**LUNG CT IMAGE TUMOR SEGMENTATION FOR NON-SMALL-CELL LUNG CANCER USING DEEP LEARNING**

*Roy Huang1  Daniel Meggo2 Jordan Mimms2*

From the 1. University of Iowa Department of Electrical and Computer Engineering,

2. University of Iowa Department of Biomedical Engineering

Lung cancer remains the leading cause of cancer-related deaths worldwide, with non-small-cell lung cancer accounting for most cases. Early detection and accurate segmentation of lung tumors from chest CT scans are critical for improving patient outcomes. Manual segmentation, however, is time-consuming and prone to human error. In this project, we propose an automated lung tumor segmentation approach using deep learning, specifically U-Net architecture, to identify and delineate tumors in CT images. Our model utilizes a dataset of 62 patients, consisting of 1,648 cancerous slices in axial plane. We compare four models, each trained with different loss functions and layer normalization strategies. We explored various training techniques, such as data augmentation, early stoppage, and learning rate scheduler. Our results successfully demonstrated the effectiveness of U-Net in segmenting lung tumor with high accuracy, using the Dice Similarity Coefficient (DSC) for evaluation.

# 1. Introduction

Lung cancer is the leading cause of cancer-related deaths globally and the most common form of cancer [1]. Non-small-cell lung cancer (NSCLC) is the most common type of lung cancer, accounting for about 80% of cases [2]. Early and accurate detection of NSCLC tumors using CT (Computed Tomography) imaging is critical for improving patient outcomes [3]. However, manually segmenting lung tumors in CT scans is labor-intensive, time-consuming, and prone to subjective decisions based on the radiologist’s experience. Automated segmentation applications can assist radiologists in making more informed decisions to produce faster, more consistent, and precise results, ultimately enhancing patient care through early-stage detection, accurate diagnoses and personalized treatment planning [4].

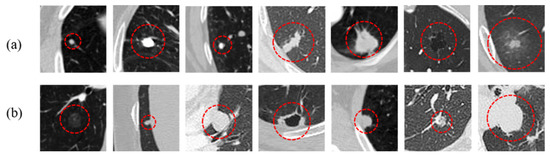


Fig. 1: Lung nodule examples [9]

The U-Net architecture is a deep learning model that has shown great efficacy in the segmentation of features across various imaging modalities [5]. When performing segmentation, a deep learning algorithm must use the spatial information provided within an image to assign each voxel with an appropriate label. The amount of spatial information used to make this decision is closely related to the receptive field of the model. In the U-Net architecture, convolution kernel size and the number of max-pooling layers are two hyper-parameters that contribute to this receptive field. In this study, we investigate the effect of different loss functions on the detection and segmentation of lung cancer lesions in 2D chest CT slices.

# 2. Dataset

Our dataset, consisting of chest CT images, was acquired from Saeid Rasouli, MD, who preprocessed the original dataset sourced from the Medical Segmentation Decathlon [6][7]. This dataset includes scans from 62 patients. During preprocessing, the intensity values of the images were normalized by dividing by the maximum intensity (3071), resulting in a normalized range of -1 to 1. Each scan was cropped to isolate the lung region, reducing unnecessary data. To optimize computational efficiency, each CT volume was split and stored as 2D slices, resized to (256, 256), and saved as numpy files in either the training or validation folders.

In total, the entire dataset contains 15767 axial slices, of which 1648 are cancerous and 14119 slices are non-cancerous. We utilized only slices containing cancer which gave us access to 1648 total slices. For the purposes of this study, the data was split into three categories: training, validation, and testing. The training dataset consisted of 1153 slices and accounted for 70% of our total data. The validation (247 slices) and testing (248 slices) datasets accounted for 15% each of our total dataset.

References for tumor segmentation on each slice are provided as binary label masks, where pixels contain cancerous nodules are labeled with a value of 1, other pixels labeled with 0.

A close-up of a scan

Description automatically generated

Fig. 2: An image and its corresponding mask in our dataset

# 3. Approach

In this project, we created 4 models for comparison, each using a modified U-Net architecture [5]. In the bottle neck layer, instead of cropping the spatial dimensions like the original U-Net, we preserve the spatial dimensions. Our output segmentation map consists of only 1 channel.

A diagram of a diagram

Description automatically generated

Fig. 3: Diagram of U-Net Architecture [5]

Our models are listed below:

1. Model 1: U-Net, trained with Dice Loss + Focal Loss [11] [10]
2. Model 2: U-Net, with layer normalization after each convolution layer, trained with Dice Loss + Focal Loss
3. Model 3: U-Net, with layer normalization after each convolution layer, trained with Dice Loss
4. Model 4: U-Net, with layer normalization after each convolution layer, trained with Focal Loss

Model 1 and Model 2 will compare the effects of adding layer normalization after each convolution layer; Model 2, Model 3, and Model 4 will compare the effects of each loss function on the training process. Our initial idea was to compare only the effects of different loss functions. However, after some experimenting, we realized that layer normalization will significantly improve the training results.

# 4. Training

Some training techniques we experimented included data augmentation, early stoppage, and learning rate scheduler.

Data augmentation was utilized to provide more variety in data since our dataset was relatively small, due to us only using cancerous slices. Each of the following transformations could happen with 50% chance: flipping image along x-axis, flipping image along y-axis, adding Gaussian noise, or rotating 90 degrees.

Early stoppage refers to the training process being stopped once validation loss stagnates after a set number of epochs. Once the training process is stopped, the best model state will be reloaded into the model. All our models have patience of 10, meaning if improvement does not happen for 10 consecutive epochs on validation loss, our model will execute early stoppage. This allows our models to only train for as long as they improve and prevents wasting computation time and energy consumption.

The learning rate scheduler works to adjust the learning rate between epochs for better model training. All our models have patience of 1 for learning rate, meaning if improvement does not improve on validation loss, the model will wait for 1 epoch before adjusting the learning rate. So, if validation loss does not improve for 2 consecutive epochs, the learning rate will adjust for the next epoch. Our learning rate adjustment is set to 0.9 (*new learning rate =0.9 \* previous learning rate*).

For optimizer, all 4 models use the Adam optimizer. Hyperparameters are all the same across all 4 models, ensuring that we only compare the effects of the differences between each model.

# 5.Results

Models 1, 2, 3, and 4 required 99, 93, 146, and 75 epochs to reach convergence, respectively. The receiver operating characteristic (ROC) curve for each model is shown in Figure 4.

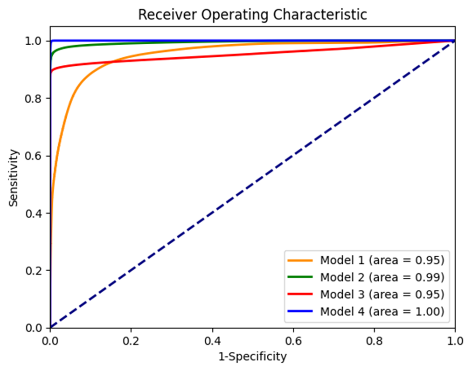


Fig. 4. Receiver operating characteristic curves of the four models.

The test dataset was provided as input to each trained model and their performance was evaluated using the Dice Similarity Coefficient (DSC). Two threshold selection criteria were also used with each model for comparison. The first criterion (Criterion I) selected a threshold that maximizes the difference between the true positive rate (TPR) and the false positive rate (FPR). This corresponded to the upper-left corner of the ROC curve. The second criterion (Criterion II) selected a threshold from the set {0.0, 0.1, 0.2, …,0.9, 1.0} that maximized the DSC. The results of these tests are shown in Table 1.

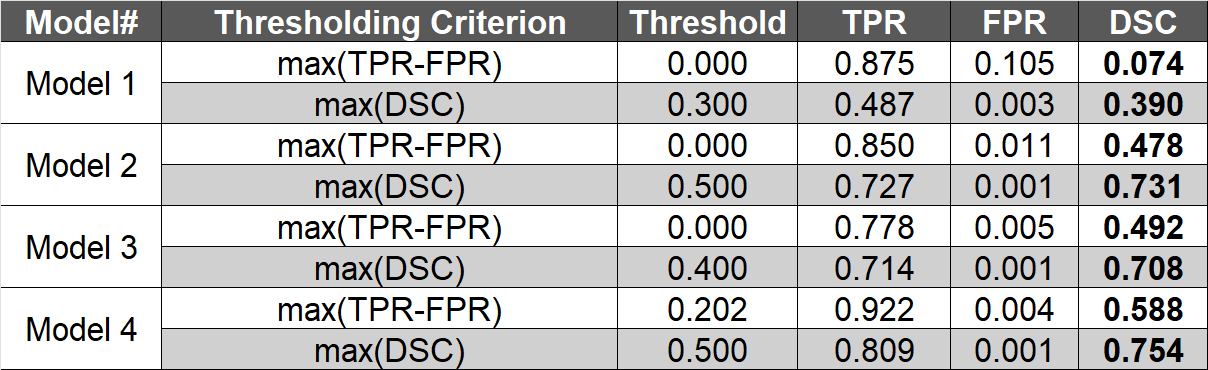


Table 1. Results of each model when evaluated using the test dataset.

For all four models, the thresholding criterion, max(*DSC*), outperformed the alternate criterion, max(*TPR-FPR*). Moreover, with the max(*DSC*) criterion, model 4, which utilized  both layer normalization and the focal loss function, was the best performing model with a DSC of 0.754. Using the same thresholding criterion, the DSC for models 1, 2, and 3 were 0.390, 0.731, and 0.708, respectively. An example of the segmentation results is shown for each model in Figure 5.

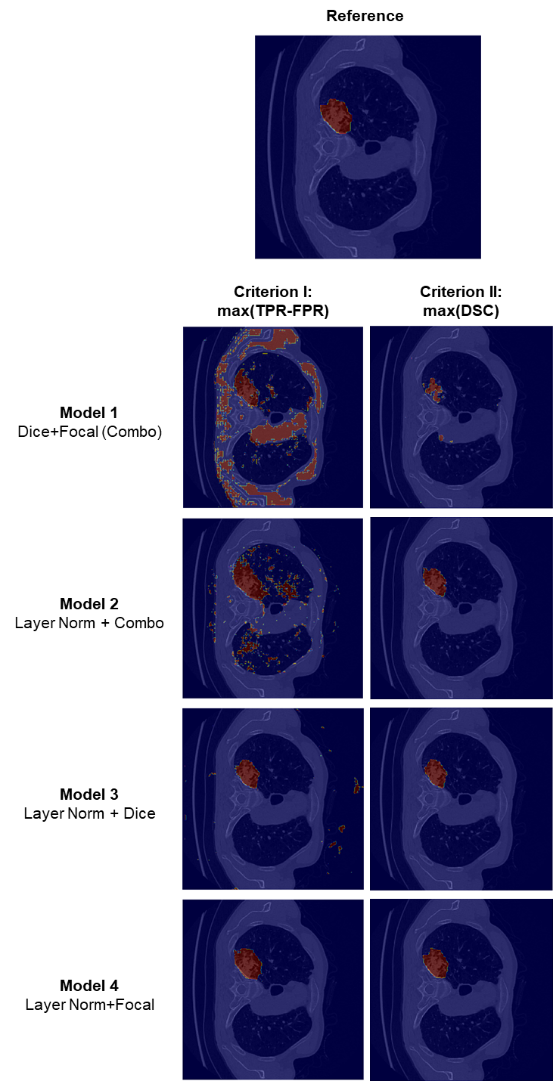


Fig. 5. Reference image and segmentation results of each model.

# 6.Discussion

Due to the massive class imbalance in our datasets, the focal loss and Dice loss functions were used for the segmentation task. The focal loss function addresses class imbalance by down-weighting the contribution of easy to classify features (i.e. the many background pixels present within the chest CT slices) to the loss function such that greater emphasis is placed on correcting features that are harder to classify (i.e. the few pixels that make up cancerous tumors in each slice) during optimization. On the other hand, the Dice loss function, based on the Dice Similarity Coefficient (DSC), inherently possesses a good balance between precision and recall, making it suitable for cases of class imbalance. Each of these loss functions comes with its own pros and cons, so to take advantage of each of their strengths, the two loss functions were combined and used with the U-Net architecture in model 1. However, as shown in Table 1 and Figure 5, model 1 had the poorest performance of all models. Hence, on their own, the loss functions were unable to segment the lung nodules appropriately. This prompted the addition of layer normalization in model 2 which helped to stabilize the model, smooth its convergence, and dramatically improve its performance. In fact, of the different strategies employed in this study, the addition of layer normalization was associated with the greatest improvement in segmentation results.

With layer normalization, each loss function was then tested individually in models 3 and 4 to observe their independent performances. It was originally believed that the individual implementations of the loss functions would perform worse than their combination. However, to our surprise, the model using focal loss with layer normalization (model 4) outperformed the model using the combination of focal loss and Dice loss functions with layer normalization (model 2). The reason for this observation is still unclear and should be studied further.

The selection of an appropriate threshold for the probability maps output by the models was perhaps the greatest challenge faced on this project, and this was once again due to the issue of class imbalance. The imbalance between background and cancerous lesions resulted in very low false positive rates despite the misclassification of several background pixels as cancerous. Additionally, since perfect segmentation of a tumor and over segmentation of that same tumor are both associated with a high true positive rate, Criterion I was not very sensitive to increases in the number of false positives and did not generate good segmentation results. On the other hand, Criterion II, which used the DSC, was more robust to this class imbalance and therefore produced significantly better results. In retrospect, using precision-recall curves instead of ROC curves for the selection of threshold values probably would have produced even better results, however, the investigators did not have the opportunity to try this. This should be attempted in future implementations.

Limitations to this study include the fact that the dataset was already preprocessed and did not have metadata associated with it. As a result, further post-processing such as the calculation of tumor size was not possible. In addition, since segmentations were only performed on two-dimensional slices, the model did not make full use of the three-dimensional spatial information provided by medical images. Since information about the third dimension was missing in our images, the model struggled to distinguish cancerous lesions from blood vessels. The use of a 3D U-Net with 3D CT volumes probably would have improved the performance of our models. However, this would have come at the cost of increasing the computational load for training. Other improvements to the model include increasing the batch size and using batch normalization, as well as implementing K-fold cross validation to improve generalizability. Finally, in the present study, the focal and dice loss functions were weighted equally in their contribution to the overall loss function. In future implementations, it would be interesting to see if weighting the contribution of each function to the overall loss using a formula such as Loverall = βLfocal + (1-β) LDice, can improve model performance.

# 7. Acknowledgments

We gratefully acknowledge our professor, Dr. Sarah Gerard, and course teaching assistant, Joshua Carrizales, for their guidance and support throughout this project. We also extend our thanks to the University of Iowa College of Engineering for providing the essential computing resources needed to carry out our work.

# 8. References

[1] Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today, accessed [18 October 2024].

[2] American Cancer Society. Cancer Facts & Figures 2024. Atlanta: American Cancer Society; 2024.

[3] Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W. E., Nicholson, A. G., Groome, P., Mitchell, A., Bolejack, V., International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions, & International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee Advisory Boards and Participating Institutions (2016). The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer, 11(1), 39–51. https://doi.org/10.1016/j.jtho.2015.09.009

[4] Riaz, Z., Khan, B., Abdullah, S., Khan, S., & Islam, M. S. (2023). Lung Tumor Image Segmentation from Computer Tomography Images Using MobileNetV2 and Transfer Learning. Bioengineering (Basel, Switzerland), 10(8), 981. https://doi.org/10.3390/bioengineering10080981

[5] O. Ronneberger, P. Fischer, T. Brox, (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351.Springer , Cham. https://doi.org/10.1007/978-3-319-24574-4\_28

[6] M. Antonelli, A. Reinke, S. Bakas, et al. “The Medical Segmentation Decathlon,” Nature Communications, vol. 13, no. 4128, 2022. https://doi.org/10.1038/s41467-022-30695-9

[7] S. Rasouli (2023). Lung Cancer Segment [Data set]. Kaggle. https://www.kaggle.com/datasets/rasoulisaeid/lung-cancer-segment

[8] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016. Available: http://www.deeplearningbook.org

[9] Tang, T.; Li, F.; Jiang, M.; Xia, X.; Zhang, R.; Lin, K. Improved Complementary Pulmonary Nodule Segmentation Model Based on Multi-Feature Fusion. Entropy 2022, 24, 1755. https://doi.org/10.3390/e24121755

[10] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). Focal loss for dense object detection. arXiv. arXiv:1708.02002. https://arxiv.org/abs/1708.02002

[11] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). 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 (pp. 240–248). Springer International Publishing. https://doi.org/10.1007/978-3-319-67558-9\_28