Deep learning methods for segmentation of lines in pediatric chest radiographs

Ryan Sullivan[ac], Greg Holste[bc], Jonathan Burkow[c], Adam Alessio[c*]
[a]Purdue University, [b]Kenyon College, [c]Michigan State University

ABSTRACT

Surgical procedures often require the use of catheters, tubes, and lines, collectively called lines. Misplaced lines can cause serious complications, such as pneumothorax, cardiac perforation, or thrombosis. To prevent these problems, radiologists examine chest radiographs after insertion and throughout intensive care to evaluate their placement. This process is time consuming, and incorrect interpretations occur with notable frequency[1]. Fast and reliable automatic interpretations could potentially reduce the cost of these surgical operations, decrease the workload of radiologists, and improve the quality of care for patients. We develop a segmentation model which can highlight the medically relevant lines in pediatric chest radiographs using deep learning. We propose a two-stage segmentation network which first classifies whether images have medically relevant lines and then segments images with lines. For the segmentation stage, we use the popular U-Net architecture[2] substituting the encoder path with multiple state-of-the-art CNN encoders. Our study compares the performance of different permutations of model architectures for the task of highlighting lines in pediatric chest radiographs and demonstrates the effectiveness of the two-stage architecture.

1. INTRODUCTION

Catheters, lines, and tubes including nasogastric tubes , endotracheal tubes , central venous catheters, and intercostal catheters are inserted into patients for a variety of surgical procedures. These devices (which we will refer to as lines) are commonly used in pediatric care to manage critically ill patients. For example, endotracheal tubes are inserted to allow for artificial ventilation, and nasogastric tubes are used to provide nutritional support. These lines can be placed incorrectly or move during care, leading to serious complications such as pneumothorax, hemorrhaging, and other forms of trauma. Radiographic images are often taken post-procedure to assess the placement of these lines.

Multiple studies have been conducted to determine the incidence of line misplacements, complications due to misplacement, and line misplacements when assisted by radiographic images. For instance, nasogastric tubes misplacement results in complications 1-3% of the time[3], and endotracheal tubes are incorrectly intubated 23.5% of the time in emergency care[4], though reported error rates vary. Overall, it has been shown that 26% of catheters are incorrectly placed, and 6% of placements results in complications that are radiographically visible[5]. To the best of our knowledge, there are few other studies that have attempted to identify and segment lines automatically on pediatric chest radiographs. One other study has performed line segmentation on adult radiographs using artificially generated lines, not real lines present in patients[6]. Another group developed a segmentation model for detecting peripherally inserted central catheters (PICC) using fully convolutional networks (FCNs)[7]. This study focused on PICC tip detection and does not take advantage of more recent developments in biomedical image processing.

We develop a two-stage segmentation model, using a classification network followed by a segmentation network, to achieve visually accurate segmentations of catheters in pediatric chest radiographs. We test a number of popular CNN models for the classification stage, and develop of a U-Net based architecture for segmentation. Our segmentation model takes advantage of recent developments in CNN encoders to enhance the performance of the standard U-Net model. Similar approaches using a VGG16 encoder have been demonstrated on a number of other segmentation applications[8]. Additionally, short residual connections, like those found in the ResNet model, have led to increased performance on

***aalessio@msu.edu; phone 1-517-432-1708; https://www.egr.msu.edu/~aalessio/**

biomedical segmentation applications[9]. We train each stage of the network separately, using categorical cross-entropy loss for the classification stage and Dice coefficient loss for the segmentation stage[10].

2. METHODS

The two-stage detector includes two main components: a binary classification stage and a segmentation stage.

2.1 Architectures for classification and segmentation

First, we developed two CNN architectures to perform binary classification of pediatric chest radiographs. The first model used a ResNet50 encoder, and the second used a ResNext101 encoder. Both encoders were followed by a dense layer of 1024 nodes, a dense layer of 256 nodes, both of which used the relu activation function. The final layer was a dense layer of 2 nodes using softmax activation to produce the classifier output. Each model was instantiated with the pretrained ImageNet weights, and every layer of the model was retrained.

For the segmentation task, we developed a U-Net style model with a modern CNN encoder replacing the default U-Net encoding (downsampling) path. We also test a U-Net model with feature pyramid networks (FPN)[11] applied to the decoding path. In the U-Net model, skip connections were added after each of the first 4 stages of the downsampling path to the U-Net's upsampling path. For the FPN model, feature pyramids using 256 filters were added at each of these stages. The results from the feature pyramids are then upsampled to the size of the inputs and concatenated as in the original FPN paper.

2.2 Data Description

A summary of the data for training and testing is presented in Table 1. The data set was divided into medically relevant line present and line absent cases based on visual inspection. These cases were used for training and testing the classification model. A portion of the line present cases were manually annotated with segmentation masks for supervised training of the segmentation model. To annotate the images, an annotator used a digital tool to place points along an image. A matlab script is used to connect these points with a simple, smooth curve. This curve is then expanded to a set width, perpendicular to the curve. As such, the annotations are not pixel-perfect segmentation masks, but are close approximations. This allowed us to rapidly label many samples, at the cost of some label accuracy.

Table 1. Summary of data sets employed for each stage. The dataset is a hand annotated IRB-Approved Custom Ped AP Chest Radiographs:

Model	Data Set Labels	Percent of Samples with Lines Present	Train / Validation / Test Set Size
Classification	Binary class labels	75%	771 / 165 / 166
Segmentation	Catheter segmentation masks	100%	66 / 15 / 15
Two-Stage	Binary class labels and catheter segmentation masks	50%	132 / 30 / 30

2.3 Two-Stage Model

By combining the two stages, we create our full model for semantic segmentation of lines. Images are first passed through the classification network, generating a class prediction. For images without lines, we simply predict an empty segmentation mask. For images with lines, we use the segmentation network to predict masks. Since every input to the segmentation network contains lines, we can train that model only on images that contain lines. Decreasing the number

of all background input images reduces the class imbalance between line and background pixels, thereby improving training. We train both stages using supervised learning. The classification stage is trained with categorical cross entropy loss, with each sample being assigned a weight based on its prevalence in the dataset. In this case, there are three times as many samples with lines, so weights of 0.33 and 1.0 are assigned to samples with lines and without lines respectively. The segmentation stage was trained with Generalized Dice loss[10].

2.4 Performance Metrics

After training the different models, we evaluated the performance of the final segmented masks on the test set. The Dice score and intersection over union (IoU) scores are evaluated by first rounding the pixel-wise predictions in the predicted masks to the nearest value. This converts the probabilities in the predicted mask to discrete class predictions by selecting the class with the highest predicted probability for each pixel.

3. RESULTS

Results show that the ResNet50 classification model slightly outperforms the ResNeXt101 architecture for binary classification of radiographs, as seen in Table 2. Both models achieve a high accuracy for the binary classification task.

Table 2. Classification results from the two tested architectures. These are the popular ResNet50 and ResNeXt101 CNN architectures, both pretrained on the ImageNet dataset.

Model	Test Accuracy
ResNet50	**93.4%**
ResNeXt101	92.8%

Results for the segmentation of the lines using just the segmentation model are presented in Table 3. For the segmentation of the lines, we found that the U-Net model using the EfficientNet B3 backbone trained on generalized Dice loss outperformed other models, achieving a Dice score of 0.551.

Table 3. Segmentation results from a variety of encoder and decoder model. “UNet” models are the original U-Net with encoding paths replaced by the listed CNN architecture. “FPN” models use the same encoder substitutions, but have an additional feature pyramid network on the decoding path.

Model	Dice Coefficient	IoU
UNet-ResNet50	0.527	0.699
UNet-ResNeXt101	0.524	0.696
UNet-EfficientNetB3	**0.551**	**0.706**
FPN-ResNet50	0.498	0.676
FPN-ResNeXt101	0.492	0.675
FPN-EfficientNetB3	0.508	0.684

We see that the segmentation model produces visually accurate segmentation masks on the new samples. Some of these examples can be seen in Figure 1. The left column shows the true binary masks, and the right column shows predictions from trained models. Figure 1.b demonstrates a visually correct segmentation of lines, while Figure 1.d displays a poorer prediction from one of the less successful models.

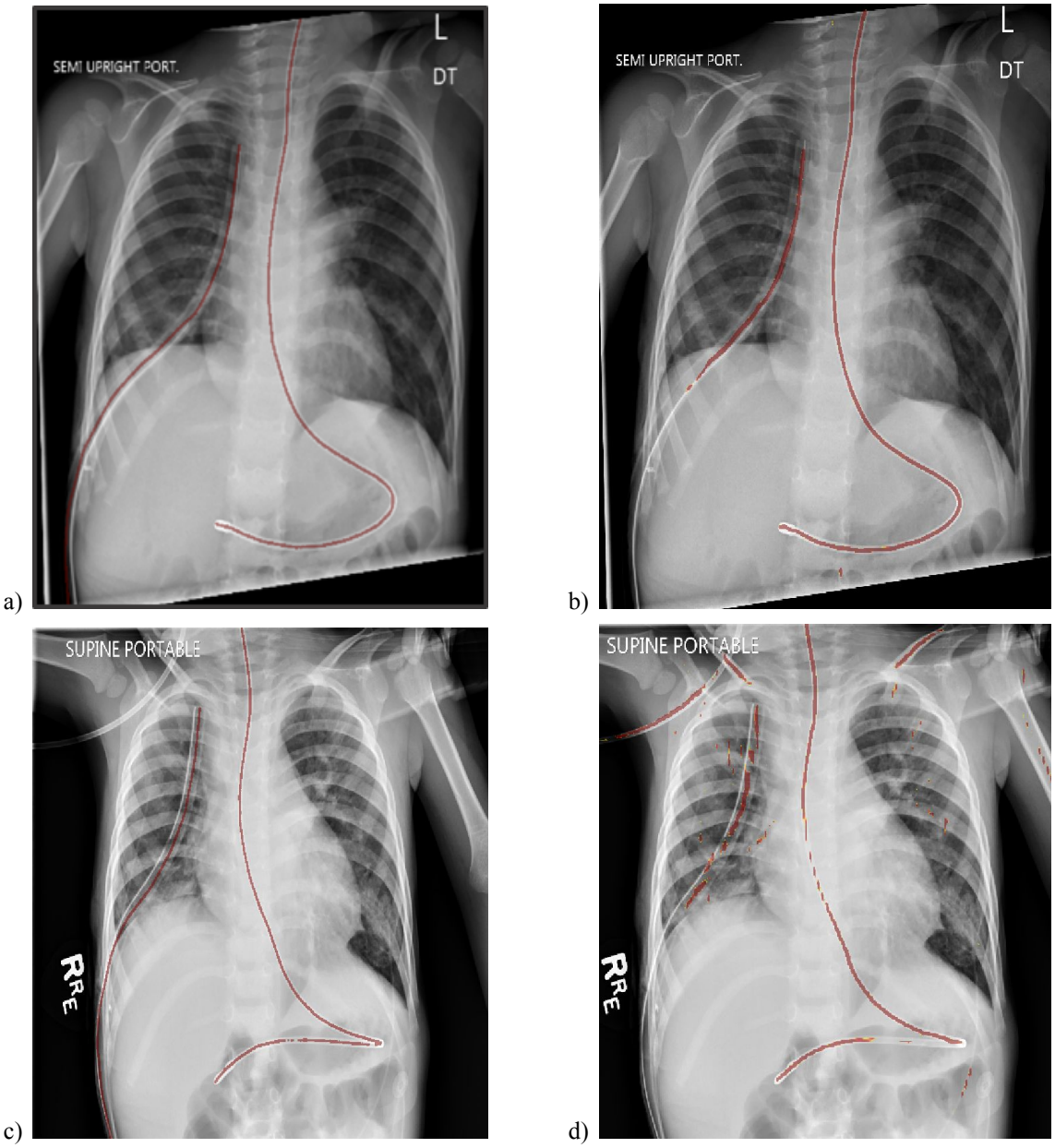

Figure 1. a) Sample ground truth and b) predicted segmentations of lines from UNet-EfficientNetB3 model. c) Sample ground truth and d) predicted segmentations of lines from FPN-ResNet50 model.

We find that the two-stage model improves overall segmentation results by using the classification network to avoid rare, but significant segmentation mistakes. Figure 2. displays some examples where the segmentation network highlighted the edges of bones, mistaking them for lines. Additionally, Table 4. compares the two-stage network's performance to our best standalone segmentation model, and demonstrates that the two-stage network results in notable improvements to the dice score and intersection over union metric. In this case, the first stage classified all but one of the images correctly. However, the segmentation network highlighted bone edges in some of the samples with no lines present, resulting in very low dice scores. By predicting entirely empty masks for

those images, which were classified correctly, we were able to achieve significant improvements to the average Dice score, despite the single misclassified sample providing a very low Dice score.

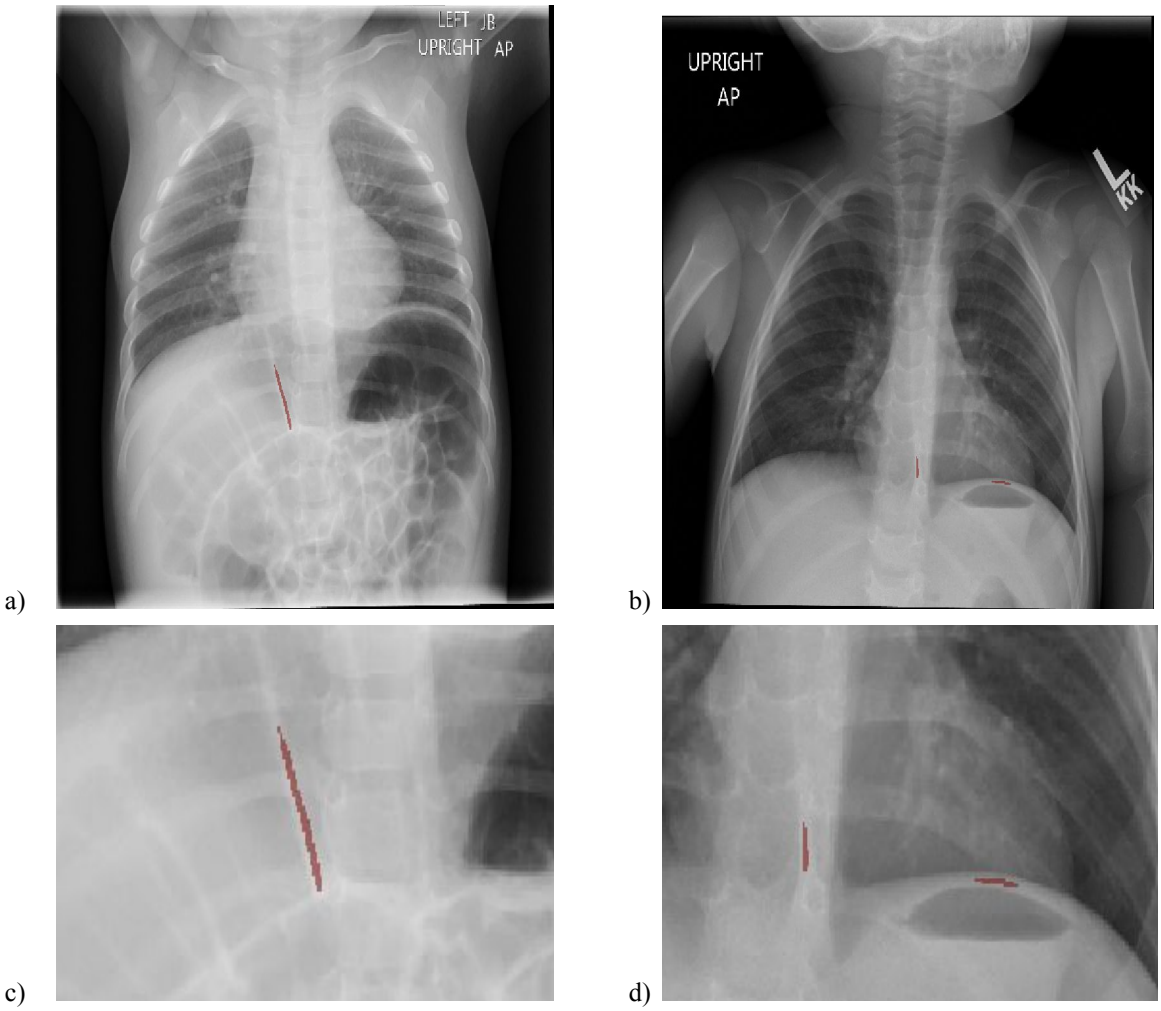

Figure 2. Examples of two segmentation mistakes made by the UNet-EfficientNetB3 model. a) and b) are predicted masks for samples with no lines present, which contain pixels classified as lines. c) Shows a close-up of the mistake in a), and d) shows a close-up of the mistake in b).

Table 4. Segmentation results comparing the two-stage network to the original segmentation network, when evaluated on the two-stage dataset (containing line present and absent images) samples listed in Table 1.

Model	Dice Coefficient	IoU (F score)
UNet-EfficientNetB3	0.698	0.623
TwoStage-UNet-EfficientNetB3	**0.737**	**0.666**

4. DISCUSSION

We see that modern fully convolutional architectures, like U-Net, perform well on our dataset when trained with Generalized Dice loss. Our proposed two-stage network demonstrates additional improvements, which can be expanded on in future research. There are a number of remedial steps we plan to introduce to further improve the performance of the network. First, we plan to annotate more data to demonstrate the impact that sample size has on the model. The dataset used in this experiment was relatively small, so we expect that the model will improve significantly with more samples. Second, we plan to attempt data augmentation to increase the utility of the existing dataset. The current method of generating segmentation labels from points does not closely approximate lines of all sizes. Lines vary in size from roughly 1mm in diameter (peripherally inserted central catheters) to 9mm (intercostal catheters for children). As a result, our fixed-width generated lines do not accurately cover lines of all sizes The utility of our dataset, and the accuracy of the model could be greatly improved by improving the quality of the generated labels. Third, we may segment the lines into individual classes. By allowing the model to select against visually different line types, we hope to reduce the range of objects that are misclassified as lines and improve overall accuracy. Similarly, to address the incorrect segmentation of bone edges, we could annotate ribs in our dataset to allow the model to train against that information.

Our work approaches this problem with a novel two-stage network. In this study, both stages were trained independently, then combined to make predictions. A training method which allows both networks to be trained simultaneously would allow each stage to share error signals during backpropagation, potentially improving the overall accuracy of the model. In addition, we performed a systematic comparison of different classification and segmentation architectures in order to determine optimal architectures for this data and task. This progress is our first step towards the larger goal of determining if lines are in a detrimental position. Future work will use these results to locate the tips of various lines and identify incorrectly place lines automatically. This would lead to a model that could be effectively employed in hospitals to improve patient outcomes.

5. CONCLUSION

We have developed a model capable of producing accurate predictions of line segmentations in pediatric chest radiographs. We compared multiple models to determine the best combination of model architectures and encoders to achieve this goal. We trained a ResNeXt101 classification model, achieving a binary classification accuracy of 93.4% for identifying line present versus line absent images. We trained U-Net and FPN models using a variety of CNN encoders, and found that the U-Net with EfficientNet-B3 model resulted in the highest Dice score. This system produced visually accurate masks on our dataset with severe class imbalance and thin, connected structures. We also combine these models into our two stage detector, benefiting from whole image classification first then segmentation only on images with lines. We demonstrate that this leads to notable improvements in both the Dice score and IoU score. Future work will improve the quality of predictions, and use this methodology to determine if line placement is normal or abnormal based on the location of the end of each line.

6. ACKNOWLEDGEMENTS

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] Brady, A. P., "Error and discrepancy in radiology: inevitable or avoidable?," Insights into Imaging **8**(1), 171–182 (2016).

[2] Ronneberger, O., Fischer, P., Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 234–241 (2015).

[3] Bickle, I. "Nasogastric Tube Positioning: Radiology Reference Article." Radiopaedia Blog RSS, October 2019. <https://radiopaedia.org/articles/nasogastric-tube-positioning> (21 January 2020).

[4] Ono, Y., Kakamu, T., Kikuchi, H., Mori, Y., Watanabe, Y., Shinohara, K., "Expert-Performed Endotracheal Intubation-Related Complications in Trauma Patients: Incidence, Possible Risk Factors, and Outcomes in the Prehospital Setting and Emergency Department," Emergency Medicine International 2018, 1–9 (2018).

[5] Bekemeyer, W. B., Crapo, R. O., Calhoon, S., Cannon, C. Y., Clayton, P. D., "Efficacy of Chest Radiography in a Respiratory Intensive Care Unit," Chest 88(5), 691–696 (1985).

[6] Yi, X., Adams, S., Babyn, P., Elnajmi, A., "Automatic Catheter and Tube Detection in Pediatric X-ray Images Using a Scale-Recurrent Network and Synthetic Data," Journal of Digital Imaging (2019).

[7] Lee, H., Mansouri, M., Tajmir, S., Lev, M. H., Do, S., "A Deep-Learning System for Fully-Automated Peripherally Inserted Central Catheter (PICC) Tip Detection," Journal of Digital Imaging **31**(4), 393–402 (2017).

[8] Iglovikov., Vladimir., Shvets., Alexey., "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," arXiv.org, 17 January 2018, <https://arxiv.org/abs/1801.05746> (22 January 2020).

[9] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., "The Importance of Skip Connections in Biomedical Image Segmentation," Deep Learning and Data Labeling for Medical Applications Lecture Notes in Computer Science, 179–187 (2016).

[10] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M. J., "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support Lecture Notes in Computer Science, 240–248 (2017).

[11] Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).