2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340181

Quantitative Assessment of COVID-19 Lung Disease Severity: A Segmentation-based Approach*

Edward P. Booker, Mehdi Paak, Mohammadreza Negahdar

***Abstract*— We present the use of mean Hounsfield units within lungs as a metric of disease severity for the comparison of image analysis models in patients with COPD and COVID. We used this metric to assess the performance of a novel 3D global context attention network for image segmentation that produces lung masks from thoracic HRCT scans. Results showed that the mean Hounsfield units enable a detailed comparison of our 3D implementation of the GC-Net model to the V-Net segmentation algorithm. We implemented a biomimetic data augmentation strategy and used a quantitative severity metric to assess its performance. Framing our investigation around lung segmentation for patients with respiratory diseases allows analysis of the strengths and weaknesses of the implemented models in this context.**

***Clinical Relevance*— Mean Hounsfield units within the lung volume can be used as an objective measure of respiratory disease severity for the comparison of CT scan analysis algorithms.**

I. Introduction

In this study, we aimed to develop a lung segmentation for patients with severe respiratory diseases. The relevant datasets used were annotated thoracic CT scans from COVID-19 and chronic obstructive pulmonary disease (COPD) patients. Labeled CT data are expensive and time consuming to produce, and as such the publicly available datasets are small and inconsistently labeled. Therefore, there is a need for methods to compensate for limitations of available datasets with biologically-inspired data augmentation e.g., Sousa et al.[1] There has been criticism of using generated data for medical imaging applications, however. Chang [2] and Vallon et al. [3] discussed the limitations of data augmentation strategies in medical image analysis and state that clinical expertise is needed to guide training and determine relevant features for model training.

Image segmentation models are useful for aiding medical decision making, accelerating diagnosis, and patient treatment planning [4-8]. Many reports on segmentation models based on U-Net [9] and V-Net [10] architectures have shown promising performance [11-14]. Of particular interest are models utilizing the attention mechanism to aid image segmentation, such as UNETR [15].

One version of the attention mechanism, global context, is calculated by taking a weighted average from all positions in an image embedding [16]. This weighted average can be used to weight an input from another layer. This allows lower-level features to be affected by broader context; we refer to this as the global context attention mechanism (GCA). GCA was incorporated into a segmentation algorithm, GC-Net, which was demonstrated in 2D by Ni [17]. We implement this method for 3D image segmentation and compare it to the state-of-the-art V-Net model. GC-Net was chosen as a promising non-transformer approach to the attention mechanism in image segmentation.

There have been many other reports of 2D segmentation on CT slices from COVID patients [14, 18, 19]. Of particular note is the model produced by Hofmanninger et al. [14], which is well-documented, is transparent over data sources, and performs with 98% Dice. Despite its performance on even severe COVID patients, this model does not correlate well between layers in CT volumes. This lack of correlation leads to jagged edges in dimensions perpendicular to the scan direction and under segmentation sandwiched between segmented layers.

3D segmentation models can correlate between layers. Successful 3D segmentation models have been reported in COVID scans [19, 20], but there is a lack of shared models, diverse performance statistics, under-reporting of performance on severe cases, and a quantification or qualification of what severe means. These deficits make it challenging to compare models. Lung segmentation masks for COVID, either severe or mild, are scarce.

Established methods to assess the severity of respiratory disease via CT scans using standardized scoring systems, such as those proposed by van der Ven et al. [21] and Brody et al. [22] rely on subjective assessments, expert annotation or electronic medical record data for objective scoring. In this study, we suggest a novel objective severity score that does not require expert annotation, and can be applied across diverse datasets when there are differences in clinical data.

High X-ray attenuation in CT scans is observed in lungs of the most severe patients with COVID [23] or pneumonia [24]. X-ray attenuation is measured in Hounsfield Units (HU). Yamada found that high HU alone did not predict patient outcomes, and therefore it can only be used as an approximate measure of patient severity [23]. This observation enables the use of mean HU in the lungs as a measurement of patient severity to compare model performance, but not to predict outcomes.

In this report we:

- Implemented two model architectures: V-Net and a novel 3D implementation of GC-Net
- Implemented novel biomimetic data augmentation to improve segmentation on severe cases.
- Quantitatively compared the performances of these models with respect to disease severity

II. Methods

All models were trained using 216 CT scans from the COPD [25] CT dataset and 70 scans from a Genentech dataset of COVID patients. The mean HU of this dataset and

[1]*This work was supported by Genentech, Inc. All authors are employees of Genentech, Inc. and shareholders in F. Hoffmann La Roche, Ltd. This research was conducted using human subject data available from previous clinical trials. Ethical approval was not required as confirmed by the license attached to the original studies.

E. Booker, M. Paak, M. Negahdar are with Genentech, 600 East Grand Ave, South San Francisco, CA 94080 (Phone: 236-979-0846; E-mail: edward.bookere@contractors.roche.com).

This work is licensed under a Creative Commons Attribution 3.0 License.
For more information, see http://creativecommons.org/licenses/by/3.0/

the training, test, and validation splits are shown in Fig. 1a. Images were stored in the nifti format, voxel values were normalized in HU. These two diseases are complementary for training lung segmentation models, as both can cause consolidation (areas of significantly increased attenuation). Consolidation on the periphery of the lung can obscure the lung boundary and make the boundary hard to detect.

Some COVID cases can have attenuation values as high as -100 HU in the lungs (Fig. 1b), in comparison to values of -800 to -600 HU in healthy lungs [23]. Severe COVID data is scarce and model exposure for training presents a challenge. Fig. 1a shows that the mean HU value in the vast majority of the dataset is less than -600 HU. COVID cases are shown in lighter colors.

To make up for the scarcity of severe COVID data, we implemented data augmentation to mimic the trends seen in COVID cases. Fig. 1 shows the changes of HU values as scans progress through the lungs of COVID patients of a variety of severities. Moderate-to-severe COVID cases have a constant offset in HU values from the baseline, often with a hump in values at some point in the scan. Our augmentation method adds a random offset to raw lung image data and a Gaussian hump centered within the lungs of the patient (Fig. 2a) as well as patches of pathological tissue sampled from a publicly available annotated dataset of 100 individual CT slices from COVID patients [26]. This type of domain-knowledge-based data augmentation has been successfully implemented [27], using patches from other data samples [28] or GANs [29] to generate appropriate synthetic data.

We did an ablation study to compare the performance of models with or without this augmentation strategy, and with no data augmentation. In both augmentation approaches, generic augmentation strategies are used (rotation, gaussian noise, patch swapping, etc.). This method cannot be used for identification of pathological tissue as the biomimetic augmentation will alter the characteristics of those tissues.

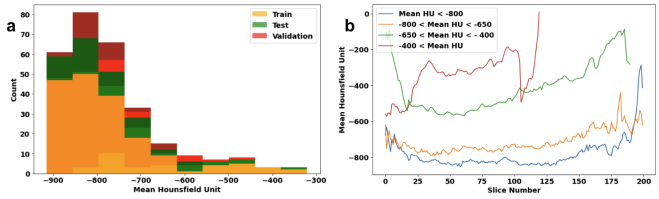

Figure 1. a) Distribution of average HU value across lung volumes in the train, test, and validation datasets. Light colors show COVID patients. b) Variation in mean HU values of slices along the height of scans. The scan for the patient in red (-400 < mean HU) has fewer slices than the others.

The models used in this study have similar architectures. An encoder path that compresses a 3D image array into an embedding and a bottleneck layer. The embedding then gets decoded to the original image size. The encoder and decoder layers of the same size are connected by skip connections (Fig. 2b). The two models used are V-Net and a novel 3D implementation of the GC-Net architecture.

The loss function was a linear combination of Dice and cross entropy as implemented in Monai [30]. In GC-Net the bottleneck layer is a squeeze-and-excitation spatial pyramid pooling (SEPP) layer, and the encoder layers are GCA layers followed by upsampling convolutional layers. The SEPP layer (with original expansion rates) and GCA layers were implemented in 3D by substitution of the layers in Ni et al. [21] with their 3D equivalents. The skip connections include a buffer layer.

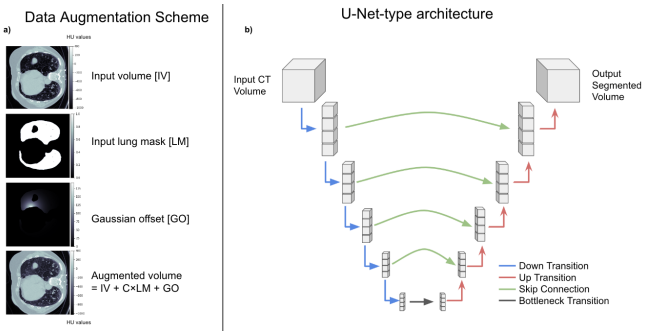

Figure 2. Model and biomimetic augmentation method schematics. a) Steps in data-inspired data augmentation. A Gaussian peak and a constant offset are added within the input lung mask. The texture-based augmentation is not shown here. b) Architecture. Blue arrows indicate down transitions, red arrows up transitions, black the bottleneck transition and green the skip connections.

For GC-Net, the learning rate used was 1×10^{-3}, the dropout rate was 0.6, the cross-entropy/Dice ratio in the loss function was 0.4, and the training weight decay was 1×10^{-5}. The V-Net implementation by Monai [30] is used here. The V-Net learning rate used was 2×10^{-3}, dropout rate was 0.6, the cross-entropy to Dice ratio in the loss function was 0.4, and the training weight decay was 1×10^{-5}.

The GC-Net model has 122,113,196 parameters, whereas the V-Net model has 45,597,898. All models were compared after training for 23 epochs with fixed random seeds.

For model comparison identical postprocessing was applied: the largest connected components in the central 50% of the volume were taken as the lung mask (components smaller than 10% of the largest component are ignored), and 3D holes were filled.

III. Results

Variations in COVID patient vital signs with the mean HU in the lungs calculated from a thoracic CT scan taken on the same day as the measured vital signs are shown in Fig. 3. The correlation between oxygen saturation and mean HU in the lungs was the strongest, which is expected as a reduction in lung function from increased fluid would be expected. The relationships between mean HU and both pulse and respiratory rates was weak, but both confirm the correlation between the severity of the state of a patient and their CT scan. These correlations do not suggest the mean HU could be used to predict patient outcomes, but we take this as justification to use the mean HU to compare model performance.

The performance of the biomimetic data augmentation, basic augmentation, and no data augmentation models are shown in Fig. 4 and Table 1.There was a drop-off in model performance as the mean HU in the lung volume increases for all the models. Comparing Fig. 4c and 4e, except for some lower cases at low mean HU, the GC-Net model

maintained performance as mean HU increases better than the V-Net model.

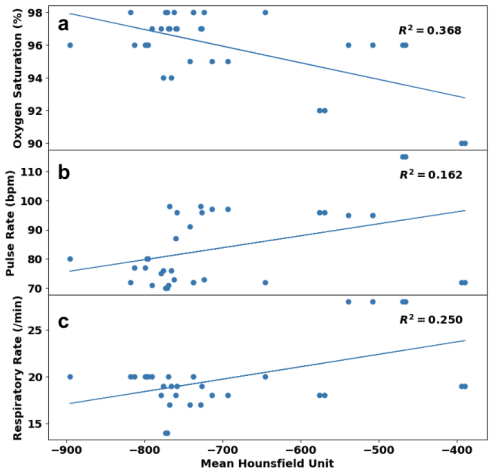

Figure 3 a)-c) Graphs of variation in oxygen saturation, pulse rate, and respiratory rate with mean HU in patient scans.

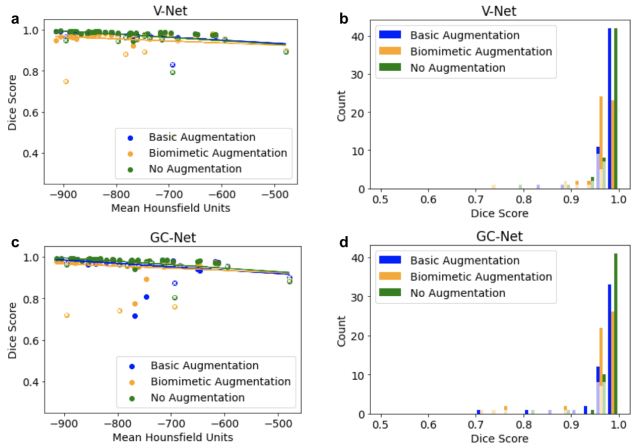

Figure 4. Model performance statistics for the different architectures investigated. a) and c) show how the Dice score varies with mean HU value within the lungs for V-Net and GC-Net respectively. Points represent individual image results, points with white crosses are COVID patients. Straight lines are lines of best fit. b) and d) show histograms of Dice scores for V-Net and GC-Net respectively.

TABLE I. MODEL PERFORMANCE STATISTICS

Model - Augmentation	***Dice (%)*[a]**	***Covid Dice (%)*[b]**	***Severe Dice (%)*[c]**
GC-Net - No Aug.	97.6 ± 0.4	93.9 ± 1.6	92 ± 2
GC-Net - Basic Aug.	96.6 ± 0.7	**94.9 ± 1.0**	**93 ± 2**
GC-Net -Biomimetic Aug.	95.4 ± 0.8	89 ± 3	92 ± 2
V-Net - No Aug.	97.6 ± 0.4	93.5 ± 1.5	92 ± 2
V-Net - Basic Aug.	**97.7 ± 0.3**	94.7 ± 1.3	**93 ± 2**
V-Net -Biomimetic Aug.	95.4 ± 1.0	88 ± 4	**93 ± 2**

a. The mean Dice score for all samples. b. The mean Dice score for all Covid samples. c. The mean Dice score for the two Covid samples over -600 mean HU. The error is the standard error in the mean.

The V-Net model yielded the highest performance averaged over all cases (Table 1). The GC-Net model achieved comparably high performance (within the standard error in the mean performance) to the best V-Net implementation. It can further be seen that GC-Net with basic augmentation achieved the highest performance in COVID cases, with over 97% Dice for lung segmentation averaged over all patients.

A comparison of results from various data augmentation methods showed that the biomimetic method did not provide an improvement in average score in either of the models investigated. The performances on both cases above -600 mean HU suggest that biomimetic augmentation was able to generalize better. Further, there was a smaller drop in Dice score as the mean HU increases going from no augmentation to basic augmentation to biomimetic augmentation. This type of comparison requires a quantitative measurement of patient severity. It is noted that the scarcity of patients above -600 HU makes these conclusions tentative.

Fig. 5 shows select qualitative results comparing the performances of the models studied on different severity patients. Fig. 5a is -761 HU and Fig. 5d is -479 HU. GC-Net preserves more of the shape of the segmented tissue than V-Net, and also identifies areas of higher HU values than the V-Net model, but misses some components.

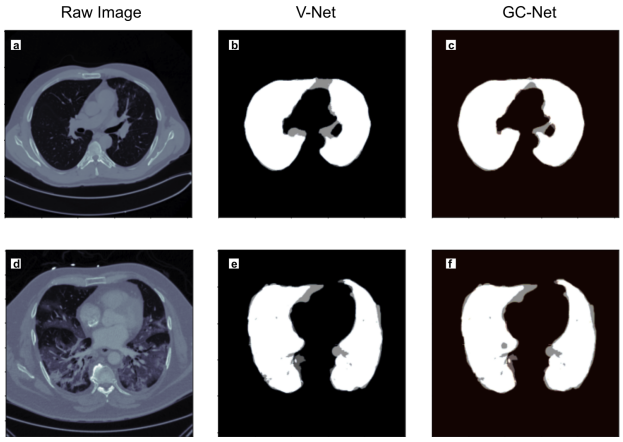

Figure 5. Qualitative results from models trained with biomimetic augmentation. Rows show results on the same volume, columns show results on the same model. Pink backgrounds show ground truth labels.

IV. DISCUSSION

Fig. 4 and Table 1. suggest that the use of this domain-inspired data augmentation may be a useful tool that can be implemented in scenarios where edge cases (peripheral consolidation in our example) are scarce and generic data augmentation is insufficient to allow the models to generalize, although it does weaken performance at lower cases. This weakening may be due to this larger model requiring significantly more training with such augmented data, or due to reduced exposure to low mean HU patients from the augmentation.

Fig. 5 shows that highly consolidated regions of the lungs may be under segmented in all models. The relatively low drop-off in performance of the GC-Net model as severity increases suggests that GC-Net is better able to generalize than V-Net. We see qualitatively from Fig. 5, and quantitatively in Fig. 4 and Table 1, that GC-Net generalizes better to higher mean HU than V-Net, but with poorer results on average. Fig. 5 shows that highly consolidated regions

are penetrated better by GC-Net, and Fig. 4 shows that there is a smaller dropoff in performance as we increase the mean HU in the lung to be segmented. This ability to generalize better may be due to the SEPP bottleneck layer in GC-Net preserving more spatial information, or due to the global context in the GCA layers. Ni [21] suggested that the GCA in their model allows longer ranged associations to be formed between regions of the image. This can be seen in Fig. 5, where segmented regions are more continuous than in the other model. Inspecting how the results vary with domain-relevant data is critical. We can only tell that the drop-off in Dice with mean HU is lower in the GC-Net model compared to the V-Net by inspecting the mean HU distribution (Fig. 4).

This difference in performance loss with mean HU between the methods used here demonstrates the importance of comparing model performance with a continuous variable. Further, our clinical results validate the use of mean HU as a way to assess how lung segmentation model performances vary. In contexts where there is a variable parameter that may be used to compare model performance across the domain of interest this should be done and justified using domain-specific understanding. This better identifies model performance trends in a way that arbitrary bins, such in Table 1 (COVID or severe COVID), do not.

V. Conclusion

This study confirms that domain-specific knowledge is needed to build robust, generalizable models and the data augmentation schemes for training these models. Model performance should be compared continuously across the training and testing domain if possible, such as by utilizing a quantitative metric for patient severity.

References

[1] A. M. Sousa et al., Improving Automated Lung Segmentation in Ct Images by Adding Anomalies Adjacent to the Pleura, 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022, pp. 1-5,

[2] E. Y. Chang, Knowledge-Guided Data-Centric AI in Healthcare: Progress, Shortcomings, and Future Directions, https://arxiv.org/abs/2212.13591

[3] J. J. Vallon et al., Patient-Level Clinical Expertise Enhances Prostate Cancer Recurrence Predictions with Machine Learning, medRxiv 2022.03.22.22272635

[4] N. Sharma, L. M. Aggarwal, Automated medical image segmentation techniques. J Med Phys. 2010 Jan;35(1):3-14.

[5] F. Renard, et al. Variability and reproducibility in deep learning for medical image segmentation. Sci Rep 10, 13724 (2020)

[6] A, Mezer, et al. Quantifying the local tissue volume and composition in individual brains with magnetic resonance imaging. Nat Med 19, 1667–1672 (2013).

[7] M. Silveira et al., Comparison of Segmentation Methods for Melanoma Diagnosis in Dermoscopy Images, IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 1, pp. 35-45, (2009)

[8] V. Fortunat et al., Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling. Med. Phys., 40: 07190 (2013)

[9] O. Ronneberger et al., U-Net: Convolutional Networks for Biomedical Image Segmentation, https://arxiv.org/abs/1505.04597

[10] F. Milletari et al., V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation https://arxiv.org/abs/1606.04797

[11] NIHR Themed Review: Living with Covid19 - Second review; March 2021

[12] S. Walvekar, S. Shinde, Efficient Medical Image Segmentation Of COVID-19 Chest CT Images Based on Deep Learning Techniques, 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), (2021)

[13] K. Gao et al., Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images,Medical Image Analysis, vol. 67, (2021)

[14] J. Hofmanninger et al., Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 4, 50 (2020).

[15] A. Hatamizadeh et al., UNETR: Transformers for 3D Medical Image Segmentation https://arxiv.org/abs/2103.10504

[16] Y. Cao et al., GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond https://arxiv.org/pdf/1904.11492

[17] J. Ni et al., GC-Net: Global context network for medical image segmentation, Computer Methods and Programs in Biomedicine, vol. 190, (2020)

[18] S. Tilborghs et al., Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients https://arxiv.org/pdf/2007.15546.pdf

[19] S. J. Yoo et al., Automated Lung Segmentation on Chest Computed Tomography Images with Extensive Lung Parenchymal Abnormalities Using a Deep Neural Network. Korean J Radiol. 22(3):476-488 (2021)

[20] Y. Qiblawey et al., Detection and Severity Classification of COVID-19 in CT Images Using Deep Learning, Diagnostics, vol. 11, no. 5, p. 893, (2021)

[21] A.A.J.M. van de Ven, et al. A CT Scan Score for the Assessment of Lung Disease in Children With Common Variable Immunodeficiency Disorders, Chest, 138, 371-379 (2010)

[22] A.S. Brody et al., Reproducibility of a Scoring System for Computed Tomography Scanning in Cystic Fibrosis, Journal of Thoracic Imaging 21(1):p 14-21, (2006.)

[23] D. Yamada et al. Visual classification of three computed tomography lung patterns to predict prognosis of COVID-19: a retrospective study. BMC Pulm Med 22, 1 (2022).

[24] P. Konietzke et al., Consolidated lung on contrast-enhanced chest CT: the use of spectral-detector computed tomography parameters in differentiating atelectasis and pneumonia. Heliyon. 26;7(5):e07066 (2021)

[25] COPDGENE: A study to investigate the underlying genetic factors of Chronic Obstructive Pulmonary Disease. Supported by the National Institutes of Health (NHLBI U01 HL089897 and U01 HL089856) and by the COPD Foundation.

[26] http://medicalsegmentation.com/covid19/ (accessed February 2023)

[27] W. Wang et al., Exploring Cross-Image Pixel Contrast for Semantic Segmentation, https://arxiv.org/pdf/2101.11939.pdf

[28] C. N. Vasconcelos, B. N. Vasconcelos, Convolutional Neural Network Committees for Melanoma Classification with Classical and Expert Knowledge Based Image Transforms Data Augmentation https://arxiv.org/pdf/1702.07025.pdf

[29] H. Li, X. Zhang, Q. Tian and H. Xiong, "Attribute Mix: Semantic Data Augmentation for Fine Grained Recognition," 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 243-246, (2020)

[30] https://docs.monai.io/en/stable/networks.html (accessed February 2023)