton Square, London, United Kingdom
[1] alex.ter-sarkisov@city.ac.uk
* corresponding author

ARTICLE INFO

Article history
Received October 2, 2020
Revised October 15, 2020
Accepted November 25, 2020

Keywords
COVID-19
Lesion Segmentation
Pneumonia Classification
Mask R-CNN

ABSTRACT

This paper compares the models for the detection and segmentation of Ground Glass Opacity and Consolidation in chest CT scans. These lesion areas are often associated both with common pneumonia and COVID-19. We train a Mask R-CNN model to segment these areas with high accuracy using three approaches: merging masks for these lesions into one, deleting the mask for Consolidation, and using both masks separately. The best model achieves the mean average precision of 44.68% using MS COCO criterion on the segmentation across all accuracy thresholds. The classification model, COVID-CT-Mask-Net, learns to predict the presence of COVID-19 vs. common pneumonia vs. control. The model achieves the 93.88% COVID-19 sensitivity, 95.64% overall accuracy, 95.06% common pneumonia sensitivity, and 96.91% true-negative rate on the COVIDx-CT test split (21192 CT scans) using a small fraction of the training data. We also analyze the effect of the Non-Maximum Suppression of overlapping object predictions, both on the segmentation and classification accuracy. The full source code, models, and pre-trained weights are available on https://github.com/AlexTS1980/COVID-CT-Mask-Net.

1. Introduction

Many publications showed commonalities and differences in the manifestation of COVID-19 and common pneumonia (CP) in chest CT scans. Both conditions give rise to lesions like Ground Glass Opacity (GGO) and Consolidation (C), but they manifest differently. In COVID-19 patients, GGO is present more often (number of lesions/scan slice) and tends to be bilateral. Subsegmental C areas are also present more often than the patients with CP [1], [2]. The absolute majority of patients with COVID-19 display either GGO, or Consolidation, or a mix of both [3], and GGO lesions are more diffused, larger in size, and spread over larger areas [2]. The problem with these findings is that many of them are not statistically significant, e.g., the difference in the location of lesions in [2] and sample sizes are relatively small (e.g., n=34 in [4]). As a result, several machine learning methods were recently developed to help experts determine the diagnosis using chest CT scans.

The two-class problems (COVID-19 vs. CP, COVID-19 vs. Control) are inherently easier to solve due to fewer false positives than the three-class problem (COVID-19 vs. CP vs. Control). Some of the best solutions for the two-class problems presented in [5], [6] include DenseNet169, ResNet50, and DRE-Net [7]. Solutions for the three-class problem using chest CT scans include ResNet18 [8], ResNet50 [9], COVIDNet-CT [10] and multiscale spatial pyramid [1] as feature extractors. The disadvantage of most COVID-19 detectors is either evaluating the model on a small amount of data [8], [9], implying weak capacity for generalization, or dependence on a large dataset and data balancing tricks [1], [10] for training models.

Semantic segmentation predicts the object's masks from images by predicting the class at a pixel level. Semantic segmentation models like FCN and U-Net are widely used to segment GGO, C, and other lesions. These predicted masks are often used in combination with the extracted features to predict the image's class [6], [11], improving the final prediction over the baseline feature extractor. Models like Mask R-CNN [12] solve the combined problem of object detection (localization) using bounding boxes and predicting the object's mask, known as instance segmentation. In this paper, we compare three ways to predict instances of lesions' masks. First, we use only masks for GGO areas, merging C with the background. Secondly, we merge GGO and C masks in a single 'lesion' mask. Finally, we keep separate masks for GGO and C instances (this approach was first presented in [13]). The first two are 1+1 class problem (1 object class + background, the latter is a 2+1 problem (2 object classes + background). The observations explain our choices that areas with GGO have larger sizes and are observed more frequently than areas with C in COVID-19 patients, hence GGO class alone may be sufficient for COVID-19 prediction.

We implement the following novelties in our solution and achieve the following results:

1. Merge of GGO and C masks into a single "Lesion" class; both improve the segmentation precision and the accuracy of the classification model built on top of the segmentation model compared to using only the GGO mask.
2. Mask R-CNN segmentation achieves a precision of 61.92%@0.5IoU, 45.22%@0.75IoU, and mean average precision of 44.68% (across all IoU thresholds).
3. The classifier built on top of the model with separate masks achieves a COVID sensitivity of 93.88% and overall accuracy of 95.64% on the COVIDx-CT test split of the CNCB CT scans dataset.
4. Compared to other solutions for a 3-class problem, we use only a small fraction of the dataset to get these results: 5% of the COVIDx-CT training split and 3% of the total data.

2. Method

2.1 Datasets

The segmentation problems solved in the paper are shown in Fig. 1. The 2-class problem, Fig. 1. (b), was first solved in [13]. We compare this problem to two 1-class problems: For the first one, Fig. 1. (c), we only consider GGO as the positive class and train the model to detect its instances (predict the bounding box coordinates and segment the positive area within it). Consolidation (C) masks are discarded (merged with the background). For the second problem, Fig. 1. (d), we merge the masks for GGO and C into one class ('lesion'), thus increasing the positive class's prevalence in the error space, compared to only GGO.

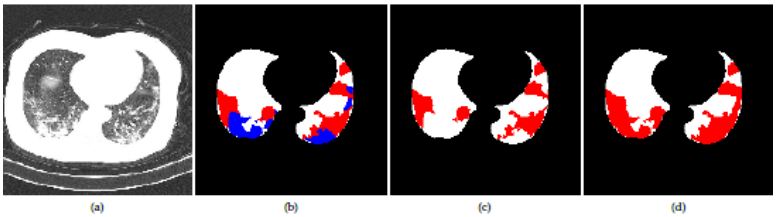

Fig. 1. Segmentation masks for the same CT scan slice. (a) input raw image. (b) 2-class problem, red: GGO masks, blue: C masks. (c): 1-class problem (only GGO). (d) 1-class problem (merged masks for GGO and C). White masks are the lungs areas. Best viewed in color.

We use the same dataset split of 600 training + 150 validation images with the varying representation of either class in each image as in [13]. Many images are purely negative (only background mask). To train the Mask R-CNN model to solve these problems, we extract bounding box coordinates of each lesion object from the masks and either use 3 (2 positives + 1 background label) or 2 (1 positive+1 background) labels for objects.

We define each object as isolated from other areas of the same class, either by background or different class areas. The lung mask is merged with the background for all problems. Except for the usual normalization using global mean and standard deviation, no other data augmentations or balancing (resampling, class balancing, image rotations, etc.) were applied to the data at any stage, unlike in many other solutions, e.g. [10]. For the classification problem us re-use the train/validation/test splits in [10], [13].

We sample 3000 images from the COVIDx-CT [10] train split (1000 images/class) and use their full validation (21036 CT scans) and test (21191 CT scans) splits. As a result of our approach, we use less than 5% of the COVIDx-CT training data split and 3% of the total CNCB CT scans data [6]. Each image is the same size as in the segmentation data, 512x512x3 pixels, all alpha-channels removed. The training split used in this paper is the same as in [13], to have a fair comparison. As with the segmentation problem, no other data normalization techniques were used apart from the global image normalization.

2.2 Model

We study in-depth the effect of the non-maximum suppression (NMS) threshold, a criterion for discarding overlapping bounding box predictions in the Region Proposal Network (RPN) at train and test stages and Region of Interest (RoI) at the test stage. High threshold values mean that a larger number of overlapping predictions is kept in the model. At the training stage of the segmentation model, low NMS in the RPN implies that a lower number of high-scoring predictions will be passed to RoI, and a lower number of high-scoring predictions will be processed by RoI, both at train and test stages. It is because of RoI. After passing the region of interest through the classification' head' (two fully connected layers and a class+bounding box layer), we can still classify this region as background, even if the prediction was derived from the 'positive' anchor [12]. The hyperparameters of the segmentation model are set in Table 1.

Table 1. Key hyperparameters of the segmentation models. RPN output is the number of predictions after the NMS step, RoI output is the maximum number of predictions at test stage after the NMS stage, RPN scoreθ is the threshold for positive predictions at train time, RoI scoreq is the threshold for object confidence at test time. In RoI, NMS threshold is used only in testing.

Backbone	Anchor Sizes	Anchor Scales	RPN NMS_θ	RoI NMS_θ	RPN Sample	RoI Sample	RPN Output	RoI Output	RPN $Score_\theta$	RoI $Score_\theta$
ResNet50 +FPN	$2^{2:5}$	0.1, 0.25, 0.5, 1, 1.5, 2	0.25/0.75	- 0.25/0.75	256 -	256 -	1000	128	0.75 -	- 0.75

The model computes 4 loss functions: two by RPN (objectness and bounding box coordinates) and two by RoI (class and bounding box coordinates). For our training and evaluation, we use the torchvision v0.3.0. In COVID-CT-Mask-Net, see Fig. 2 and Fig. 3, Mask R-CNN layers, including RPN and RoI, are set to test mode: they do not compute any losses. Therefore, RoI uses the NMS threshold to filter predictions. A larger number of overlapping positive predictions can prompt the model to learn to associate them with a particular class, e.g., more prevalent in COVID-19 than common pneumonia. If the NMS threshold is low, the model will have to learn to associate a small number of distant predictions with the particular condition, which is likely to be a more challenging problem because of the similarities between COVID-19 and common pneumonia. RoI score q is set to -0.01 to accept all predictions regardless of confidence score, to keep the input size in the classification module S of fixed size. The classification model ***S*** architecture details (including the batch to feature conversion) are presented in Fig. 2 and Fig. 3 and [13], and its hyperparameters in Table 2.

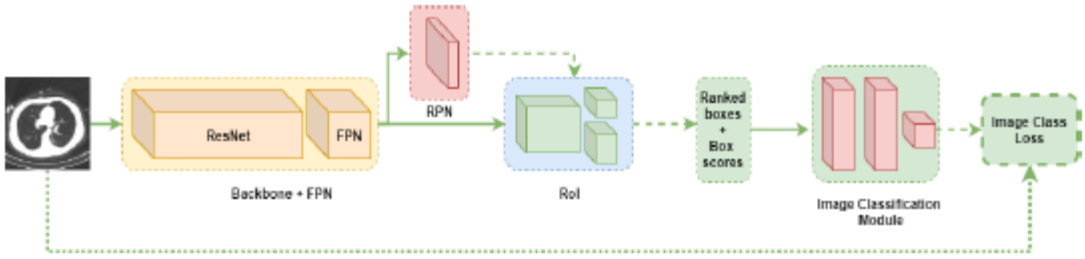

Fig. 2. Architecture of the COVID-CT-Mask-Net classification model.

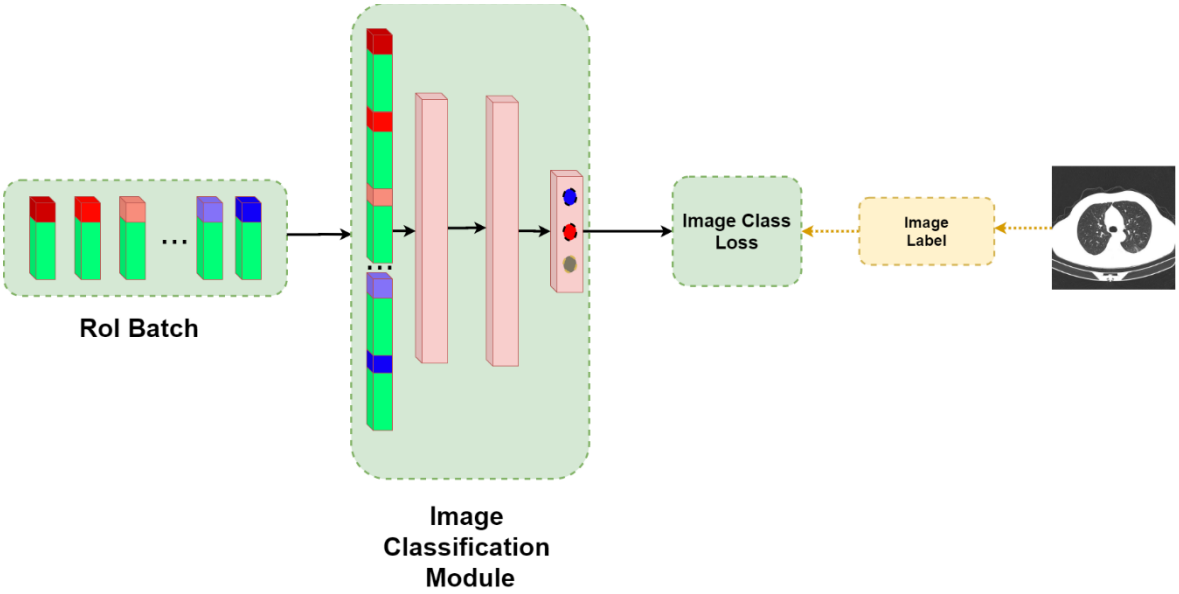

Fig. 3. Architecture of the Image Classification Module S.

Table 2. Key hyperparameters of COVID-CT-Mask-Net. Backbone network, anchor scales and sizes are the same as in Table 1. Both RPN and RoI modules are set to the test mode. RoI scoreθ is set to -0.01 to accept all predictions, even with low scores, to maintain the fixed batch size that is passed to the classification module S. The value of S is the total number of trainable parameters in it.

Backbone	Anchor Sizes	Anchor Scales	RPN NMS_θ	RoI NMS_θ	RPN Output	RoI Output/ Batch Size	RPN $Score_\theta$	RoI $Score_\theta$	Classifier Module S
ResNet50 +FPN	$2^{2:5}$	0.1, 0.25, 0.5, 1, 1.5, 2	0.25/0.75	0.25/0.75	1000	256	-	- 0.01	2.26M

Table 3. Average precision of segmentation models. Best results in bold.

Model	AP@0.5	AP@0.75	AP@[0.5:0.95]
Only GGO mask + NMS@0.25	0.4575	0.3777	0.3542
Only GGO mask + NMS@0.75	0.4588	0.3982	0.3610
Merged mask + NMS@0.25	0.5682	0.4134	0.4310
Merged mask + NMS@0.75	**0.6192**	**0.4522**	**0.4468**
Separate mask + NMS@0.25 [13]	0.4741	0.3895	0.3641
Merged mask + NMS@0.75 [13]	0.5020	0.4198	0.3871

3. Results and Discussion

Each segmentation model was trained using Adam optimizer with the same learning rate of 1e-5 and weight regularization coefficient 1e-3 for 100 epochs. The best models for each configuration are reported in Table 3. Training of each model took about 3h on a GPU with 8 Gb VRAM. All classifiers were trained with the same configuration for 50 epochs, which took about

8 hours on the same GPU. The sizes of the models are presented in Table 4. The difference in size between all segmentation models presented here is minuscule (< 1000 parameters). The architecture and the training of models with separate masks are the same as in [13]. The only difference that explains better results in Table 3, Table 5 dan Table 6 is due to the removal of tiny objects (less than 10 10 pixels) and reduction of unnecessary large sample sizes during the training of RPN and RoI, from 1024/1024 in [13] to 256/256 in this paper.

Table 4. Comparison of the models' sizes and data splits used to training, validation and testing. The number of the trainable parameters in COVID-CT-Mask-Net is due to the fact that we only train the module S and batch normalization layers in the backbone.

Model	Total #parameters	#Trainable parameters	Training	Validation	Test	Ratio Test/Train
Mask R-CNN	31.78M	31.78M	600	150	-	-
COVID-CT-Mask-Net	34.14M	2.36M	3K	20.6K	21.1K	7.06
COVIDNet-CT (best) [10]	1.8M	1.8M	60K	20.6K	21.1K	0.353
COVNet [9]	25.61M	25.61M	3K	370	438	0.129
ResNet18 [8]	11.69M	11.69M	528		90	0.17

Table 5. Sensitivity (specificity) and overall accuracy results on COVIDx-CT test data (21192 images). Best results in bold.

Model	COVID-19	Pneumonia	Normal	Overall
Only GGO mask + NMS@0.25	93.39% (95.73%)	95.27% (93.08%)	97.30% (97.95%	95.77%
Only GGO mask + NMS@0.75	86.98% (92.26%)	94.27% (69.70%)	71.12% (94.75%)	82.45%
Merged mask + NMS@0.25	93.56% (97.92%)	**97.20% (90.99%)**	95.12% (98.34%)	95.52%
Merged mask + NMS@0.75	92.68% (96.29%)	96.69% (93.63%)	**97.74% (98.54%)**	**96.33%**
Separate mask + NMS@0.25	92.22% (95.51%)	93.06% (90.11%)	95.15% (96.08%)	93.82%
Merged mask + NMS@0.75	**93.88% (95.88%)**	95.06% (93.00%)	96.91% (97.66%)	95.66%
COVID-CT_Mask-Net (best) [13]	90.80% (94.75%)	91.62% (87.07%)	91.10% (94.33%)	91.66%

Table 6. Comparison to other models. The results for COVIDNet-CT were obtained by running the publicly available model (https://github.com/haydengunraj/COVIDNet-CT) on the same test split using inference method and differs from the one reported in the publication, results for the other two models are taken from the publication. Last column is the share of COVID observations in the test split. Test split for COVNet has 438 images, ResNet18 90 images.

Model	COVID-19 Sensitivity	Overall Accuracy	COVID Prevalence
Ours(best COVID-19 Sensitivity)	**93.88%**	95.64%	20%
Ours (best Overall Accuracy)	92.68%	96.33%	20%
COVID-CT_Mask-Net [13]	90.80%	91.66%	20%
COVIDNet_CT (best) [10]	92.48%	**97.57%**	20%
COVIDNet [9]	90.00%	89.04%	30%
ResNet18 [8]	81.30%	86.70%	35.79%

To measure the accuracy of the segmentation models, we use the average precision (AP), a benchmark tool for datasets labeled at an instance level like MS COCO [14] and Pascal VOC [15]. We adapt the MS COCO convention and report values for three thresholds: AP@0.5, AP@0.75, and AP (primary challenge metric). The first two use Intersect over Union (IoU) between predicted and ground-truth bounding boxes with thresholds equal to 0.5 and 0.75. The latter averages over thresholds between 0.5 and 0.95 with a 0.05 step (a total of 10 thresholds). For details, see [14]. We adapt the implementation of average precision computation from https://github.com/matterport/Mask_RCNN. The confidence threshold for considering the object (RoIθ hyperparameter) is 0.75 across all models. Only predictions with confidence scores >RoIθ are considered for computing (m)AP. The rest are discarded. RoI NMSθ is always the same as the RPN.

Images in Fig. 4 illustrate the difference between the two NMS thresholds across all mask types. Each column corresponds to a particular CT scan slice. The bottom row is the ground

truth masks with both segmented lesion regions. Rows 1,3,5 are models that use an NMS threshold of 0.25, rows 2,4,6 use an NMS threshold of 0.75. Rows 1,2 are models that were trained only with the GGO mask. Models in rows 3,4 were trained with merged masks. Models in rows 5,6 were trained using both masks. Models with a higher NMS threshold produce a larger number of predictions overall (except, for example, in Fig. 4. (e), the models with the merged GGO and C masks, row 3 with low NMS, and row 4 with high NMS), many of them overlapping. It is a consequence because a particular predicted region can have a high enough RPN confidence score to be passed on to RoI. However, then RoI classification 'head' outputs a confidence score lower than the RoI score. Hence that region will be classified as background. In a low NMS, an overlapping prediction with a slightly lower score would be discarded at the RPN stage. The high NMS would be added to the pool of predictions, and RoI could extract a confidence score exceeding the RoI score from this second prediction. Therefore, models with high NMS produce more predictions overall, both true and false positives.

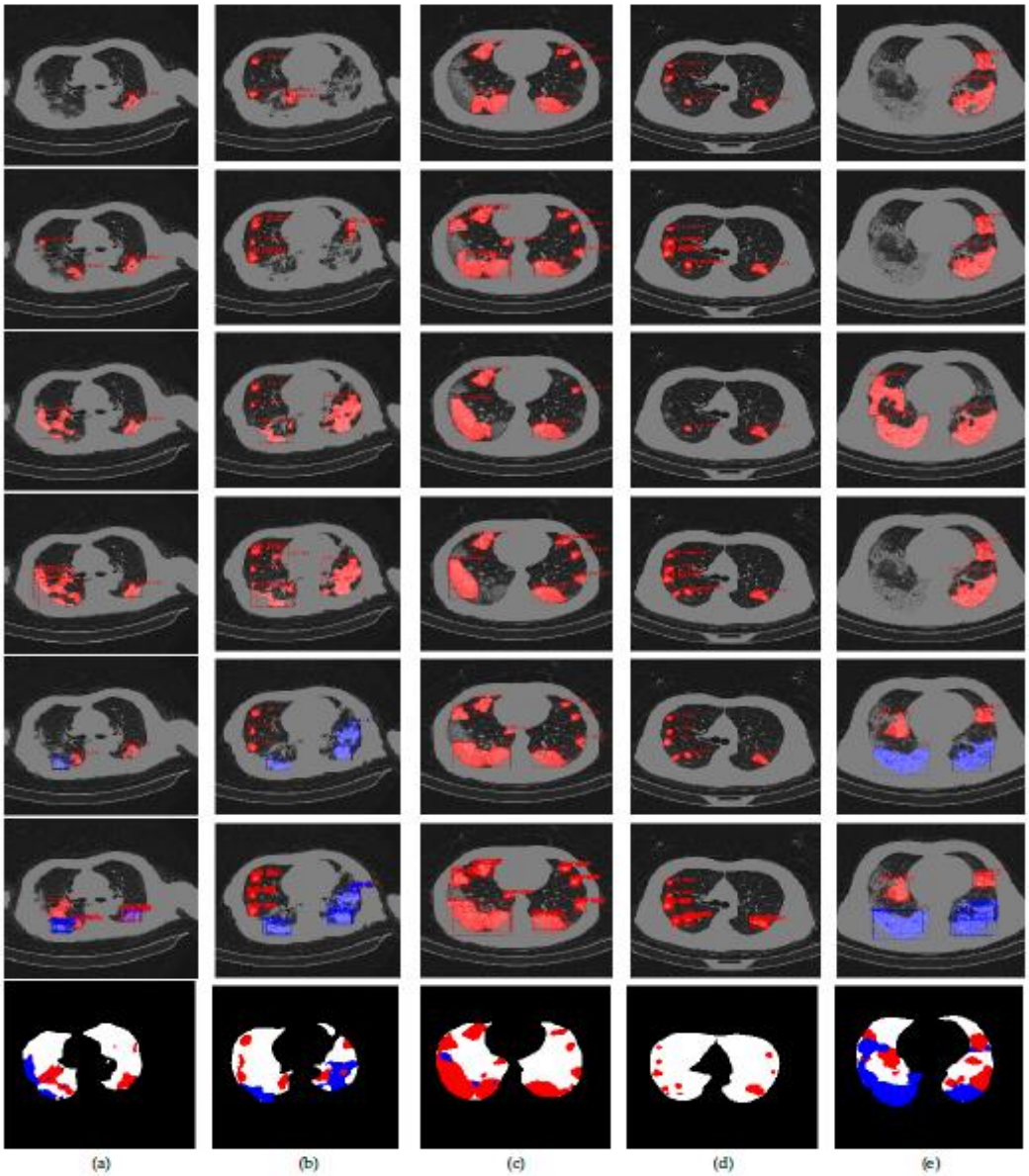

Fig. 4. Predicted masks for a number of CT scans. Row 7: ground truth masks, red: GGO, blue: C. Rows 1,3,5: models with NMS=0.25. Rows 2,4,6: models with NMS=0.75. Rows 1,2: models trained only with the GGO mask, Rows 3,4: models trained with the merged GGO and C masks. Rows 5,6: models trained with separate masks for both classes. All mask predictions are overlaid with bounding boxes and RoI confidence scores. Best viewed in color.

Evaluation results of the segmentation model are summarized in Table 3. Models using a high NMS threshold of 0.75 outperform the ones with a low NMS threshold of 0.25 across all mask types. The model that learns from merged GGO and C masks with high NMS confidently outperforms GGO-only at every IoU threshold level. Apart from the NMS effect described above, GGO and C areas in CT scans have many commonalities, so if the model learns to segment GGO only, then Consolidation and background have the same label. As a result, the model associates some important patterns with the background rather than the object class. Results for separate GGO and C masks are mostly better than for GGO but worse than for the merged masks. We explain this because overall, C is not very well represented in the dataset (see [13] for details of the data analysis). Therefore the model often confuses it with GGO features or fails to learn certain essential features because of their under-representation in the data.

The COVID-CT-Mask-Net evaluation results are presented in Table 5, and the comparison of the best models we trained (highest COVID sensitivity and highest overall accuracy) in Table 6. All results are a significant improvement over the baseline COVID-CT-Mask-Net model in [13], which we beat by 3.08% (COVID sensitivity) and 5.10% (overall accuracy). Comparing the segmentation and classification results, though, the advantage of the segmentation models learning from merged masks does not immediately translate into the advantage for solving the classification problem. Overall, the classifiers derived from these models are slightly better than the classifiers derived from the segmentation models for two classes and noticeably better than GGO-only models. This advantage is much smaller than the gap in the AP and mAP metrics for the corresponding segmentation problems.

4. Conclusion

This paper compared some Mask R-CNN models that detect and segment instances of two types of lesions in chest CT scans. We established that merging lesion masks for Ground Glass Opacity and Consolidation into a single lesion mask dramatically improves the predictive power and the precision of the instance segmentation model compared to other approaches. We extended these models to predict COVID-19, common pneumonia, and control classes using COVID-CT-Mask-Net architecture. On a sizeable COVIDx-CT dataset (21192 chest CT scan slices), the classification model derived from the best segmentation model achieved the COVID-19 sensitivity of 92.68% and overall accuracy of 96.33%, and the model derived from the segmentation model using both masks achieved a COVID-19 sensitivity of 93.88% and an overall accuracy of 95.64%. The source code and the pre-trained models are available on https://github.com/AlexTS1980/COVID-CT-Mask-Net.

References

[1] T. Yan, P. K. Wong, H. Ren, H. Wang, J. Wang, and Y. Li, "Automatic distinction between COVID-19 and common pneumonia using multi-scale convolutional neural network on chest CT scans," *Chaos, Solitons & Fractals*, vol. 140, p. 110153, Nov. 2020, doi: 10.1016/j.chaos.2020.110153.

[2] X. Li, X. Fang, Y. Bian, and J. Lu, "Comparison of chest CT findings between COVID-19 pneumonia and other types of viral pneumonia: a two-center retrospective study," *Eur. Radiol.*, vol. 30, no. 10, pp. 5470–5478, Oct. 2020, doi: 10.1007/s00330-020-06925-3.

[3] W. Zhao, Z. Zhong, X. Xie, Q. Yu, and J. Liu, "CT scans of patients with 2019 novel coronavirus (COVID-19) pneumonia," *Theranostics*, vol. 10, no. 10, pp. 4606–4613, 2020, doi: 10.7150/thno.45016.

[4] D. Zhao *et al.*, "A comparative study on the clinical features of coronavirus 2019 (COVID-19) pneumonia with other pneumonias," *Clin. Infect. Dis.*, vol. 71, no. 15, pp. 756–761, Jul. 2020, doi: 10.1093/cid/ciaa247.

[5] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: a CT scan dataset about COVID-19," *arXiv Prepr. arXiv2003.13865*, 2020. Available: arXiv.

[6] K. Zhang *et al.*, "Clinically applicable AI System for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423-1433.e11, Jun. 2020, doi: 10.1016/j.cell.2020.04.045.

[7] Y. Song *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *medRxiv*, doi: 10.1101/2020.02.23.20026930.

[8] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Appl. Intell.*, pp. 1–7, Apr. 2020, doi: 10.1007/s10489-020-01714-3.

[9] L. Li *et al.*, "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, Aug. 2020, doi: 10.1148/radiol.2020200905.

[10] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images," *arXiv Prepr. arXiv2009.05383*, 2020. Available: arXiv

[11] Y.-H. Wu *et al.*, "JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation," *arXiv Prepr. arXiv2004.07054*, 2020. Available: arXiv

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[13] A. Ter-Sarkisov, "COVID-CT-Mask-Net: Prediction of COVID-19 from CT scans using regional features," *medRxiv*, 2020, doi: 10.1101/2020.10.11.20211052.

[14] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Eur. Conf. Comput. Vis.*, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.