Comparative Analysis on Multiple Methods to Identify and Segment Lung Cancer Tumors

Salem AlAthari, Zehui Chen, Kat Desai, Manjit Singh

CSC19100 Final Group Report

Abstract.

Lung cancer, affecting 1 in 16 people, necessitates early detection techniques to improve patient outcomes. Deep neural networks, particularly segmentation models, play a crucial role in identifying lung tumors using medical imaging. This study utilizes a dataset derived from the LIDC-IRDI database, consisting of CT scans and corresponding tumor masks, to develop and evaluate four segmentation models: a baseline per-pixel classification model, two 2D segmentation models, and one 3D segmentation model.

The models were trained and tested using Dice Metrics to assess the similarity between predicted tumor regions and ground truth masks. A small testing set was employed to generate predictions and visualize results, providing insights into model effectiveness and hyperparameter tuning.

In the background, existing literature on medical imaging segmentation was reviewed, highlighting the performance of models like DeepLabv3+ and MobileNetV2 combined with UNET for efficient and accurate lung tumor segmentation. Data for this study were sourced from the Kazakh Research Institute of Oncology and Radiology and the LIDC-IDRI dataset, comprising 972 CT scans labeled with the Lung-RADS system.

The methodology involved evaluating four deep neural networks. The baseline model performed poorly with a Dice score of 0.1. The UNet with MobileNetV2 achieved high training accuracy but struggled with validation due to small sample sizes and tumor variability. The FPN with ResNet34 encoder showed robust performance, achieving a Dice score of 0.89, precision of 0.95, and recall of 0.9. The UNET-R 3D model, using a Visual Transformer encoder, underwent extensive hyperparameter tuning, achieving a mean Dice score of around 25%.

In conclusion, the FPN with ResNet34 encoder emerged as the most effective model, demonstrating potential to assist radiologists in lung tumor detection. Future work will focus on training with larger, more diverse datasets to enhance model robustness and incorporating tumor classification capabilities.

1. Introduction

The Lung Cancer Research Foundation estimates that 1 in 16 people will be diagnosed with Lung Cancer in their lifetime. Thus the development of techniques to detect Lung Cancer as early as possible is quite important. Segmentation models are very commonly used Deep Neural Networks that the medical industry has both used and had their part in developing. Using

medical imaging technology, the National Cancer Institute had initiated the development of the LIDC-IRDI; a database of lung images that is widely used in Lung Cancer detection.

Using a dataset based off a section of the LIDC-IRDI database, that included both lung images and masks depicting the location of lung tumors, we've tackled the problem of Lung Cancer Detection by developing four different Segmentation models, a baseline model which used a a per-pixel classification basis method, two 2-D Segmentation models to analyze per segmentation of each 2-D image frame, and one 3-D Segmentation model that analyzes the data by splitting the 3-D images into voxels. After normalizing the input images and making sure that the masks would only have binary values, these models were all trained, measured and compared using Dice Metrics, which measures the similarity of a predicted lung-cancer pixel vs the cancer pixels in the associated masks, over the total area (or volume) of both images.

For all models a small testing set, separate from the validation/training data that was put into the model, was used to generate predictions and visualized using the matplotlib. This when combined with the measured data during the testing/validation stage would not only allow us to see how effective each model was, but also allow us to see how these model hyperparameters would need to be tuned.

2. Background

For the research we decided to focus on medical imaging segmentation papers in order to get an understand of what kind of models were considered

In the paper "Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures," the authors explore the performance of various deep learning models for segmenting lung cancer tumors from CT images. The study specifically evaluates the efficiency and accuracy of these models in segmenting malignant lung tumors, which is crucial for the early diagnosis and treatment of lung cancer.The research investigates several state-of-the-art deep learning architectures, including the U-Net, SegNet, and DeepLabv3+, among others. These models are assessed based on their ability to accurately delineate tumor boundaries in CT images. The key findings highlight that DeepLabv3+ achieved the highest performance, with significant improvements in precision and recall compared to the other models. Specifically, DeepLabv3+ demonstrated a superior balance between accuracy and computational efficiency, making it particularly suitable for clinical applications where both factors are critical. Overall, the study concludes that while all the evaluated models show promise, DeepLabv3+ stands out due to its robust performance metrics, which include a high Dice similarity coefficient, precision, and recall. These results suggest that employing advanced deep learning techniques can significantly enhance the segmentation accuracy of lung cancer tumors in medical imaging, thereby aiding in more effective diagnosis and treatment planning

In the paper titled "Lung Tumor Image Segmentation from Computer Tomography Images Using MobileNetV2 and Transfer Learning," the authors propose a hybrid neural network that combines MobileNetV2 and UNET architectures for the segmentation of malignant lung tumors from CT images. The approach leverages transfer learning by using a pre-trained MobileNetV2 as the encoder for feature extraction within a UNET framework. This combination aims to improve efficiency and accuracy in segmenting lung tumors. The model employs lightweight filtering to reduce computational demands and pointwise convolutions to enhance feature building. Skip connections with ReLU activation functions link encoder layers of MobileNetV2 to decoder layers in UNET, facilitating the concatenation of feature maps with different resolutions. The model was trained and fine-tuned using the Medical Segmentation Decathlon (MSD) 2018 Challenge dataset. When tested on a portion of this dataset, the proposed network achieved a dice score of 0.8793, recall of 0.8602, and precision of 0.93. These results indicate that the proposed model outperforms existing methods that require multiple phases of training and testing, highlighting its potential for efficient and accurate lung tumor segmentation in medical imaging

In the paper titled "Deep Learning Ensemble 2D CNN Approach towards the Detection of Lung Cancer," the authors propose a method using a deep learning ensemble of 2D convolutional neural networks (CNNs) to detect lung cancer from CT scans. This approach focuses on leveraging multiple CNN models to enhance the accuracy and reliability of lung cancer detection. The study presented a novel ensemble learning method combining multiple 2D CNN architectures to improve the detection of lung cancer from CT images. By integrating various CNN models, the ensemble approach aimed to capitalize on the strengths of each individual model, thus enhancing overall performance. The experimental results demonstrated that this ensemble method achieved a significant improvement in accuracy compared to single-model approaches. Specifically, the proposed ensemble model exhibited a high detection accuracy, underscoring its potential as a robust tool for early lung cancer diagnosis. This method's success highlights the efficacy of ensemble techniques in medical image analysis and suggests a promising direction for future research in automated cancer detection systems. The study underscores the potential of deep learning ensemble methods in improving diagnostic accuracy, suggesting that incorporating multiple models can effectively address the variability and complexity of medical imaging data. This research contributes to the growing body of evidence supporting the use of advanced AI techniques in enhancing the early detection and diagnosis of lung cancer, ultimately aiming to improve patient outcomes through timely and accurate medical interventions

3. Data

The dataset used for this project is a combination of images sourced from the Kazakh Research Institute of Oncology and Radiology and the publicly available LIDC-IDRI dataset. The images

used are sourced from 972 CT scans of 40 unique lungs that are at various stages of lung cancer progression. Each image is accompanied with a label according to the Lung-RADS system which categorizes tumors by incremental stages of maturity, along with a binary mask which spatially identifies the tumor within the image. Labels and masks have been produced by physicians specializing in radiology.
Each model is trained on a set of 708 of these CT scan images, and model performance on unseen data is evaluated on a test set of 264 images. In cases where hyperparameters of the model are tuned, an additional validation set is produced from a portion of the training images.

4. Methods

Our approach to this task involved evaluating 4 different deep neural networks to understand which model architecture is best able to segment tumors within the lung images. We utilized a simple semantic segmentation model to serve as a baseline, two deep neural networks with pre-trained encoders for 2D segmentation, and UNet-R on collections of CT scans derived from the same lungs for the task of 3D segmentation.

4.1 Simple Semantic Segmentation Model

Based on the UNet architecture we developed a model that would perform semantic segmentation on a CT slice of a lung.The model as a single layer convolutional network that was meant to perform as a baseline to what would be the simplest model. The data was loaded from a pickle file into a pandas dataframe that had data from the ct slice and the ground truth mask. The data was then converted to tensors and normalized at 0.5. The hyper parameters ranged from 0.1 - 0.001 for the learning rate. Adam was used as an optimizer. Dice loss was ultimately chosen but binary cross entropy and mean squared error loss was tested as well.

4.2 UNet with MobileNetV2 Encoder

For this task, I combined a 4-layer deep UNet design with MobileNetV2, which is a pre-trained encoder. The 4-layer deep UNet excels in this task because it captures both local and global details crucial for precise detection. Its skip connections preserve spatial information, aiding accurate tumor localization within segmentation masks. MobileNetV2 was chosen as the encoder due to its efficiency on devices with limited resources, which is crucial given our RAM constraints. This UNet-MobileNetV2 setup aims to balance accuracy and computational efficiency.

First, data was loaded into a custom PyTorch dataset, with each sample consisting of a 2D image. During preprocessing, each sample is normalized by min-max rescaling the pixel values to a range between 0 and 1:

$$I_{Normalized} = \frac{I - I_{min}}{I_{max} - I_{min}}$$

Additionally, each image is resized to 256x256 pixels with random horizontal flips, vertical flips, and rotations to increase the diversity of the dataset and prevent model overfitting.

The model was then trained with a learning rate of 0.0001. A combination loss function of Binary Cross-Entropy (BCE) and Dice loss was used to produce both accurate and spatially coherent segmentations. The Adam optimizer was chosen for its ability to efficiently find the global minima. The model was trained for 240 epochs.

4.3 FPN with ResNet34 Encoder

This modeling approach for 2D per-slice segmentation makes use of a Feature Pyramid Network (FPN) architecture with a ResNet34 pre-trained encoder. The FPN model consists of 4 layers and is particularly performant on small object detection tasks. The ResNet34 encoder is pre-trained on the CIFAR-10 dataset, and the entire model is then fine-tuned on our training set images. The model is trained over 100 epochs with a batch size of 16 and an Adam Optimizer learning rate of 0.0001. Model errors are measured using the Dice loss function.

4.4 UNET-R 3-D Segmentation model

The model used for the 3-D Segmentation task was a model known as UNET-R which is a UNET Model that uses a Visual Transformer as the encoder. This was done on Kaggle as it hosts more RAM than Google Colab, and has easier access to local storage. The version used was from the MONAI Library. Since the Lung Cancer RADS dataset was a pickle file of 2-D images, the data had to be preprocessed into a dictionary of NIFTI file, a common way to store 3-D medical images, and a requirement to make use of the MONAI Library functions.

The pixel images were normalized from [-1024, 1024] to [0,1]. The image groups, (number of image slices in a group represents depth of lung) for each lung then had to be converted into a NIFTI file, each of which would occupy a place in the dictionary, with lung groups that had too small of a depth being filtered out.

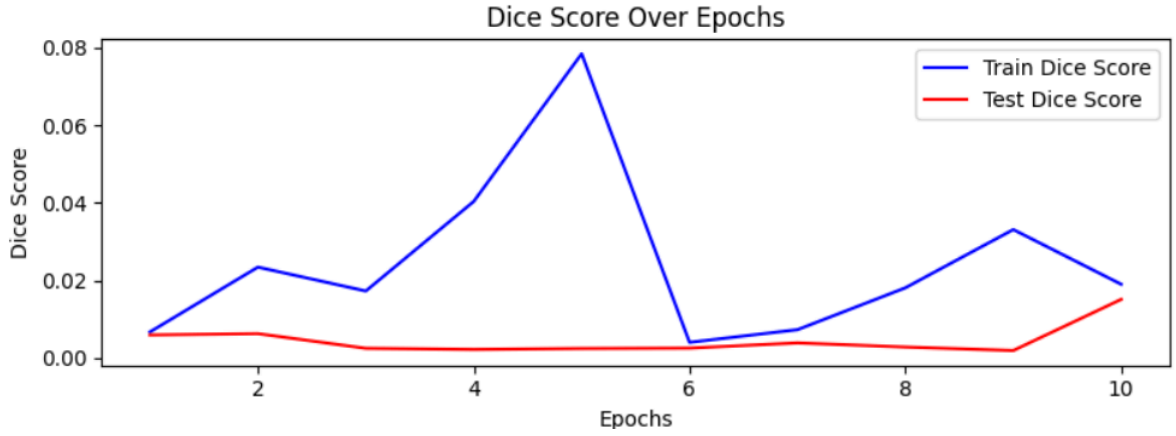
The MONAI Library Transformations were then used to make each 3-D Lung group have consistent dimensions from (512 x 512) into dimension (128x128x128), in order to keep the voxels consistent and to save on RAM. Since the amount of data we had to train the model was relatively small, each 3-D training image would also be split into 4 partitions of (64x64x64), which made the training dataset dimension to be (4 partitions , 1 channel,64 H,64 W,64 D), as the input for the UNET-R model, which would split the input into (16x16x16) voxels. The training dataset would also occasionally be rotated/flipped.

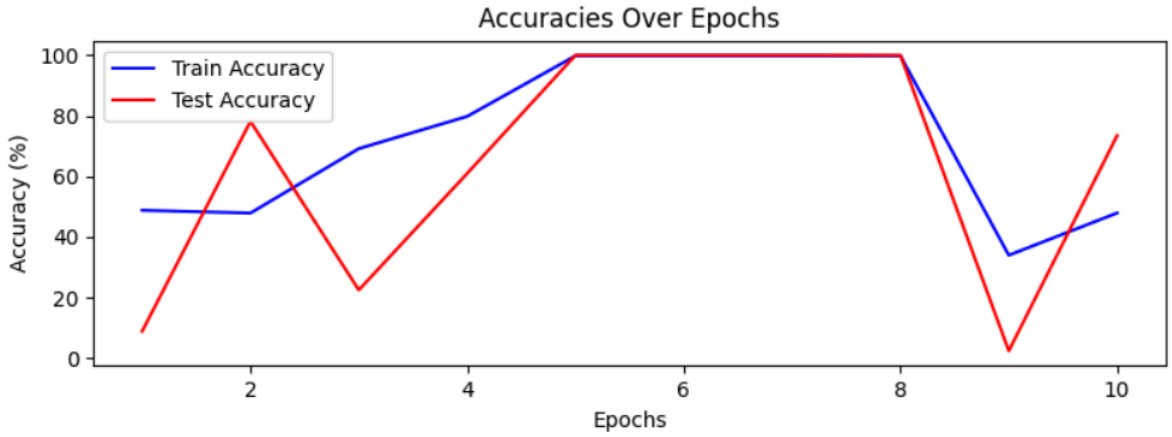
Using the Mean Dice Value function, different hyperparameters, such as weight decay for L2 Regularization, Learning rates, Dropout rates, Dice Loss functions, and the effects of including the background as an output were used to evaluate and tune the model.

5. Evaluation

5.1 Simple Semantic Segmentation Model

The model performed poorly as it was too simple to extract any meaningful features. With a dice score of around 0.1. This was to be expected as the model seemed to be randomly guessing what pixel was to be assigned to what without assigning any significance. The model held a high accuracy however accuracy is a relatively unimportant metric as most of the ground truth mask had background zero value pixels.



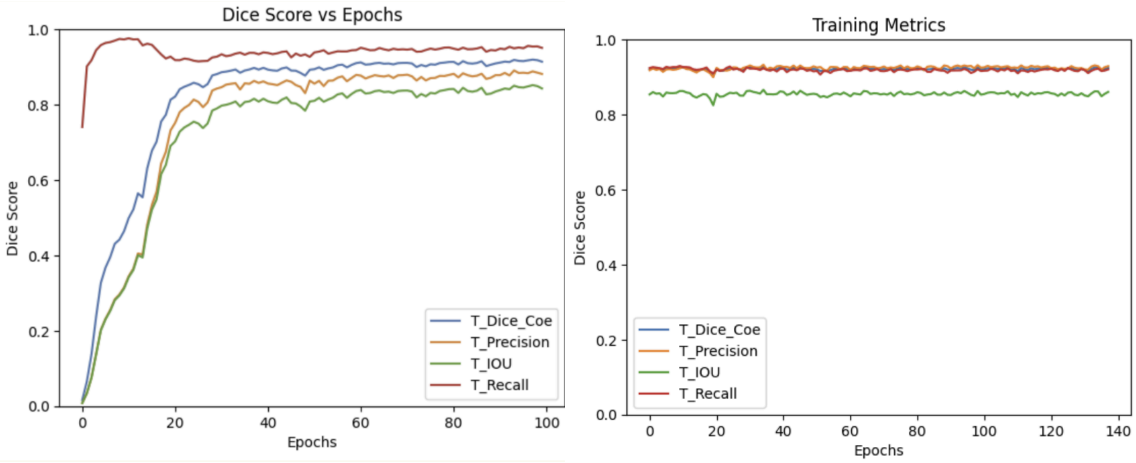


5.2 UNet with MobileNetV2 Model

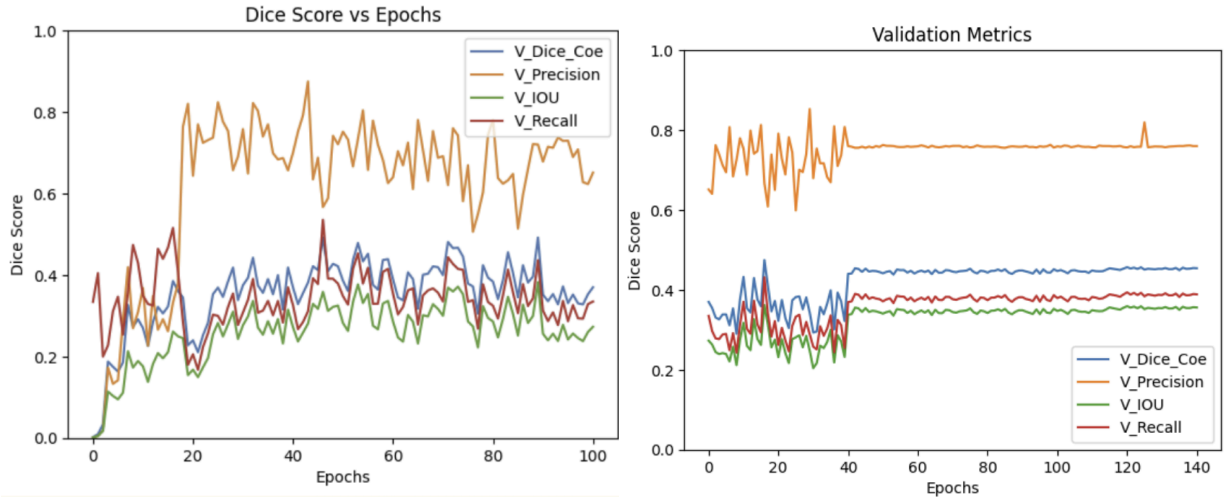
After 240 epochs of training, the UNet with MobileNetV2 model achieves a Dice score of more than 0.90 on the training set but only 0.45 on the validation set. The precision score for

the validation set is relatively high because this model fails to detect tumors in some CT scans with relatively small tumor sizes. This means most of this model's predictions are correct tumor pixels for large tumors, while small tumors are often missed. This model also sometimes predicts a small tumor and sometimes a large tumor when both small and large shaded areas exist in a CT image. This confusion highlights the challenge this model faces when both small and large shaded areas are present in the same slice of the CT scans. Additionally, the low performance on the validation set is due to the imbalance in our dataset and the insufficient amount of data, which consists of only 706 samples in the training set and 264 samples in the validation set.

Training Evaluation



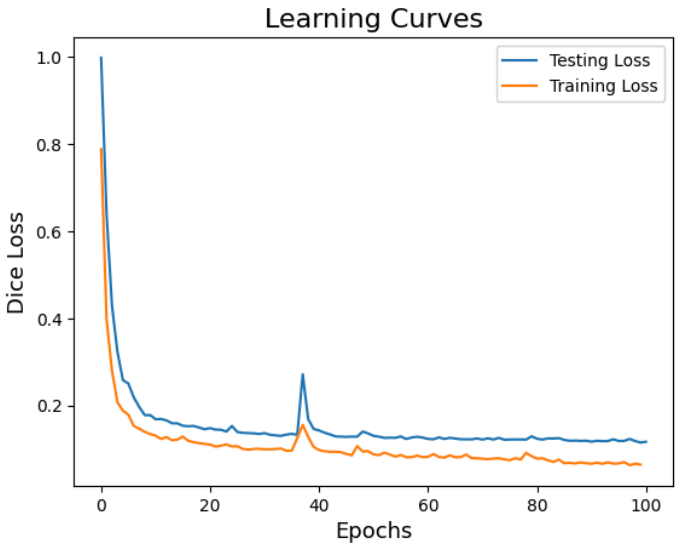
Validation Evaluation

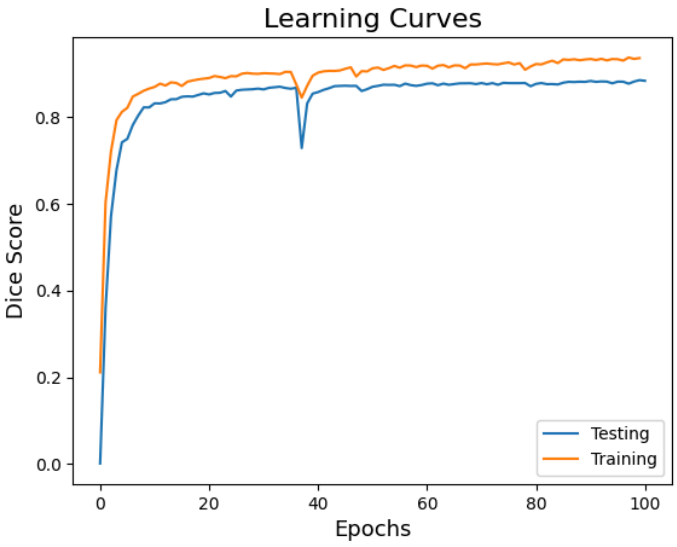


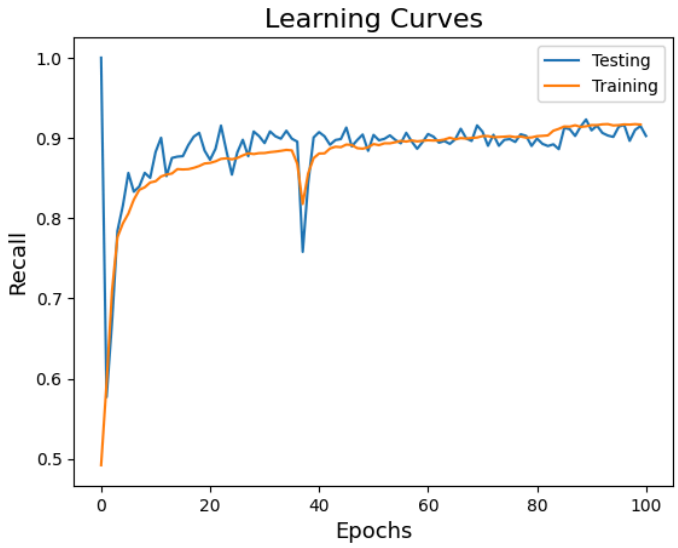
5.3 FPN with ResNet34 Encoder

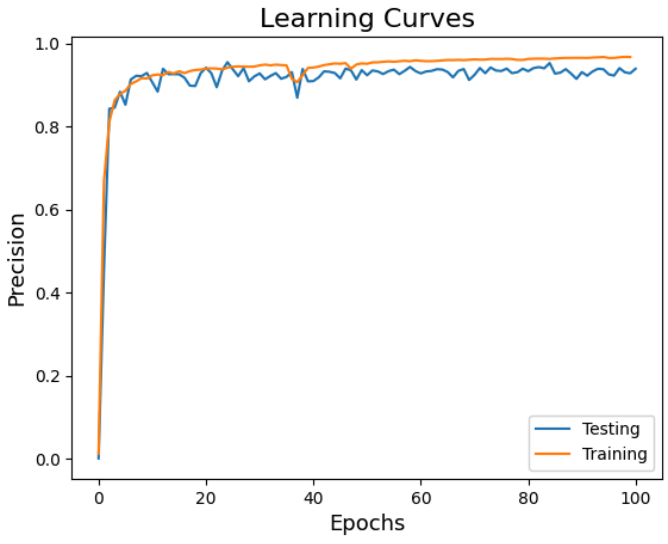
The FPN model is trained for 100 epochs, and performance is measured using Dice Loss, Precision, and Recall. We observe a Test set Dice score of 0.89, Precision of 0.95, and Recall of

0.9. Our testing metrics generally track the training curves, indicating that the model is performant on unseen data and has not yet reached the threshold of overfitting. These metrics are generally consistent with Dice scores achieved in related research.

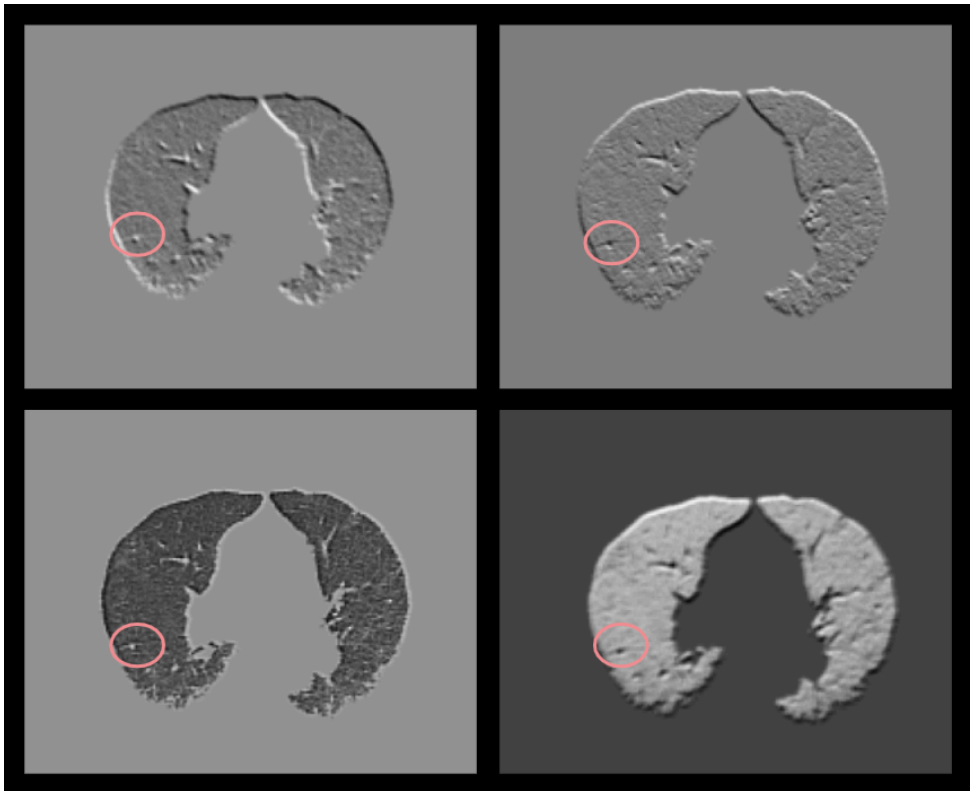




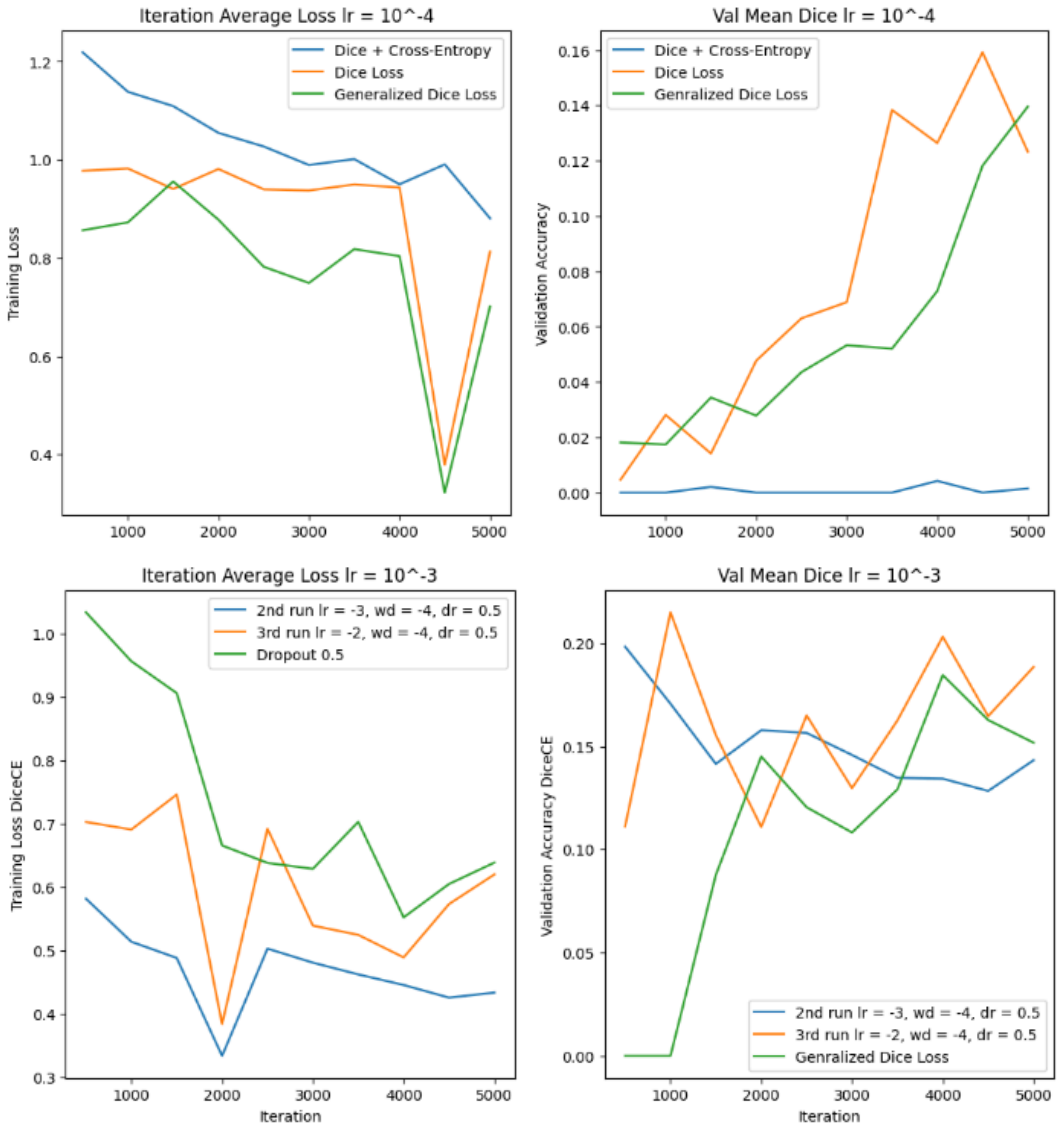




Additionally, we inspected the activations of the first convolutional layer of the FPN model. These feature maps help us to understand which aspects of the lung CT scans are particularly salient to the model and help drive the model's decision making process.



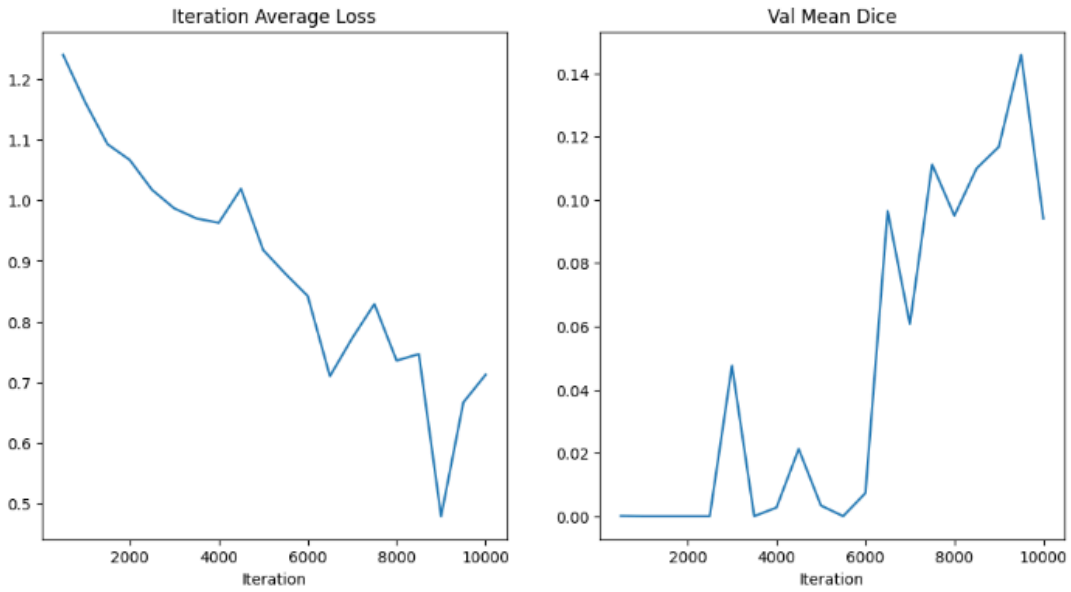
5.4 UNET-R 3-D Model



Most of the Hyperparameters were chosen after running the model for 5000 epochs for each change in the Hyperparameter. The Training Loss function that I've evaluated that yielded the best results was the Dice + Cross Entropy Loss function, as opposed to just Dice Loss or a Generalized Dice Function. Though the training loss (left) starts higher and the validation accuracy takes longer to rise, this Loss function would yield the highest Mean Dice Score when given enough iterations. In this model we've also determined that the model performed best with a learning rate of 10^-4, a weight decay of 10^-5 and a dropout rate around 50%.

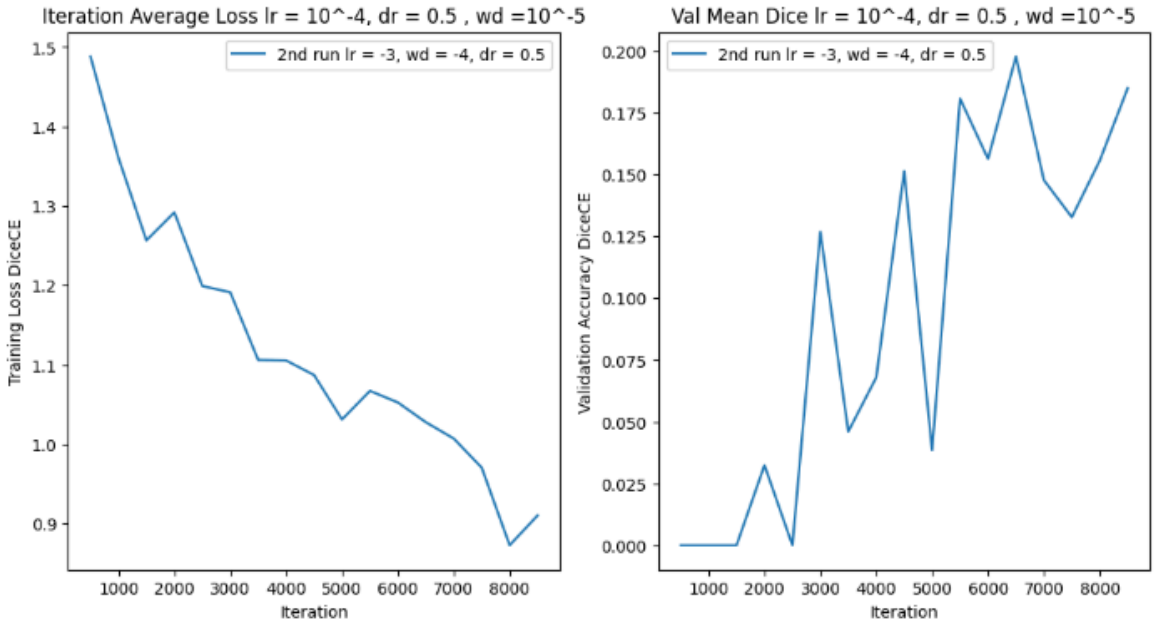
The model was then tested using two approaches on 10,000 epochs, one with 1-channel outputs, and another with 2-channel outputs, with one hot encoding. In the 1-channel approach the weight

tensor for the loss was set to 100 to account for imbalances, and the prediction values before being compared to the masks were discretized with a threshold of 0.15



The other method had the model use a 2-channel output to account for a background classification. The outputs of this model were converted into one-hot encoded values, and the one-hot encoded output was then discretized using an argmax function to compare to the ground truth masks. The weight tensor was also set to [0.1,0.9] (background 10% , Cancer Cell 90%). This model yielded the highest Validation Mean Dice Score of around 25%.

Using the Testing image to visualize the model, we can see that while the model may not properly match the ground truth mask in terms of locating the cancer cell, we do see that the model is able to make out some segmentation of where it thinks the cancer cell is. The reasons for why the segmentation location is off is still being investigated, but this could be due to the data still being unbalanced, which may mean that we may need stronger weights for the cancer cell. We also note that some of the MONAI Transformations can generate padding, which may also contribute to the imbalance of data.



6. Conclusion & Future Work

For this project, we focused on lung tumor segmentation from CT scans using four different models. The FPN with ResNet34 encoder demonstrated the best performance, achieving a Dice score of 0.89 on the test set. This highlights the capability of a well-trained model to assist radiologists in detecting lung tumors from CT scans, a crucial task given that radiologists must review thousands of CT scans daily. This model can help increase the effectiveness and efficiency of reviewing CT scans for radiologists.

However, we trained our models on a small and imbalanced dataset. Consequently, our models are biased towards the dataset and lack diversity in representing different types of lung tumors. To improve the robustness of our models, a much larger and more diverse dataset is needed.

For future work, we will focus on enhancing the model's robustness by using a much larger dataset that includes various types of lung tumors and is more balanced in terms of tumor size. Additionally, we will aim to improve the model's capability not only in segmenting lung tumors but also in classifying the types of lung tumors it detects.

Attribution

Salem AlAthari: I did around 20 code hours of work mainly trying to get models to produce understandable results. I also organized the group and tried to keep everyone on task for the project as well as consulting with the professor.

Zehui Chen: I spent 3 to 5 hours every week researching ways to improve my models. I also dedicate around 10 hours weekly to data processing and model training on Kaggle to experiment with various hyperparameters.

Khyatee Desai: I built the FPN/ResNet34 model, preprocessed the dataset, tuned model hyperparameters, and generated visualizations of the underlying model activations. I spent about 10 hours per week on this project.

Manjit Singh: I worked on the 3-D UNET-R model , preprocessing and cleaning the data to a usable format to make the smaller dataset robust, and experimenting with the model hyperparameters on Kaggle to get better segmentation results. I worked on my part for 6-hours a week.

Bibliography

Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., & Jaeger, P. F. (2024). nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv preprint arXiv:2404.09556.

Primakov, S.P., Ibrahim, A., van Timmeren, J.E. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nat Commun 13, 3423 (2022). https://doi.org/10.1038/s41467-022-30841-3

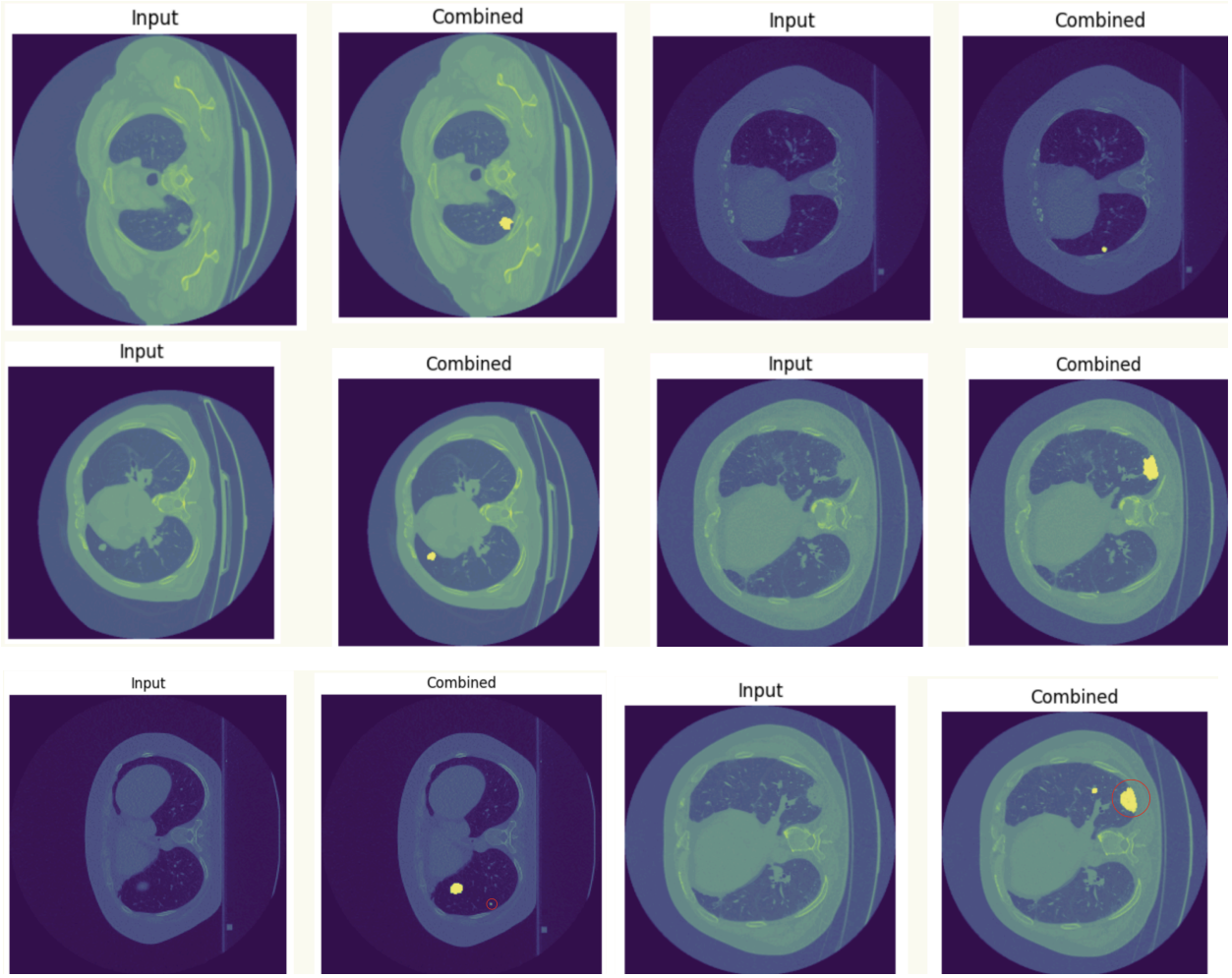
Riaz, Z., Khan, B., Abdullah, S., Khan, S., & Islam, M. S. (2023). Lung Tumor Image Segmentation from Computer Tomography Images Using MobileNetV2 and Transfer Learning. Bioengineering (Basel, Switzerland), 10(8), 981. https://doi.org/10.3390/bioengineering10080981

Said, Y., Alsheikhy, A. A., Shawly, T., & Lahza, H. (2023). Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures. Diagnostics (Basel, Switzerland), 13(3), 546. https://doi.org/10.3390/diagnostics13030546

Hatamizadeh, A., “UNETR: Transformers for 3D Medical Image Segmentation”, arXiv e-prints, 2021. doi:10.48550/arXiv.2103.10504.

Appendix

UNet with MobileNetV2 Predictions



UNET-R 3-D Segmentation: