CONVOLUTIONAL NEURAL NETWORKS FOR SEGMENTATION OF MALIGNANT PLEURAL MESOTHELIOMA: ANALYSIS OF PROBABILITY MAP THRESHOLDS (CALGB 30901, ALLIANCE)

Mena Shenouda
Department of Radiology
The University of Chicago
Chicago, IL, USA
shenmena@uchicago.edu

Eyjólfur Gudmundsson
Icelandic Radiation Safety Office
Reykjavik, Iceland
eyjolfur88@gmail.com

Feng Li
Department of Radiology
The University of Chicago
Chicago, IL, USA
feng.li@uchospitals.edu

Christopher M. Straus
Department of Radiology
The University of Chicago
Chicago, IL, USA
cstraus@uchicago.edu

Hedy L. Kindler
Department of Radiology
The University of Chicago
Chicago, IL, USA
hkindler@medicine.bsd.uchicago.edu

Arkadiusz Z. Dudek
Metro Minnesota Community Oncology Research Consortium
St. Louis Park, MN, USA
adudek021@outlook.com

Thomas Stinchcombe
Duke Cancer Institute
Duke University
Durham, NC, USA
thomas.stinchcombe@duke.edu

Xiaofei Wang
Alliance Statistics and Data Management Center
Duke University
Durham, NC, USA
xiaofei.wang@duke.edu

Adam Starkey
Department of Radiology
The University of Chicago
Chicago, IL, USA
starkey@uchicago.edu

Samuel G. Armato III
Department of Radiology
The University of Chicago
Chicago, IL, USA
s-armato@uchicago.edu

ABSTRACT

Malignant pleural mesothelioma (MPM) is the most common form of malignant mesothelioma, with exposure to asbestos being the primary cause of the disease. To assess response to treatment, tumor measurements are acquired and evaluated based on a patient's longitudinal computed tomography (CT) scans. Tumor volume, however, is the more accurate metric for assessing tumor burden and response. Automated segmentation methods using deep learning can be employed to acquire volume, which otherwise is a tedious task performed manually. The deep learning-based tumor volume and contours can then be compared with a standard reference to assess the robustness of the automated segmentations. The purpose of this study was to evaluate the impact of probability map threshold on MPM tumor delineations generated using a convolutional neural network (CNN). Eighty-eight CT scans from 21 MPM patients were segmented by a VGG16/U-Net CNN. A radiologist modified the contours generated at a 0.5 probability threshold. Percent difference of tumor volume and overlap using the Dice Similarity Coefficient (DSC) were compared between the standard reference provided by the radiologist and CNN outputs for thresholds ranging from 0.001 to 0.9. CNN annotations consistently yielded smaller tumor volumes than radiologist contours. Reducing the probability threshold from 0.5 to 0.1 decreased the absolute percent volume difference, on average, from 43.96% to 24.18%. Median and mean DSC ranged from 0.58 to 0.60, with a peak at a threshold of 0.5; no distinct threshold was found for percent volume difference. The CNN exhibited deficiencies with

specific disease presentations, such as severe pleural effusion or disease in the pleural fissure. No single output threshold in the CNN probability maps was optimal for both tumor volume and DSC. This study emphasized the importance of considering both figures of merit when evaluating deep learning-based tumor segmentations across probability thresholds. This work underscores the need to simultaneously assess tumor volume and spatial overlap when evaluating CNN performance. While automated segmentations may yield comparable tumor volumes to that of the reference standard, the spatial region delineated by the CNN at a specific threshold is equally important.

Keywords MPM · Probability maps · Tumor volume

1   Introduction

Malignant pleural mesothelioma (MPM) is an aggressive form of cancer present in the pleural lining of the lungs. It is often the result of exposure to asbestos and has a poor prognosis [1]. Computed tomography (CT) is the most common imaging modality used to stage and assess patients with MPM. The current standard to evaluate tumor response to therapy is the modified Response Evaluation Criteria in Solid Tumors (mRECIST). Using this protocol, observers make up to six measurements of "tumor thickness residing perpendicular to the chest wall or mediastinum" as presented on a CT scan [2]. This protocol has been recently updated as mRECIST 1.1 [3].

In contrast to the linear measurements of mRECIST, manual volumetric analysis conducted by radiologists can be used to better estimate tumor burden. Tumor volume has displayed strong predictive power in patient assessment in terms of overall and progression-free survival [4, 5]; however, acquisition of manual tumor volume is too time-consuming and burdensome to be systematically used in routine clinical care. The difficulties of tumor volume calculations can be mitigated using artificial intelligence and machine learning, specifically convolutional neural networks (CNNs). Previous studies have successfully implemented CNNs in MPM segmentation, which is a crucial step in automating tumor volume measurements as the laborious part is performed automatically, and simple pixel counting remains for the volume calculation [6, 7].

For a CNN to properly segment MPM tumor on CT sections, the network must be trained and validated using a labeled set of images. Due to the fundamental statistical nature of machine learning, the outputs of a CNN are probabilities. For instance, in identifying the location of tumor on a CT scan, the CNN assigns each pixel a probability of being tumor. Therefore, for each CT section input to the network, a probability map is generated that displays the likelihood of tumor in a pixel-wise fashion. In practice, a threshold is set to binarize these maps so that any pixel with probability equal to or greater than 0.5, for example, is set to 1 ("tumor"), and all other pixels are set to 0 ("not tumor"). Given that the 0.5 probability of might not generate the most accurate segmentation for a disease as complicated as MPM tumor, the purpose of the present study was (1) to better understand the probability values returned by the CNN and (2) to investigate whether different probability thresholds could improve pixel-wise class labeling in this complicated tumor. The tumor segmentations obtained using different thresholds were evaluated against the reference standard using the Dice Similarity Coefficient (DSC) and the percent difference of volume as the figures of merit. Therefore, we investigated the impact that varying thresholds would have on the predicted tumor segmentation/probability maps, given the inherent complexities of this tumor.

Overall, the choice of threshold may have considerable impact on the final tumor segmentation. In terms of MPM tumor volumetry, a lower threshold applied to the probability maps would result in a larger volume as more pixels are now considered tumor; however, this may result in an increase of pixels erroneously labeled as tumor, thus negatively impacting the CNN accuracy of tumor segmentation. In contrast, increasing the threshold may produce a segmentation that is too restrictive, substantially decreasing the tumor volume calculated. Furthermore, the change in threshold alters the overlap of tumor identified by the CNN with the ground truth (as determined by a trained observer). It is important to note that the purpose of this work was not to assess an observer's performance when given initial computerized outlines (as our group has previously shown the significant effect of initial outlines on observer performance [8]) but rather to demonstrate the impact a choice of probability threshold has on defining tumor. Therefore, this study investigated the impact of probability thresholds on tumor volume and the overlap of tumor contours by applying a broad range of thresholds, recording the volumes and the DSC and studying the resulting trends.

2   Methods

2.1   Patient population

The patient cohort was compiled from a previous study performed by the Cancer and Leukemia Group B (CALGB 30901) [9]. CALGB is now part of the Alliance for Clinical Trials in Oncology. The CALGB 30901 study evaluated 49

patients with confirmed unresectable epithelioid, sarcomatoid, or mixed-type MPM, and patients were without disease progression after 4 to 6 cycles on first-line therapy with pemetrexed and cisplatin or carboplatin. Patients were randomly assigned to either the treatment arm (continued therapy with pemetrexed alone) or the observation arm. The patients underwent CT scans at baseline and then every 6 weeks for the first 6 months. Each participant signed an IRB-approved, protocol-specific informed consent document in accordance with federal and institutional guidelines.

2.2 CT image acquisition and tumor contours

The present study was retrospectively conducted on 88 baseline and follow-up CT scans of 21 patients from the CALGB 30901 study. There was an average of 106 sections per scan (range: 39-418 sections) and 3.7-mm thickness (range: 1.25-5 mm). The scans were acquired using 21 different scanner models.

Tumor contours were generated automatically using a CNN previously published [7] without any additional training or validation. The CNN employed the U-Net deep CNN (2D) architecture. Specifically, the deep CNN architecture consisted of a downsampling and upsampling path. For the downsampling, a Visual Geometry Group 16 (VGG16) model was pre-trained on the ImageNet database using scale-jittering [10, 11]. Layers of the downsampling path were initialized using the weights acquired from the VGG16 training scheme. A 2×2 max pooling operation with stride 2 was applied to the feature maps at each step of the downsampling path. A dropout layer of probability 0.5 was used to prevent overfitting. During the upsampling path, a 2D operation using nearest-neighbor interpolation was applied to the feature maps. The network output a segmentation mask the same size as the input image size (e.g., 512×512 pixels). A rectified linear unit (ReLU) activation function was applied after the convolutional layers, except for the last layer, which used a sigmoid activation function that returned pixel-wise probabilities on the range [0,1] for the segmentation task, i.e., whether a pixel contains tumor. We applied a threshold of 0.5 to the output of the sigmoid layer during the validation step so that any pixel with a probability 0.5 or greater was labeled "tumor." During its training phase, the network minimized the binary cross-entropy, which was averaged over all pixels of each predicted segmentation and the provided reference standard. Adam, an algorithm for first-order gradient-based stochastic optimization, was used to optimize the network during training using a learning rate of 10^{-5}.

The VGG16/U-Net deep CNN architecture was previously trained on 126 MPM patients, some presenting with pleural effusion [7]. In this earlier study, the CNN was tested on 34 patients, who all presented with both tumor and pleural effusion, and on 43 patients, some of whom presented with both tumor and pleural effusion and some only with tumor; the median DSC and median average Hausdorff distance (AHD) for that method were 0.690 and 5.1 mm, respectively, as previously reported [6]. For the present study, no additional training or validation was performed before evaluating the performance of the trained CNN on the external CALGB dataset.

A research radiologist [F.L.] presented with the CNN contours (acquired using a predetermined probability threshold value of 0.5) was able to modify/redraw the contours using in-house software to provide the reference standard. Due to the time-consuming nature of adjusting the contours, however, the radiologist was presented with sections separated by approximately 5 mm. This process resulted in an average of 51 reviewed sections per scan (range: 32-70 sections). Section comparisons and tumor volumes were performed only on contours of sections that the radiologist reviewed.

2.3 Tumor volume and Dice similarity coefficient

2.3.1 Tumor volume calculation

Tumor volume was defined as:

$$\text{Volume [mm}^3] = \sum \text{Number of pixels within a contour}$$

$$\times \text{ pixel dimension}^2 \text{ [mm}^2] \times \text{ inter-section distance [mm]} \quad (1)$$

where the summation is over all sections containing a contour. The number of pixels within a contour was equal to the number of nonzero pixels within the binary mask created after applying a threshold to the probability maps generated by the CNN. Pixel dimension (in units of mm^2) was acquired from the DICOM header. Inter-section distance corresponded to the difference in table position between two sections on which the radiologist provided reference contours. All tumor volumes computations were performed using MATLAB (Mathworks Inc., Natick, Massachusetts).

2.3.2 Dice similarity coefficient

Another metric used to compare the CNN tumor contours generated at the different thresholds with the radiologist's reference standard was the Dice similarity coefficient (DSC) [12]:

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (2)$$

Where $|A|$ and $|B|$ are the number of elements within a set, e.g., pixels within a contour. The DSC is a measure of overlap between segmentations: a comparison between contours generated by the CNN and the reference provided by the radiologist [13]. The DSC was calculated for each individual CT section (using MATLAB's "dice" function), and a final DSC was calculated per patient after averaging the DSC values across all sections.

2.4   Statistical methods

Comparison of the absolute percent difference of volumes and comparison of DSC values across thresholds were first checked for normality using the one-sample Kolmogorov-Smirnov test [14]. Since the null hypothesis was rejected, none of the data was deemed to come from a standard normal distribution. Therefore, the Wilcoxon signed rank test was used to compare DSC or absolute percent volume differences computed between the reference standard contours and contours generated across a range of CNN probability map threshold values. Statistical significance was considered at $p = 0.05$; the p-value was not adjusted to account for multiple comparison due to the small sample size in this hypothesis-generating study. Data collection was conducted by the Alliance Statistics and Data Management Center. Data quality was ensured by review of data by the Alliance Statistics and Data Management Center and by the study chairperson following Alliance policies.

3   Results

Figure 1 displays a visual representation of a change in the probability thresholds and its impact on the tumor contour. Pleural effusions present were difficult for the CNN to fully appreciate as shown. Overall, the thresholds ranged from 0.001 to 0.9; however, the CNN never assigned a pixel a probability of 0.75 or greater. Figure 2 shows boxplots of the DSC values comparing the reference contours to six of the thresholds applied to the CNN generated probability maps. Except for the 0.01 threshold, the range of DSC values decreased with the incremental reduction of probability thresholds. The median does not substantially change (see Table 1).
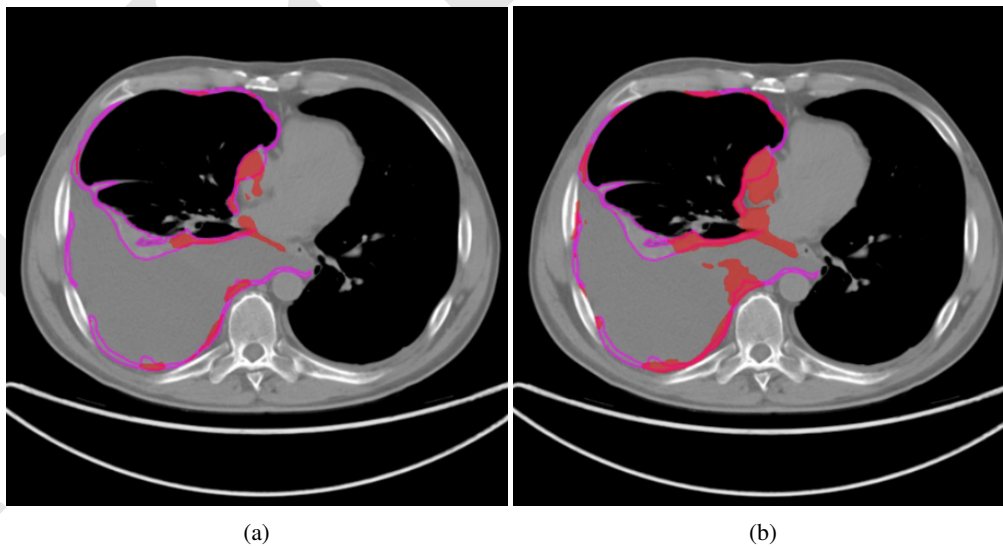

| (a) | (b) |

Figure 1: Differing contours on the same section of the same patient created with an adjustment in the CNN probability threshold. Purple represents the radiologist's reference outline, and red represents the CNN pixel-wise segmentation prediction of tumor with in (a) a probability threshold of 0.5 (average DSC: 0.357) and (b) a probability threshold of 0.001 (average DSC: 0.476).
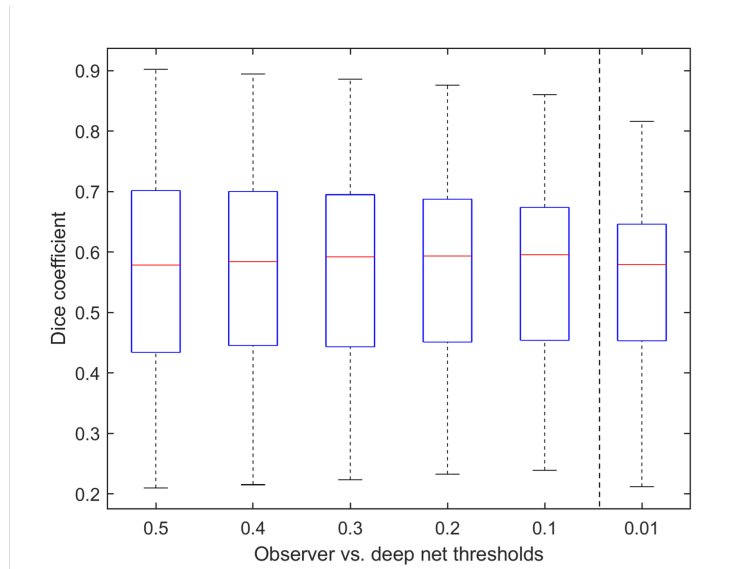
Figure 2: Boxplots showing the DSC values obtained for tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. The red lines display the median of the data.
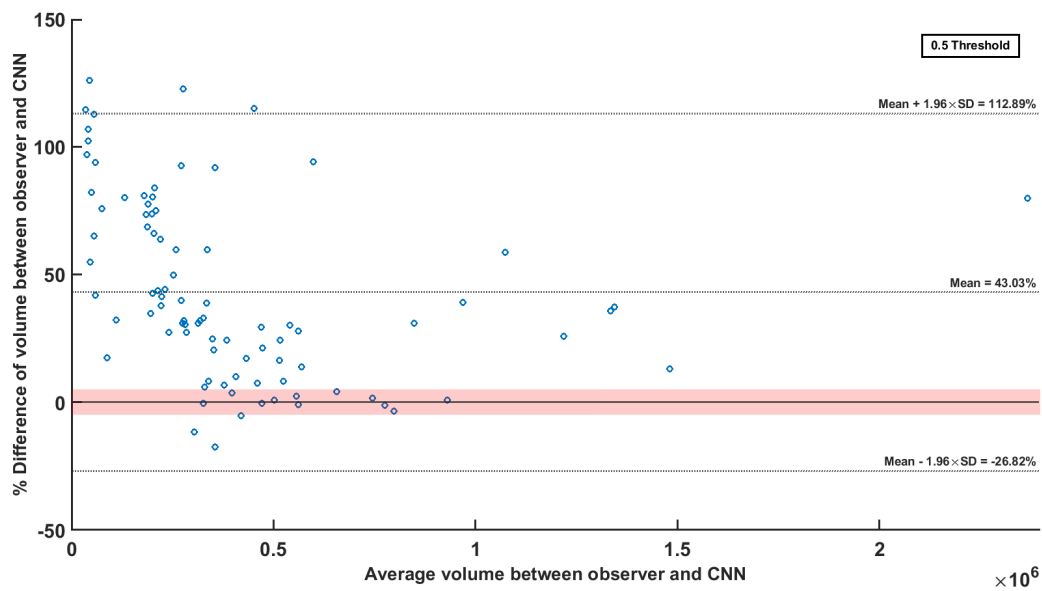

Figure 3: Bland-Altman plot of the relative differences between manual and CNN-based tumor volumes. The red band highlights differences within 5% of 0.

Figure 3 is a Bland-Altman plot [15] at the 0.5 threshold displaying the relative percent difference of volume as calculated using the radiologist's and the CNN contours. The CNN consistently undermeasured the tumor at that threshold, which resulted in a mean percent difference of 43% (range from -17% to 126%, median of 34%). Eleven (12.5%) scans were within $\pm5\%$ of 0% difference as shown by the red band in Figure 3.

Table 1: Absolute percent difference (± standard deviation) of volume, average, and median DSC at six thresholds. (IQR = interquartile range).

	Threshold = 0.5	Threshold = 0.4	Threshold = 0.3	Threshold = 0.2	Threshold = 0.1	Threshold = 0.01
Average absolute % difference of volume	43.96 ± 34.48	38.05 ± 31.78	32.68 ± 28.69	27.26 ± 25.38	24.18 ± 18.87	28.16 ± 18.41
Average DSC	0.58 ± 0.17	0.58 ± 0.17	0.58 ± 0.16	0.58 ± 0.15	0.58 ± 0.15	0.54 ± 0.14
Median DSC (IQR)	0.58 (0.27)	0.58 (0.26)	0.59 (0.25)	0.59 (0.24)	0.60 (0.22)	0.58 (0.19)

Table 1 shows that the average absolute percent difference of volume consistently decreased with a lowered threshold, further demonstrating the CNN's undermeasurement of the tumor at the 0.5 threshold. This trend did not occur at the 0.01 threshold, with comparable DSCs but larger percent difference. The average DSC peaked at the 0.3 threshold, while the median DSC reached its maximum at 0.1. Figure 4 below displays all the relevant p-value comparisons performed for the percent difference of volume and DSC values. Absolute percent difference of volume was affected more by changes in threshold as compared with DSC.
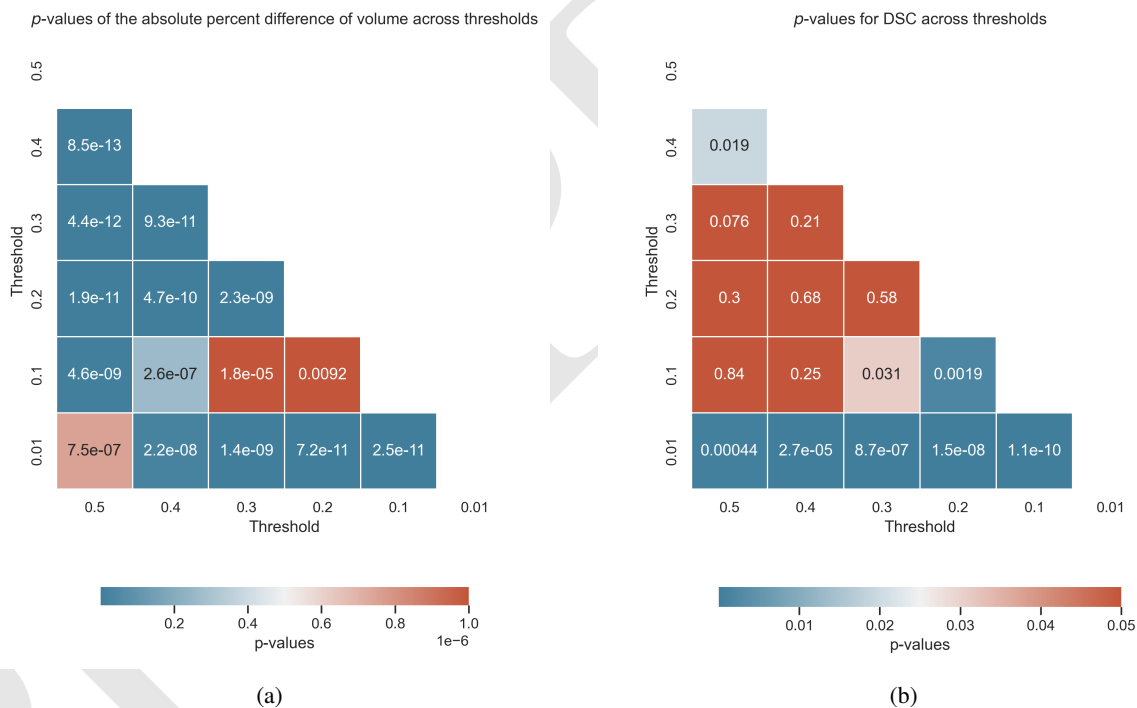

Figure 4: Matrix displaying the p-values when comparing the absolute percent difference of volume (a) and DSC (b) across thresholds.

Figure 5 shows the frequency of a threshold being chosen as "optimal," ranging from 0.001 to 0.7. An optimal threshold differed based on the metric. For example, a certain threshold being optimal for volume meant this was the threshold at which we acquired the lowest percent difference of tumor volume between the radiologist's and CNN contours; similarly, an optimal threshold for DCS indicated that it maximized DSC. Beside the substantial peak present at a threshold of 0.5 for the DSC, there does not appear to be a distinct pattern of "best" values.
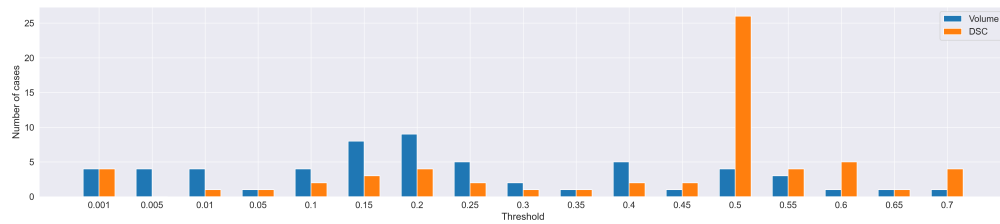
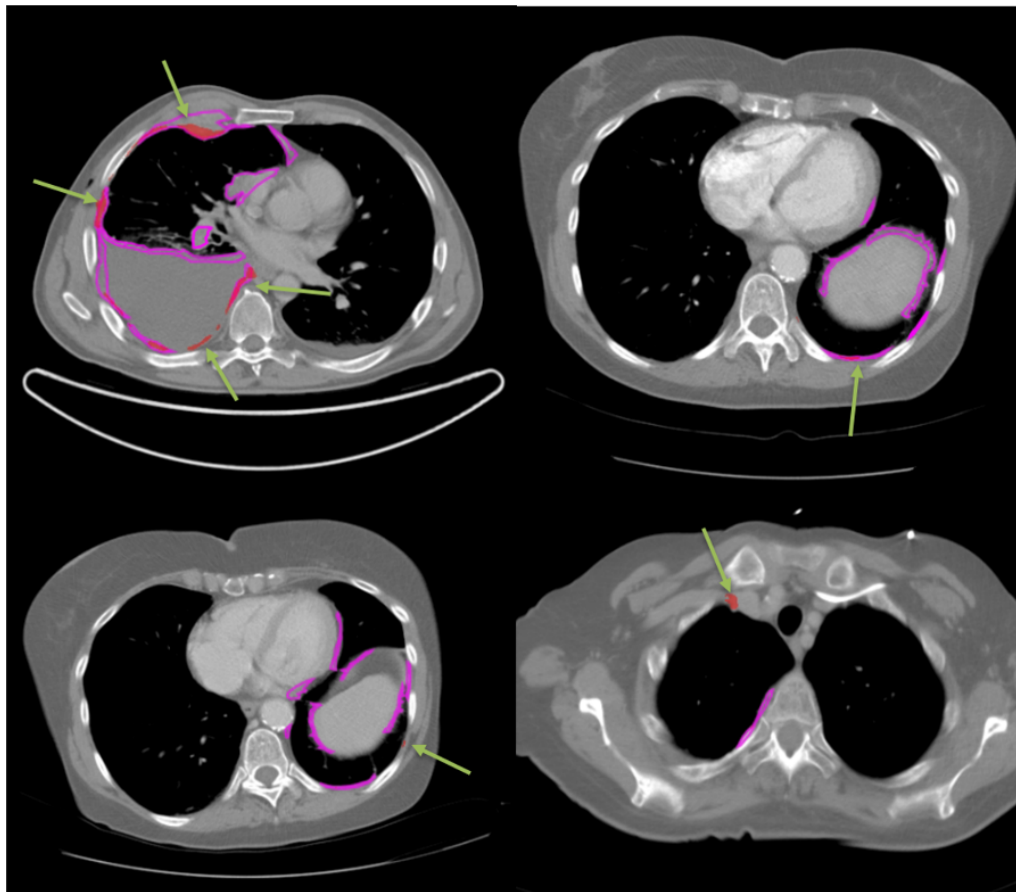Figure 5: Histogram of the CNN output thresholds that "optimize" for DSC and percent difference of volume.

Figure 6: Example images of the four cases that were greater than the 95% agreement limits as shown in the Bland-Altman plot (Figure 3). Green arrows point to regions where the CNN predicted tumor at the 0.5 threshold. Purple outlines are the radiologist's reference contours.

Figure 6 displays example images from four scans (three patients) that were greater than the 95% agreement limits as presented in the Bland-Altman plot in Figure 3. The CNN used was trained on cases that were applicable to mRECIST measurements and was not designed to consider tumors present near the diaphragm, near the lung apices, or invading other parts of thoracic anatomy. Therefore, it unreliably identified tumor for cases with severe pleural effusion as well as disease superior to the aortic arch and inferior to the pulmonary vein, where other anatomic structures complicate the morphology of MPM.

To exclude such regions, we performed the volume and DSC comparisons only considering CT sections inferior to the aortic arch and superior to the pulmonary vein. Figure 7 and Table 2 parallel Figure 2 and Table 1, showing DSC and volume values for the subset analysis. Figure 7 displays the DSC across the same six thresholds analyzed, and Table

2 displays the average absolute percent difference of volume along with average and median DSC values across the thresholds. The range of DSC values is also smaller for the subset analysis (range from 0.24 to 0.92 for the subset as opposed to 0.20 to 0.90).
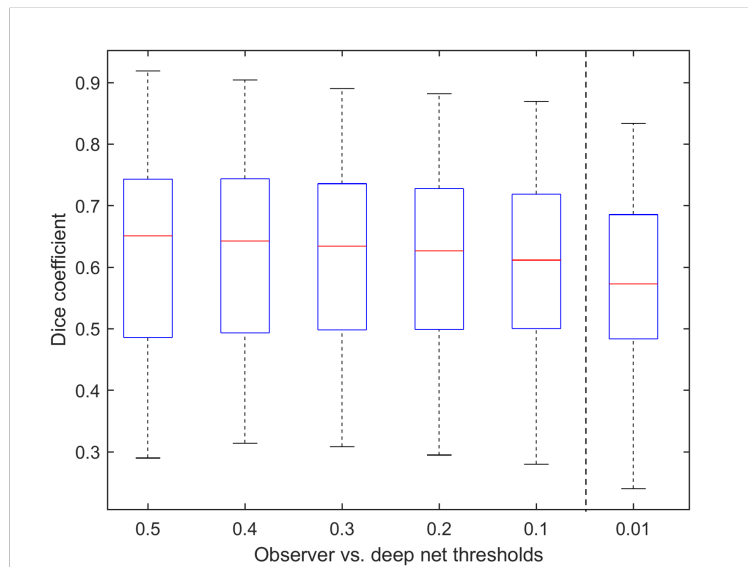

Figure 7: Boxplots showing the DSC values obtained for the subset tumor comparisons acquired between the radiologist and the deep CNN at six different thresholds. Red lines display the median of the data.

Table 2: Absolute percent difference ($\pm$ standard deviation) of volume, average, and median DSC at six thresholds for the subset analysis. (IQR = interquartile range).

	Threshold = 0.5	Threshold = 0.4	Threshold = 0.3	Threshold = 0.2	Threshold = 0.1	Threshold = 0.01
Average absolute % difference of volume	44.00 ± 31.10	37.25 ± 29.10	30.81 ± 27.12	25.35 ± 24.08	22.18 ± 19.26	30.55 ± 19.39
Average DSC	0.62 ± 0.18	0.62 ± 0.17	0.62 ± 0.16	0.62 ± 0.15	0.61 ± 0.15	0.57 ± 0.14
Median DSC (IQR)	0.65 (0.26)	0.64 (0.25)	0.63 (0.24)	0.63 (0.23)	0.61 (0.22)	0.57 (0.20)

Overall, the subset analysis displayed smaller volume percent differences and narrower standard deviations when compared with values in Table 1. Further, the average and median DSCs were consistently larger than those acquired when evaluating the entirety of the scan. The percent volume difference and median DSC value at the 0.01 threshold, however, performed worse for the subset. For instance, the percent difference of 30.55±19.39% was larger than 28.16±18.41% and the median DSC of 0.57 (0.20) was lower than 0.58 (0.19). Though, the average DSC of 0.57±0.14 for the subset was larger than 0.54±0.14 when considering the entirety of the scan at the 0.01 threshold.

4   Discussion

At lower probability thresholds, more pixels were counted as tumor within a CT section, as demonstrated in Figure 1. As a result of a lowered threshold, the computed tumor volume increased (Equation 1). Furthermore, except for eight cases, the radiologist's volumes were consistently larger than those of the CNN (Figure 3), indicating that the CNN did not capture the tumor to its true extent at the default threshold of 0.5. Lastly, as presented in Table 1, the absolute percent difference in volume significantly decreased (Figure 4a) with a lowered threshold, indicating an increase of calculated tumor volume.

While the absolute percent difference decreased at smaller thresholds, the average and median DSC were not different, both within 0.58-0.60. These values are slightly higher than in the current literature [16]. Therefore, the accuracy of contour overlap did not suffer with the change of thresholds. In other words, the newly included pixels at lower

thresholds were not at arbitrary locations within the scan but were, on average, at relevant anatomic regions that overlapped the reference contours, which demonstrated the robustness of the CNN in identifying tumor, even for lower pixel probabilities. A limitation in this analysis was the inherent bias present for the DSC calculation: rather than delineating the tumor on the original CT scans, the radiologist modified the contours generated by the CNN at a threshold of 0.5, which explains the distinct peak at the 0.5 threshold when seeking to maximize DSC (Figure 5).

We obtained large 95% limits of agreement due to the considerable standard deviation of percent differences of volume. Four cases exceeded the 95% limits of agreement (Figure 6). The CNN failed to capture disease that surrounded organs such as the spleen, vertebral column, and heart in those scans. For these cases, the radiologist provided new contours to capture regions of tumor excluded by the CNN at the 0.5 threshold, to establish a reference standard, instead of modifying a preexisting CNN output. Severe pleural effusion was erroneously identified as tumor. Overall, these four examples underly major trends of where the model underperformed. Upon closer examination of the images, the CNN had difficulty contouring disease for the aforementioned reasons along with presence of metallic artifacts, disease in the fissure, and disease present superior to the aortic arch and inferior to the pulmonary veins.

Poor performance was expected for regions that were at the level of the diaphragm or superior to the lung apices because the model was not trained using such regions. Rather, the model was evaluated using tumor contours applicable to mRECIST measurements. To account for this discrepancy, we performed a subset analysis on sections only between the aortic arch and superior to the pulmonary veins. Beside the 0.5 and 0.01 threshold, the percent differences of volume were smaller, and the DSCs were higher across the six thresholds evaluated (except for the 0.01 threshold where the entire scan's median DSC was slightly higher). Further, the median DSC was highest at the 0.5 threshold and decreased with lower thresholds. This may be due to the training of the model once more, as it was trained using the 0.5 threshold. Therefore, we acquired performance comparable to what the model was trained on previously. It is also important to note that the 0.01 threshold acquired the lowest DSC and increased the percent volume difference for the whole scan along with the subset analysis. This indicated that the delineations at low thresholds erroneously considered a substantial number of pixels as tumor, pixels that encroached on anatomy in the region. Overall, for both the subset and whole scan, the percent difference of volume, average DSCs, and median DSCs decreased with a decrease in threshold, with the subset analysis outperforming the entire scan for both metrics.

There are some limitations that should be addressed in this work. First, as previously mentioned, by displaying for the radiologist contours that had been generated at the 0.5 threshold, there was an inherent bias: this bias has been shown to impact the modified outlines produced by observers [8]. The 0.5 threshold was chosen a priori as it is often selected because it is an intuitive value to binarize output probability maps. The aim of this study was to determine the impact of the probability threshold on the two figures of merit studied, not necessarily to study the clinical implementation or the generalizability of a given threshold; therefore, this study was not hindered by having a reference standard with only a single radiologist. A second limitation was the performance of the segmentation task using a 2D CNN architecture as opposed to more advanced techniques, e.g., a 3D architecture. This is will be explored in future work, as the current dataset size may restrict model complexity (compared with a 2D architecture) and result in poor performance when implementing a 3D architecture.

Overall, the purpose of this work was to study the impact of CNN probability map thresholds on the percent volume differences and DSCs when comparing CNN-generated mesothelioma tumor outlines with a radiologist's tumor outlines. These results indicate that the investigation is slightly more nuanced, as lowering the probability threshold (1) predictably increased the resulting tumor volume (which was consistently smaller than the radiologist's) and lowered the percent difference but (2) only negligibly affected the DSC. It is important to make the distinction between CNN volume measurement and DSC. Ensuring that the CNN acquires volume comparable to the reference standard is critical, as volume can be used to capture tumor burden more completely for response assessment. However, we must also be cognizant of the spatial regions where the CNN identifies tumor; it is not sufficient only to match volumes with the reference standard, but also to match the location of the contours. Future work will train the CNN using cases for which this current model was deficient, providing it with cases that displayed pleural effusion and disease surrounding the various structures in the thorax.

5 Conclusion

This study explored the impact of changing the threshold applied to the probability maps output by a CNN when segmenting MPM tumors on CT scans. After investigating thresholds from 0.001 to 0.9, a clear peak at the 0.5 threshold was found for DSC; however, there was no definitive threshold value to minimize the percent difference of volume between the radiologist's and CNN outlines. The percent differences of volume decreased when lowering the probability threshold, while the median DSC did not significantly change. The CNN performance was deficient on scans that contained severe pleural effusion and disease that bordered other structures in the thorax. Therefore, a subset analysis

was conducted, which yielded improved results for difference of volume and DSC. Overall, this pilot study highlighted the impact of varying CNN-generated probability map thresholds on mesothelioma tumor outlines, using percent difference volume and the DSC as the figures of merit.

Acknowledgments

References

[1] M.V. Gerwenn, N. Alpert, A. Wolf, N. Ohri, E. Lewis, K.E. Rosenzweig, R.Flores, and E. Taioli. Prognostic factors of survival in patients with malignant pleural mesothelioma: An analysis of the national cancer database. *Carcinogenesis*, 40(4):529–536, 2019.

[2] M.J. Byrne and A.K. Nowak. Modified recist criteria for assessment of response in malignant pleural mesothelioma. *Annals of Oncology*, 15(2):257–260, 2004.

[3] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, and et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009.

[4] H.I. Pass, B.K. Temeck, K. Kranda, S.M. Steinberg, and I.R. Feuerstein. Preoperative tumor volume is associated with outcome in malignant pleural mesothelioma. *The Journal of Thoracic and Cardiovascular Surgery*, 115(2):310–318, 1998.

[5] D.J. Murphy and R.R. Gill. Volumetric assessment in malignant pleural mesothelioma. *Annals of Translational Medicine*, 5(11):241–241, 2017.

[6] E. Gudmundsson, C.M. Straus, and S.G. Armato III. Deep convolutional neural networks for the automated segmentation of malignant pleural mesothelioma on computed tomography scans. *Journal of Medical Imaging*, 5:034503, 2018.

[7] E. Gudmundsson, C.M. Straus, F. Li, and S.G. Armato III. Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion. *Journal of Medical Imaging*, 7:012705, 2020.

[8] W.F. Sensakovic, A. Starkey, R. Roberts, C.M. Straus, P. Caligiuri, M. Kocherginsky, and S.G. Armato III. The influence of initial outlines on manual segmentation. *Medical Physics*, 37(5):2153–2158, 2010.

[9] A.Z. Dudek, X. Wang, L. Gu, S. Duong, T.E. Stinchcombe, R. Kratzke, H. Borghaei, E.E. Vokes, and H.L. Kindler. Randomized study of maintenance pemetrexed versus observation for treatment of malignant pleural mesothelioma: CALGB 30901. *Clinical Lung Cancer*, 21(6), 2020.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, page 234–241, 2015.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[12] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[13] A.A. Taha and A. Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1), 2015.

[14] F.J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[15] J.M. Bland and D.G. Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999.

[16] A.C. Kidd, O. Anderson, G.W. Cowell, A.J. Weir, J.P. Voisey, M. Evison, S. Tsim, K.A. Goatman, and K.G. Blyth. Fully automated volumetric measurement of malignant pleural mesothelioma by deep learning AI: Validation and comparison with modified recist response criteria. *Thorax*, 77(12):1251–1259, 2022.