Multi-class semantic segmentation of pediatric chest radiographs

Gregory Holste[a,d], Ryan Sullivan[b,d], Michael Bindschadler[c], Nicholas Nagy[c], Adam Alessio[c,d]

[a]Mathematics & Statistics, Kenyon College, Gambier, OH, USA
[b]Computer Science and Statistics, Purdue University, West Lafayette, IN, USA
[c]Radiology, University of Washington, Seattle, WA, USA
[d]Biomedical Engineering and Radiology, Michigan State University, East Lansing, MI, USA

ABSTRACT

Chest radiographs are a common diagnostic tool in pediatric care, and several computer-augmented decision tasks for radiographs would benefit from knowledge of the anatomic locations within the thorax. For example, a pre-segmented chest radiograph could provide context for algorithms designed for automatic grading of catheters and tubes. This work develops a deep learning approach to automatically segment chest radiographs into multiple regions to provide anatomic context for future automatic methods. This type of segmentation offers challenging aspects in its goal of multi-class segmentation with extreme class imbalance between regions. In an IRB-approved study, pediatric chest radiographs were collected and annotated with custom software in which users drew boundaries around seven regions of the chest: left and right lung, left and right subdiaphragm, spine, mediastinum, and carina. We trained a U-Net-style architecture on 328 annotated radiographs, comparing model performance with various combinations of loss functions, weighting schemes, and data augmentation. On a test set of 70 radiographs, our best-performing model achieved 93.8% mean pixel accuracy and a mean Dice coefficient of 0.83. We find that (1) cross-entropy consistently outperforms generalized Dice loss, (2) light augmentation, including random rotations, improves overall performance, and (3) pre-computed pixel weights that account for class frequency provide small performance boosts. Overall, our approach produces realistic eight-class chest segmentations that can provide anatomic context for line placement and potentially other medical applications.

Keywords: pediatric imaging, chest radiograph, U-net, multi-class segmentation, deep learning

1. INTRODUCTION

Critically ill patients in the neonatal and pediatric intensive care units often require catheters and tubes, collectively called "lines," to sustain life. Throughout their stay, these patients undergo a series of radiographs to monitor the placement of such lines; since each radiograph must be sent off and assessed by a radiologist, this is a time-consuming, labor-intensive process. We aim to automatically segment the chest into regions that will provide anatomic context for line placement or other automated clinical applications. Previous efforts have performed segmentation tasks on chest radiographs, however these efforts typically have relied on conventional, non-machine-learned methods such as non-rigid registration of normal atlases[1] or active shape models.[2] These approaches are generally effective for binary (single-class) segmentation, commonly focused on rib cage segmentation,[3,4] but have not been scaled for segmentation of a radiograph into multiple regions.

A few previous efforts have aimed to segment chest radiographs into multiple anatomic regions,[5,6] and some have used deep learning approaches.[7,8] However, to our knowledge, the application of deep learning approaches to multi-class segmentation of pediatric chest radiographs remains relatively unexplored. Along with being a unique application, this task turns out to be particularly difficult for the following reasons: (1) it involves segmentation into eight regions, and (2) it suffers from extreme class imbalance. In any given radiograph, the carina – the juncture at which the trachea splits into each bronchus – is 2-5 orders of magnitude smaller (in number of pixels) than the background region. Our work explores deep learning strategies, including unique weighted loss functions, to overcome these challenges and achieve multi-class (eight-class) segmentations that can provide anatomic context for potential future methods.

aalessio@msu.edu; http://www.egr.msu.edu/~aalessio/

2. METHODS

2.1 Data Curation and Processing

In an IRB-approved study, pediatric chest radiographs were collected from Seattle Children's Hospital. Radiographs were annotated with a custom MATLAB tool to create a set of ground truth chest segmentations. Three independent users were instructed to draw boundaries around the lungs, spine, and bottom of the thorax, then select a single pixel where the carina was located. The software smoothed all lines drawn and used these user-drawn boundaries to divide the chest into eight regions: left and right lung, left and right "subdiaphragm" (the thorax below the lungs), spine, mediastinum, carina, and background. The 469 labeled radiographs were randomly divided into training, validation, and test sets with an approximate $75:15:15$ split (see Table 1). Patients ranged from ages 0 to 20, with a strong skew toward very young patients; for example, approximately 65% of all labeled radiographs were from patients who could be considered toddlers (ages 4 and under).*

Table 1. Age breakdown of patients in training, validation, and test sets.

	Training	Validation	Test
Ages 0-4	218	49	38
Ages 5+	110	22	32
Total	328	71	70

The ground truth segmentation for each radiograph was then one-hot encoded such that it consisted of a stack of eight binary masks – one for each output class. Finally, the contrast of all radiographs were linearly normalized such that the minimum and maximum pixel values were mapped to the interval $[0, 1]$, then all radiographs and labels were resized to be 256 pixels in height and width. Specifically, each radiograph was of size $256 \times 256 \times 1$ and each ground truth label was of size $256 \times 256 \times 8$.

2.2 Network Architecture

We use a U-Net[9]-style architecture employing repeated 3×3 convolution operations and skip connections. The contracting path consisted of five downsampling steps (with depths 32, 64, 128, 256, and 512 respectively), and the expansive path consisted of mirrored upsampling steps (with depths of 256, 128, 64 and 32 filters). Each convolution block used a 3×3 filter, a stride length of 1, and was followed by ReLU activation layer and batch normalization; the same is true for each upsampling ("transposed convolution") block, except those used a stride length of 2. We use "same" padding for all 2-D convolutions so as to remove the need for cropping as in the original U-Net. This network was implemented and trained in Keras using a Tensorflow[10] backend, the Adam optimizer[11] with default learning rate 0.001, and a batch size of 2. We terminated training when the Dice coefficient on our validation set did not improve for 100 epochs, and restored the weights from 100 epochs earlier.

We kept this approximately 8.6 million-parameter architecture fixed and performed an ablation study, comparing model performance when varying the following components: loss function, pixel weighting scheme, and use of data augmentation. The two losses we consider are weighted adaptations of categorical cross-entropy (L_{CCE}) and generalized Dice loss[12] (L_{GDL}), a continuous (and differentiable) version of the Dice coefficient. Letting $n \in \{1, \ldots, N\}$ represent the index for each pixel of an input radiograph and $c \in \{1, \ldots, 8\}$ represent each output class, we have

$$L_{CCE} = -\frac{1}{N} \sum_{n=1}^{N} w_n \left[\sum_{c=1}^{8} y_{n,c} \; log(\hat{y}_{n,c}) \right] \quad \& \quad L_{GDL} = 1 - 2 \sum_{n=1}^{N} w_n \left[\sum_{c=1}^{8} \frac{y_{n,c} \; \hat{y}_{n,c}}{y_{n,c} + \hat{y}_{n,c}} \right],$$

where $w_n > 0$ is a pre-defined pixel weight, $y_{n,c} \in \{0, 1\}$ is a binary indicator as to whether pixel n belongs to class c, and $\hat{y}_{n,c} \in (0, 1)$ is the predicted probability that pixel n belongs to class c.

We introduce pixel weights in order to combat severe class imbalance introduced primarily by the carina; in any given radiograph, the carina can be 3-5 orders of magnitude smaller than the background region. This means that when $w_1 = \cdots = w_N$, the background region's contribution to our loss is thousands of times greater than that of the carina. We generate three different weight maps for every training image, each representing

*To later compare model performance by age class, we consider this partition of patients into toddlers and non-toddlers because (1) the two groups in general have different morphologies after rapid growth of the thorax region and (2) this allows the test set to be roughly equally split.

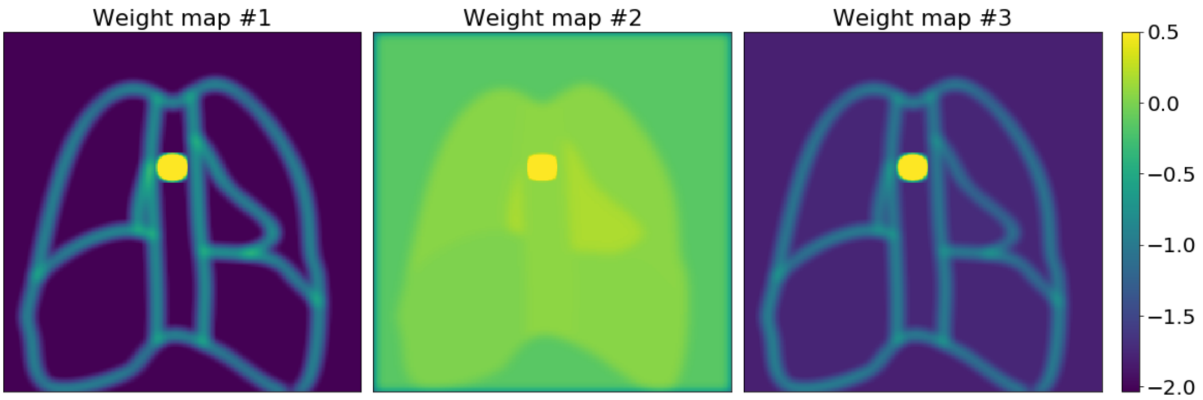

Figure 1. Example of three different weight maps created for a test set radiograph. Weight maps were natural log-transformed and "clipped" to 0.5 for aid of visualization. Actual log-transformed pixel weights around the carina were 6-8 times greater than the maximum clipped value (e.g., upwards of 3).

a different approach to emphasizing the "hard-to-learn" regions such as small classes and boundaries between classes. The first weight map produces a mask of blurred (enlarged) edges between classes, then weights those pixels by inverse class frequency; the second map generates a mask of blurred classes, then weights those pixels by inverse class frequency; the third map is like the second except it additionally weights blurred edges three times more heavily than the rest of each class. Example maps are presented in Figure 1.

Data augmentation was minimal, consisting of an elastic deformation[9,13] with $\sigma = 5$ followed by a random rotation within 15 and -15 degrees. Instead of performing augmentations "on-the-fly," we manually created and saved three transformed copies of each training image. In sum, all models that used augmentation were trained on four times as many samples as models trained on the original training set.

2.3 Evaluation and Failure Analysis

The two performance metrics of interest were accuracy, the proportion of pixels classified correctly, and the mean Dice coefficient (over all output classes), $DC = \frac{1}{8}\sum_{c=1}^{8}\frac{2|Y_c \cap \hat{Y}_c|}{|Y_c|+|\hat{Y}_c|}$, where Y_c is a ground truth binary mask for class c and $\hat{Y}_c$ is a binarized predicted mask for class c. We also consider the Dice coefficients for each class individually when further analyzing our models.

After observing that some predictions were not morphologically sound – namely, missing the carina or having disconnected regions – we subjected all predictions to a post-processing step designed to ensure the presence of all classes; this was done *after* model evaluation. In predictions with no carina present, instead of binarizing our model's outputs via a "winner-takes-all" approach (the class with the greatest probability wins), we simply lower the threshold at which the model predicts the presence of carina. Specifically, we found the threshold among $(0.25, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005)$ that produced the largest carina within a "reasonable" range of 5-75 pixels.[†] Then to prohibit the presence of multiple carinas, for all predictions, we found the largest connected component in the carina mask, set all other values in the carina mask to zero, then "filled in" those pixels with the model's next-most-confident class.

In an attempt to assess the clinical applicability of our results, we conducted a rule-based failure analysis on the post-processed predictions from our best- and worst-performing models. We defined a "failure" – a clinically unusable, morphologically inaccurate segmentation – as a predicted segmentation that (1) is missing an output class, or (2) contains disconnected regions that belong to two or more different output classes. The latter criterion refers to when either there is more than one connected component in a particular predicted mask or when a predicted mask is "interrupted" by another class in an objectively morphologically inaccurate manner. Again, post-processing was performed *after* model evaluation solely for the purpose of this comparative failure analysis.

[†]This range of reasonable carina sizes was based on the interquartile range of observed carina sizes in the training set.

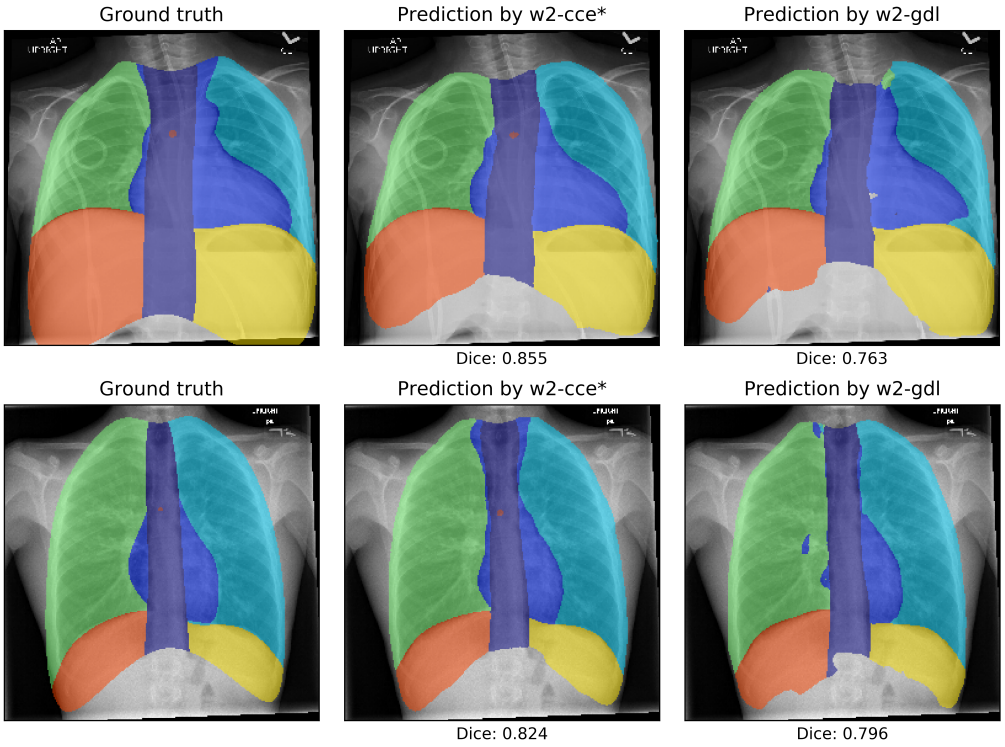

Figure 2. Representative examples of ground truth classes (left) and predicted segmentations for two patients (rows) by the best- (middle) and worst-performing (right) models as per Table 2. Mean Dice coefficient is listed below each prediction.

3. RESULTS

Table 2 summarizes overall model performance on the test set when altering the training loss function and use of data augmentation. We observe that categorical cross-entropy consistently outperformed generalized Dice loss and that data augmentation was always beneficial. Training on the sum of L_{CCE} and L_{GDL} was also one of the stronger approaches without introducing pixel weights.

Figure 2 shows predictions from the best- and worst-performing models for two radiographs in the test set. We see that the carina is missing from the predictions made by the worst model (w2-gdl), but present in those from the best model (w2-cce*). Furthermore, the boundaries between regions appear smoother and more realistic when trained on w2-cce with augmentation. The top row of Figure 3 shows the distribution of mean Dice coefficients on the test set for the same two best- and worst-performing models. Similarly, the bottom row summarizes the Dice coefficients by region for our best- and worst-performing models. We see that the Dice coefficient for the carina is always 0 when training on w2-gdl, but varied widely with a median near 0.35 when training on w2-cce with augmentation.

Example cases of qualitatively failed segmentations are shown in Figure 4, and failures are summarized in Table 3 for three different models. We see that predictions from w2-gdl never included a carina, even after drastically lowering the threshold at which the carina class "wins out" over other classes; in fact, all models trained with L_{GDL} failed to locate the carina. In contrast, models trained with L_{CCE} – even when unweighted and without augmentation – almost always predicted a reasonably sized carina after post-processing. Finally, even the best-performing model contains several disconnected class predictions after the carina post-processing step.

Table 2. Comparison of mean Dice coefficient on test set ($n = 70$) for different loss functions.

Loss Function	Dice
w2-cce*	0.832
w3-cce*	0.830
w1-cce*	0.829
cce + gdl*	0.828
cce*	0.828
w2-cce	0.825
cce + gdl	0.821
w3-cce	0.820
cce	0.815
w1-cce	0.814
w3-gdl	0.788
w1-gdl	0.787
gdl	0.784
w2-gdl	0.783

"*" = with data augmentation, "cce" = categorical cross-entropy, "gdl" = generalized Dice loss, "w<no.>" = with pre-computed pixel weights from method <no.>.

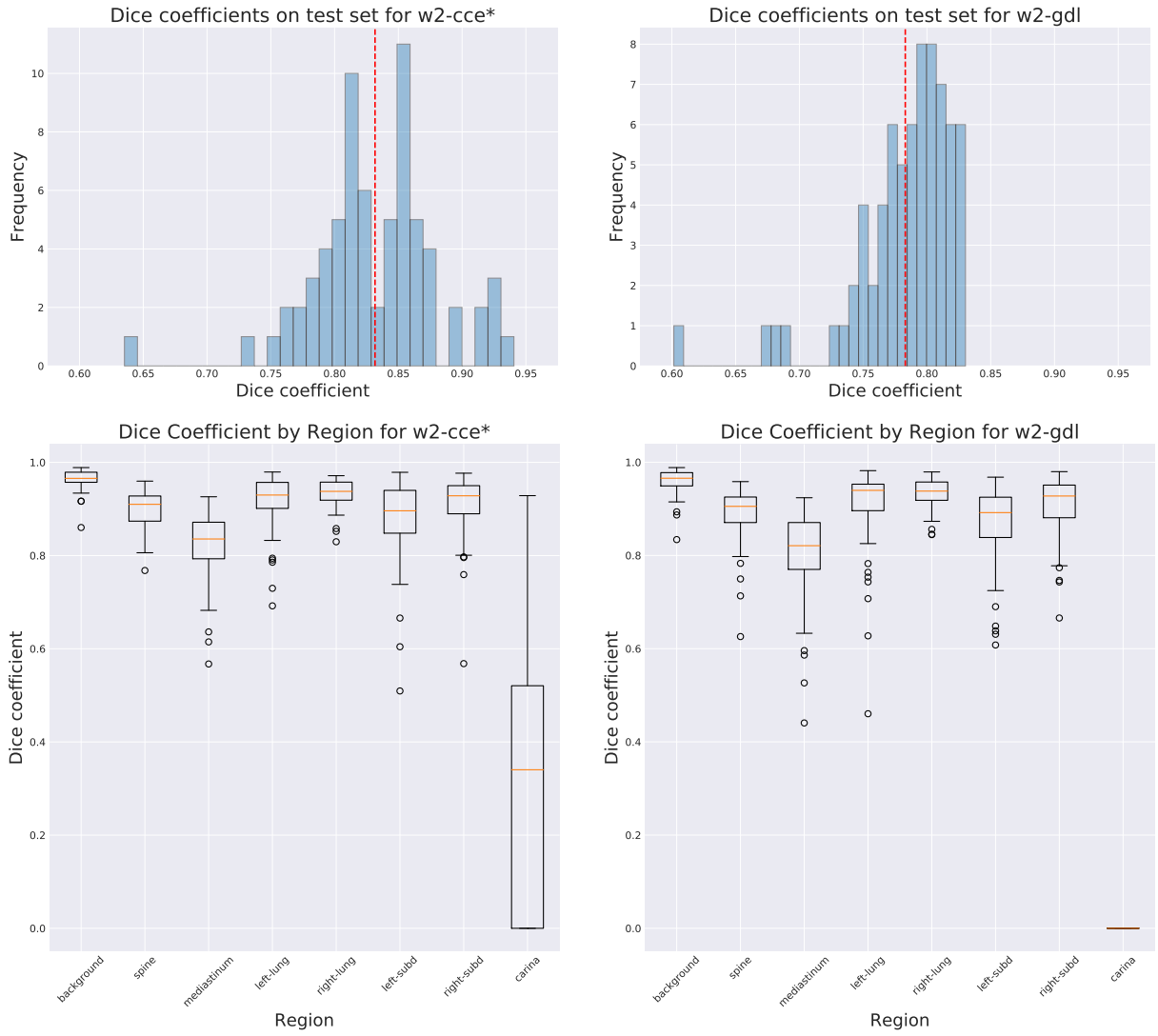

Figure 3. Results for the for best- (left) and worst-performing (right) models, as per Table 2, applied to the test set ($n = 70$). The top row contains histograms of mean Dice coefficients on the test set, where the red dashed vertical line is the overall mean Dice score. The bottom row contains box-and-whisker plots of Dice coefficients by region, where the orange horizontal bar represents the median Dice coefficient for that region.

4. DISCUSSION

Since the generalized Dice loss can be interpreted as a continuous analog of our primary performance metric, the Dice coefficient, L_{GDL} was an alluring choice of loss function. Additionally, others have found L_{GDL} to be particularly suited for segmentation problems that suffer from class imbalance[14,15] For these reasons, it is somewhat surprising that categorical cross-entropy so consistently outperformed generalized Dice loss in this task. Interestingly, even with pre-computed pixel weights to emphasize the carina class, no model trained on any version of L_{GDL} learned to locate the carina.

Table 3. Summary of "failures" made by three models on the test set ($n = 70$).

Loss Function	Missing Carina	Disconnected Classes
w2-cce*	0	6
cce	0	12
w2-gdl	70	24

Overall, the inclusion of pixel weights into either loss function turned out to be only slightly beneficial. It is interesting how well the models trained on *unweighted* cross-entropy performed on the test set, particularly that

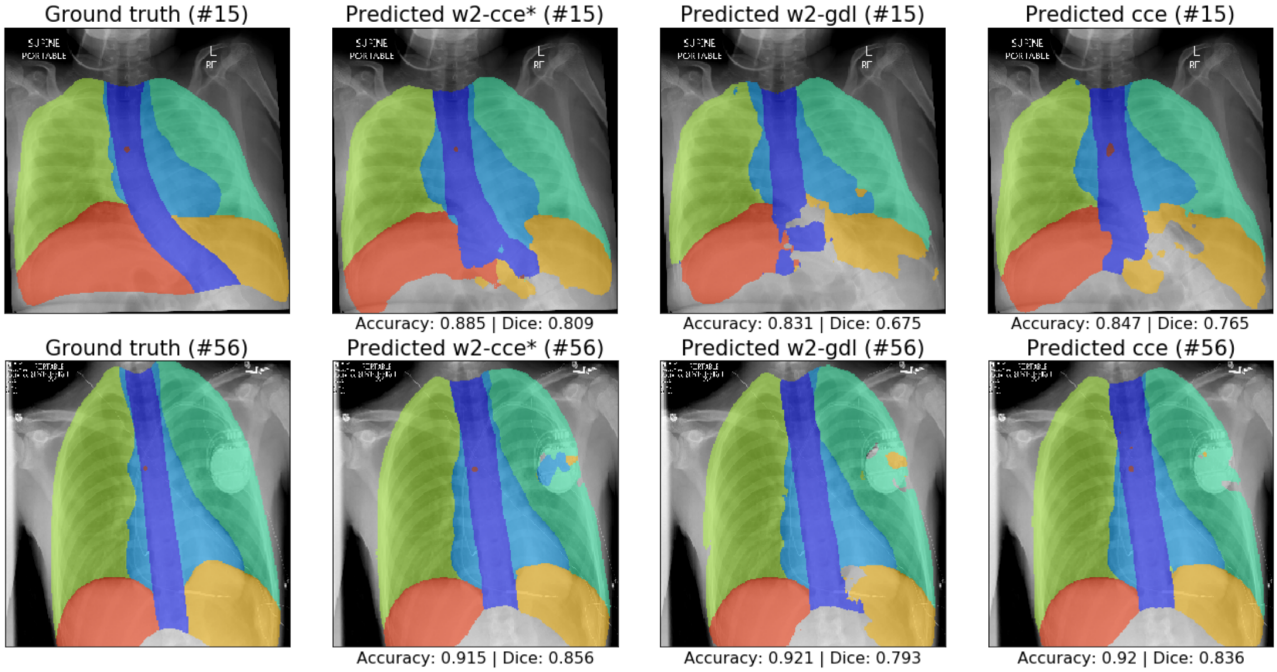

Figure 4. Examples of two "failure" cases (rows) by best-, worst-, and mediocre-performing models w2-cce*, w2-gdl, and cce. Each prediction contains multiple classes with disconnected regions, and both predictions from w2-gdl lack a carina.

they were able to handle the large class imbalance and always predict a carina region. Data augmentation proved to be helpful with roughly 5% improvements in Dice score. Ideally, each patient and radiograph film would be perfectly upright, but this is not the case in practice. The inclusion of random rotations may have made each model more robust to variation in patient angle. This leads us to believe that more aggressive augmentation – creating far more than three copies of each image and perhaps including other operations such as contrast shifts – may further improve overall performance.

Rules designed to catch morphologically inaccurate predictions added value to the carina segmentation. Since post-processing was only applied to the carina, it is reasonable to believe that the number of "failures" for many models would decrease when applying the same post-processing to all classes. There may be value to incorporating this idea into the learning process by designing a loss function that prohibits or discourages disconnected regions and promotes large connected components via dynamically computed pixel weights applied in appropriate areas.

In a preliminary evaluation, we sought to explore the generalizability of the trained models across different age groups. We considered how three different models – first presented in Table 2 – performed on test sets of toddlers (ages 0-4) *vs.* non-toddlers (ages 5+) in terms of mean Dice coefficient (Table 4). We observe comparable performance, but slightly higher Dice coefficients for older patients. Considering approximately 66% of the training set consisted of radiographs from patients aged 0-4, one would imagine each model might perform best on the youngest patients in the test set. Since Dice scores for non-toddlers were consistently higher despite being "underrepresented" in the training set, this perhaps suggests chest segmentation of very young patients is a more difficult task than in more developed patients. To extend this idea of comparing performance by age clusters, it would be interesting to compare how a model trained on only adults (or older pediatric patients) performs on pediatric radiographs and vice-versa – that is, to compare adult-specific models on pediatric radiographs and vice-versa.

Table 4. Mean Dice coefficient by age class (toddler *vs.* non-toddler) for three different models on the test set.

Loss Function	Dice Coefficient	
	Ages 0-4 (n=38)	Ages 5+ (n=32)
w2-cce*	0.830	0.834
cce	0.812	0.820
w2-gdl	0.781	0.785

As is true of almost any machine learning solution, more labeled data would likely improve the performance of our models. Beyond that however, we would benefit perhaps even more greatly from improved label quality. Radiographs were not annotated by clinicians or radiologists, so the quality of our ground truth labels inherently limits the predictive ability of any model trained on those labels, thus leading to noisy measures of performance. This being the case, it is a testament to the power of such deep learning approaches that we could produce mostly realistic, morphologically sound chest segmentations from imperfectly labeled training examples.

5. CONCLUSION

To the best of our knowledge, we are among the first to apply deep learning strategies to multi-class segmentation of pediatric chest radiographs. Our work is innovative in its use of a U-Net for eight distinct classes and its application to a problem where class imbalance between segmented regions spans 10-fold (for example, background to subdiaphragm region) and 10^5-fold (background to carina region) variations; this problem required application of the unique weighted loss functions presented here. Our approach has the potential to provide anatomic context for a variety of future automated diagnostic applications such as assessing line placement and determining the location of rib fractures. A comparison of loss functions revealed that categorical cross-entropy consistently outperformed the recently popular generalized Dice loss. Lastly, light data augmentation – elastic deformations and rotations – improved performance, and pre-defined pixel weights appeared to slightly improve performance in the hardest-to-learn regions.

ACKNOWLEDGMENTS

This work was supported by the NSF REU grant 1560168 and by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the NIH under Award Number R21HD097609.

REFERENCES

[1] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., and McDonald, C. J., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imag.* **33**(2), 577–590 (2014).

[2] Juhász, S., Horváth, Á., Nikházy, L., Horváth, G., and Horváth, Á., "Segmentation of anatomical structures on chest radiographs," in *12th Medit. Conf. Med. Biol. Eng. Comput. (MEDICON)*, 359–362, Springer Berlin Heidelberg, Berlin, Heidelberg (2010).

[3] Loog, M. and Ginneken, B., "Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification," *IEEE Trans. Med. Imag.* **25**(5), 602–611 (2006).

[4] Koehler, C. and Wischgoll, T., "Knowledge-assisted reconstruction of the human rib cage and lungs," *IEEE Comput. Graph. Appl.* **30**(1), 17–29 (2010).

[5] Tsujii, O., Freedman, M. T., and Mun, S. K., "Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network.," *Med. Phys.* **25**(6), 998–1007 (1998).

[6] Luo, H. and Foos, D., "Method for automated analysis of digital chest radiographs." U.S. Patent 7,221,787 B2, *U.S. Patent and Trademark Office* (2007).

[7] Ngo, T. A. and Carneiro, G., "Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2140–2143 (2015).

[8] Arbabshirani, M. R., Dallal, A. H., Agarwal, C., Patel, A., and Moore, G., "Accurate segmentation of lung fields on chest radiographs using deep convolutional networks," in *Proc. SPIE 10133*, 1013305 (2017).

[9] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 234–241 (2015).

[10] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X., "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from http://tensorflow.org/.

[11] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," in *3rd Int. Conf. on Learning Representations (ICLR)*, (2015).

[12] Crum, W. R., Camara, O., and Hill, D. L., "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.* **25**(11), 1451–1461 (2006).

[13] Simard, P. Y., Steinkraus, D., and Platt, J. C., "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. on Doc. Anal. Recog. (ICDAR)*, **2**, 958 (2003).

[14] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 240–248, Springer (2017).

[15] Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., and Hamarneh, G., "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Comput. Med. Imaging Graph.* **75**, 24–33 (2019).