Deep-Learning Based 3D Lung Segmentation

Kerem Erciyes
Department of Artificial Intelligence for Science and Technology
University of Milano-Bicocca
Milan, Italy
k.erciyes@campus.unimib.it

Mustafa Soydan
Department of Artificial Intelligence for Science and Technology
University of Milano-Bicocca
Milan, Italy
m.soydan@campus.unimib.it

Ozgur Gumus
Department of Artificial Intelligence for Science and Technology
University of Milano-Bicocca
Milan, Italy
o.gumus@campus.unimib.it

***Abstract*—Medical imaging plays a crucial role in diagnosing and treating lung conditions, where accurate lung segmentation in Computed Tomography (CT) scans is a fundamental step in various applications, from disease detection to surgical planning. This study presents a deep learning-based approach for lung segmentation using an artificial neural network trained on a dataset derived from the AAPM Thoracic Auto-segmentation Challenge. The study consists of preprocessing techniques, dynamic advanced data augmentation, and a U-Net-based architecture to enhance segmentation accuracy while mitigating overfitting. The results demonstrate the effectiveness of our approach, achieving high performance on key evaluation metrics.**

Keywords—Lung segmentation, medical imaging, image processing, Convolutional Neural Networks, deep learning, U-Net, Dice Similarity Coefficient, Intersection over Union, TensorFlow, Keras, ReLU, Adam Optimizer.

I. Introduction

Lung diseases, including pneumonia, lung cancer, and COVID-19, necessitate accurate medical imaging analysis. Accurate segmentation aids in disease detection, treatment planning, and surgical navigation. This study focuses on improving lung segmentation using deep learning techniques to enhance accuracy and generalization. As part of a laboratory activity, this work aims to refine an existing 3D segmentation model by reducing overfitting and increasing learning efficiency through advanced data augmentation techniques.

The primary goal of this work was to develop a highly robust lung segmentation model capable of addressing human errors and detector inconsistencies while enhancing the model's generalization ability. This was achieved through extensive data augmentation techniques, ensuring that the model performs well across diverse medical imaging conditions and maintains high accuracy despite variations in input quality. This paper shows the development of a U-Net-based segmentation model optimized for medical imaging datasets.

II. Related work

Lung segmentation in medical imaging has been an active area of research, with numerous approaches proposed to address the challenges of segmenting complex anatomical structures. Early methods were primarily based on traditional image processing techniques, such as thresholding and region-growing, which were often limited by their inability to handle the variability in patient anatomy and imaging conditions. Recent advancements, however, have been driven by deep learning models, particularly convolutional neural networks (CNNs), which have shown significant improvements in segmentation accuracy.

One of the most influential architectures for biomedical image segmentation is the U-Net, proposed by Ronneberger et al. [1]. The U-Net architecture features a symmetric encoder-decoder structure with skip connections, which allows for the preservation of spatial information while capturing hierarchical features. This architecture has been widely adopted for various medical image segmentation tasks, including lung segmentation, due to its ability to perform well even with limited training data.

Further improvements to CNN-based models include Fully Convolutional Networks (FCNs), introduced by Long et al. [2], which replaced fully connected layers with convolutional layers to allow for end-to-end segmentation of images at different scales. FCNs have also been successfully applied to medical image segmentation tasks, demonstrating robust performance on lung segmentation datasets.

The use of residual networks (ResNets) has also gained traction in recent years. He et al. [3] introduced deep residual learning for image recognition, a technique that mitigates the vanishing gradient problem and enables the training of deeper models. These models have been employed in medical image segmentation to further enhance feature extraction capabilities, especially in complex datasets such as lung CT scans.

Kamnitsas et al. [4] explored the effectiveness of ensemble learning in brain tumor segmentation, demonstrating that combining multiple models can improve robustness and accuracy. Their approach involved training multiple networks with different architectures and aggregating their predictions to mitigate individual model biases. This technique has also been extended to lung segmentation, where different segmentation models can complement each other by capturing different aspects of CT images. Ensemble-based approaches have been particularly useful in handling variations in imaging conditions and patient anatomy, which are common challenges in lung segmentation tasks. However, ensemble learning comes with increased computational costs due to the need to train and deploy multiple models simultaneously.

In the realm of 3D medical image segmentation, Shaker et al. [5] introduced the UNETR++ architecture, which combines the efficiency of the U-Net model with transformer-based mechanisms for improved segmentation in 3D datasets. This approach has shown promising results for lung CT segmentation, particularly in terms of computational efficiency and accuracy.

Furthermore, challenges such as class imbalance, the presence of noise, and variations in lung anatomy remain significant hurdles. The Lung CT Segmentation Challenge [6] has played a critical role in advancing methods to handle

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

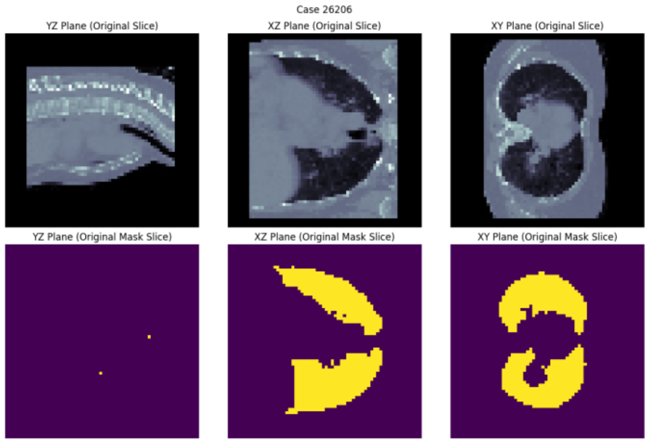
Case 26206
YZ Plane (Original Slice)
XZ Plane (Original Slice)
XY Plane (Original Slice)
YZ Plane (Original Mask Slice)
XZ Plane (Original Mask Slice)
XY Plane (Original Mask Slice)

Fig. 1. Slices of Original Images and Masks

these issues, providing a benchmark for evaluating lung segmentation algorithms on diverse datasets.

III. DATASET AND PREPROCESSING

Medical imaging datasets for lung segmentation vary due to patient anatomy, imaging protocols, and scanner noise, making preprocessing essential for consistency and robustness. 3D images exhibit intensity variations from different scanner settings and conditions, requiring HU windowing and standardization for uniform feature extraction. Gaussian smoothing helps suppress noise and artifacts, improving segmentation accuracy. Adjusting intensity ranges and normalizing pixel values enhances anatomical structures, allowing the model to focus on lung regions while reducing irrelevant details. The dataset consists of 3D (64×64×64) volumetric medical images in DICOM and NIfTI formats, which require structured preprocessing for optimal segmentation results.

- HU Windowing: Intensity values were clipped to the range [-1000, 400] HU, focusing on lung tissues while removing irrelevant structures such as bones and artifacts. This reduced the complexity of the images and improved model generalization.
- Gaussian Smoothing: A low-pass Gaussian filter was applied to remove high-frequency noise and smooth intensity variations. This preprocessing step enhanced robustness by suppressing scanner artifacts and producing cleaner segmentation boundaries.
- Standardization: Each image was standardized by subtracting the mean intensity and dividing by the standard deviation. This ensured uniform feature distribution across the dataset, reducing sensitivity to variations in intensity levels and improving training stability.
- Normalization: Scaling pixel intensities to ensure consistency.

Given that images were already standardized in dimensions, no further resizing was necessary. Test and validation datasets underwent only preprocessing to maintain integrity.

“Fig. 1” shows slices across axial, sagittal, and coronal planes for Case 26206 after preprocessing steps applied.

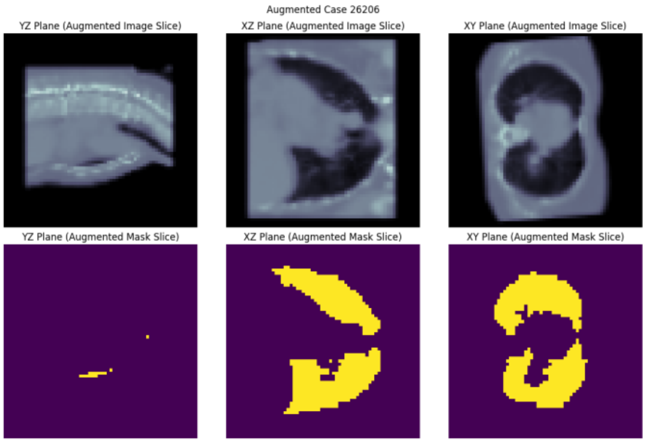
Augmented Case 26206
YZ Plane (Augmented Image Slice)
XZ Plane (Augmented Image Slice)
XY Plane (Augmented Image Slice)
YZ Plane (Augmented Mask Slice)
XZ Plane (Augmented Mask Slice)
XY Plane (Augmented Mask Slice)

Fig. 1. Augmented Image and Mask Slices

IV. DATA AUGMENTATION

To mitigate overfitting and improve the generalization of the model, dynamic data augmentation techniques were implemented directly into the TensorFlow dataset pipeline.

The following techniques were utilized:

- Elastic Deformation: This method introduces localized distortions to the lung regions in the images, simulating anatomical variability. The purpose of this augmentation is to improve the model’s ability to generalize across diverse anatomical variations, which is crucial for accurate segmentation in real-world clinical settings.
- Random Rotations and Flips: These transformations are applied to improve the spatial invariance of the model. By randomly rotating and flipping the lung regions and their corresponding masks, the model becomes more robust to variations in image orientation and symmetry, ensuring accurate segmentation regardless of the position or rotation of the lung structures.
- Random Scaling: Scaling transformations were used to accommodate varying lung sizes and imaging resolutions. By randomly altering the size of the lung structures within the images, the model is trained to generalize across images with different lung sizes, improving segmentation accuracy for a range of patient data and scanning resolutions.
- Intensity Variations & Noise Addition: Random intensity shifts and the introduction of noise are applied to simulate variations in imaging conditions, such as differences in scanner settings and patient positioning. These augmentations help the model become more resilient to imaging artifacts, ensuring accurate segmentation in noisy or imperfect datasets.
- Brightness Change: Random brightness adjustments simulate variations in the lighting and scanning conditions across different imaging devices. This augmentation ensures that the model can handle datasets with differing intensity distributions, further enhancing its generalization capabilities.

“Fig. 2” shows slices across axial, sagittal, and coronal planes for Case 26206 after all augmentations applied to demonstrate the effects of augmentation techniques.

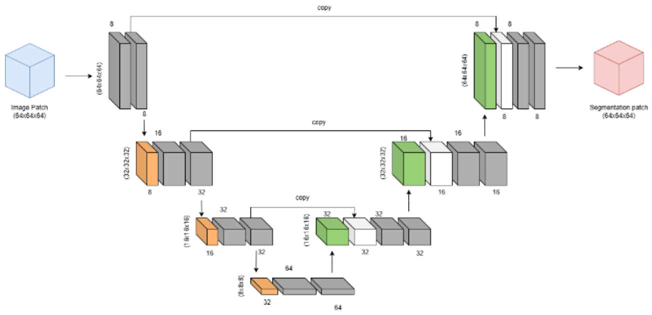

Fig. 3. Model Architecture Visualization

Each augmentation was applied with a probability of greater than 0.5 to ensure that the model encountered a sufficient diversity of training data during each epoch. Careful tuning of the augmentation intensity was essential to prevent excessive distortions that could negatively affect performance. By dynamically applying these transformations, the model's generalization ability was significantly enhanced, improving its ability to perform well on unseen data.

V. Model Architecture

A. U-Net Architecture

In this study, a 3D U-Net architecture is employed for the segmentation of lung structures from volumetric CT images. The U-Net model is a sophisticated convolutional neural network (CNN) specifically designed for biomedical image segmentation. It features an encoder-decoder structure augmented with skip connections, which facilitate the retention of spatial information during feature extraction.

B. Architectural Details

The U-Net model is composed of two main paths and fundamental components:

1) Encoder (Contracting Path): The encoder is responsible for extracting hierarchical features from the input CT volumes. It consists of multiple 3D (3x3x3) convolutional layers, each followed by batch normalization and ReLU activation. Max pooling layers are employed to progressively reduce the spatial resolution while simultaneously increasing the depth of the feature maps.

2) Decoder (Expanding Path): The decoder reconstructs the segmentation map by sequentially upsampling the feature maps. This process involves 3D transpose convolution layers, concatenation with corresponding skip connections from the encoder, and final convolution layers that generate the segmentation output.

3) Network Layers

In the proposed approach, the U-Net architecture relies on 3D convolutional layers to extract spatial features from the volumetric CT scans. By capturing information across all three dimensions, these layers enable the model to learn rich contextual details essential for accurate segmentation.

Batch normalization follows the convolutions, standardizing feature distributions at each layer to mitigate internal covariate shifts. This stabilization of feature statistics improves the training process, allowing for higher learning rates and better convergence.

To introduce non-linearity and enhance representational power, *ReLU* activation is applied after each convolutional operation. This activation function zeroes out negative values while preserving positive ones, enabling the network to learn complex decision boundaries.

Within the encoder, max pooling layers perform downsampling by reducing the spatial dimensions of feature maps. This operation shrinks the size of the representations, focusing the network on the most important features while simultaneously reducing computational requirements.

In the decoder path, upsampling and transpose convolution layers are used to recover spatial resolution. By performing the inverse of downsampling, these layers reconstruct finer details of the segmented regions, essential for precise boundary delineation.

A key innovation in U-Net is the inclusion of skip connections, which bridge the encoder and decoder at corresponding scales. These connections transfer low-level spatial features directly to the decoder, preventing the loss of fine details that often occurs during pooling operations.

Finally, a sigmoid layer at the output produces voxel-wise segmentation probabilities. This element-wise activation ensures that each voxel is assigned a probability value, enabling fine-grained and interpretable segmentation results.

In “Fig. 3” Gray blocks correspond to convolutional layers. Orange blocks represent max pooling layers, which reduce dimensions in the downsampling path. The green blocks denote upsampling layers, which restore spatial resolution in the upsampling. The skip connections (copies) help transfer high-resolution features from the downsampling path directly to the corresponding upsampling layers, preserving spatial details. This residual-style information transfer is a key feature of U-Net, ensuring precise segmentation. The process starts from an input image patch (blue cube) and results in a segmentation patch (red cube).

4) Addressing Overfitting

To mitigate the tendency of deep models to overfit, dropout layers (specifically spatial dropout) are incorporated within both the encoder and decoder. By randomly deactivating neurons during training, dropout encourages the network to learn more robust, generalized representations rather than memorizing training examples.

In conjunction with dropout, batch normalization also plays a role in reducing overfitting. By normalizing activations and constraining their dynamic range, batch normalization stabilizes training and discourages the proliferation of excessively large weights.

Additionally, data augmentation techniques are employed to further improve generalization. Random rotations, flipping, scaling, and intensity normalization increase the effective size and diversity of the training dataset, making the model less prone to overfitting specific data patterns.

An early stopping criterion is utilized to monitor validation loss and halt training once improvements plateau. This prevents the model from continuing to adapt to the peculiarities of the training set, thereby preserving its ability to generalize to unseen data.

Finally, L2 regularization is applied to the convolutional layers to penalize large weight values. By constraining the magnitude of weights, this approach fosters simpler, more parsimonious models that are less susceptible to overfitting.

VI. Training

In the training pipeline, a custom data generator is utilized to apply dynamic augmentation to each mini-batch of

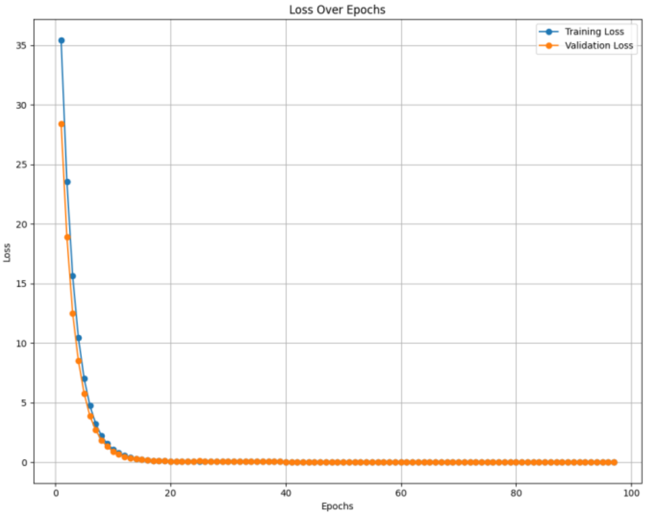

Fig. 4. Epoch Loss Graph

volumetric input data prior to feeding it into the neural network. This generator randomly selects from a set of transformations based on predefined probability thresholds and parameter ranges. Each transformation is designed to mimic real-world variations in patient positioning, scanner resolution, and tissue deformation, thereby expanding the effective diversity of the training dataset without the need for additional labeled samples. By introducing slight modifications in orientation, scale, and localized structure, the generator compels the model to learn features that are robust to subtle changes, ultimately improving generalization performance. Because these transformations are applied on-the-fly, no two training epochs use exactly the same version of the original scans, making it more difficult for the model to memorize specific augmented patterns. Simultaneously, a mini-batch gradient descent procedure is employed to update the model's parameters. Each forward pass processes a batch of augmented data, computes predictions, and compares them against the ground truth segmentation masks using a suitable loss function which is the combination Dice loss and Binary Cross-Entropy.

This hybrid approach helps address the issue of class imbalance while improving the model's segmentation accuracy. The gradients of the loss are then backpropagated to update the network weights, guided by Adam optimizer with a dynamically tuned learning rate initialized at 1e-3 to stabilize convergence. With using of *ReduceLROnPlateau* if the validation loss stops improving for a few consecutive epochs, the learning rate is reduced by a factor of 0.5 to encourage finer adjustments during training. Throughout the training epochs, performance is monitored on a validation set that remains untouched by augmentations. Early stopping is invoked when overfitting is detected, determined either by a stagnation or increase in the validation loss. A checkpointing mechanism is implemented to save the model whenever it achieves a lower validation loss, ensuring the best-performing model is retained. By halting the process at an optimal point, the network avoids learning patterns too specific to the training set. Once the training converges, the final model is primed for subsequent evaluation on an independent test set, where its robustness and accuracy can be conclusively verified.

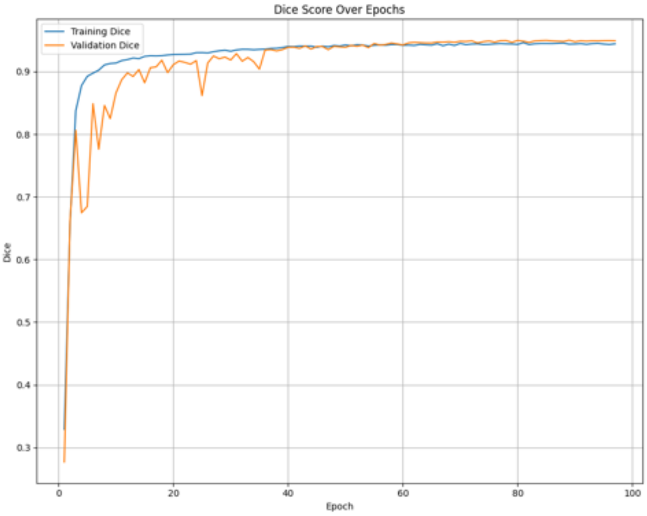

Fig. 5. Epoch Dice Graph

VII. Evaluation and Testing

In the testing phase, the best-performing model checkpoint, identified by the lowest validation loss during training, is restored to ensure optimal performance. This checkpoint-based approach minimizes the risk of performance degradation and maintains consistency between training and inference environments.

Inference is conducted in batch mode to leverage GPU parallelism and handle large-scale medical imaging data efficiently. The network generates probability maps, where each voxel is assigned a likelihood of belonging to the segmented region. These likelihoods are converted into binary masks using a fixed threshold of 0.5, thereby creating a clear distinction between segmented and non-segmented regions. Segmentation quality is then quantified using common evaluation metrics, such as the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and voxel-wise losses (for instance, binary cross-entropy).

$$\mathrm{DSC} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

$$\mathrm{IoU} = \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

In order to interpret these quantitative measures more effectively, qualitative visualizations are performed by overlaying the predicted segmentation masks on the original scans. Viewing multiple slices across axial, sagittal, and coronal planes helps reveal any systemic errors, subtle mismatches along structure boundaries, or inconsistencies between slices.

VIII. Experiment

A. Training

The experiments were conducted on a high-performance computing setup equipped with an NVIDIA Tesla P100 GPU. The model was implemented using TensorFlow/Keras, leveraging GPU acceleration for efficient training.
The model was trained using supervised learning with the following hyperparameters:

- Optimizer: Adam optimizer with β1 = 0.9, β2 = 0.999

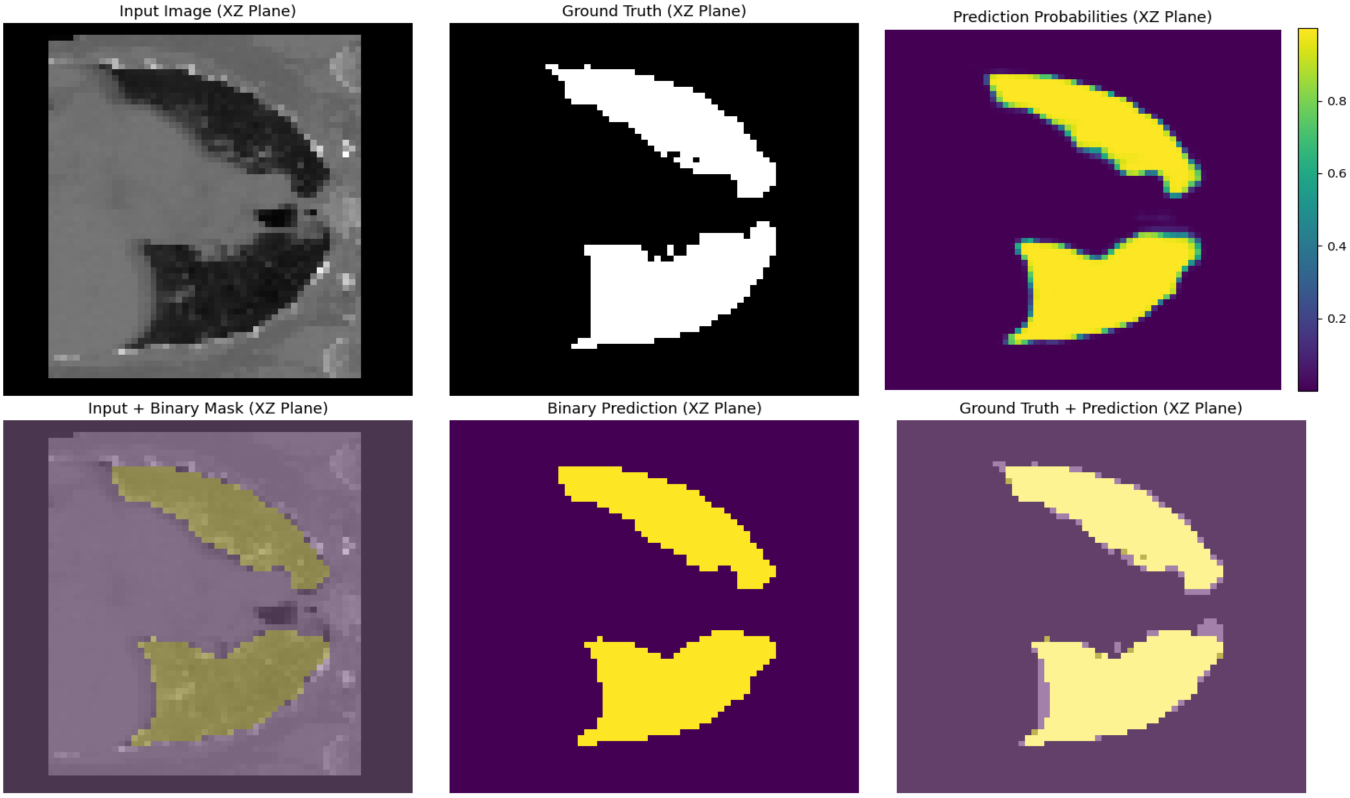
Input Image (XZ Plane)
Ground Truth (XZ Plane)
Prediction Probabilities (XZ Plane)
0.8
0.6
0.4
0.2
Input + Binary Mask (XZ Plane)
Binary Prediction (XZ Plane)
Ground Truth + Prediction (XZ Plane)

Fig. 6. Visualization of a Test Sample

- Learning Rate: 1e-3 (adaptive decay using ReduceLROnPlateau)
- Batch Size: 4
- Number of Epochs: 100
- Early Stopping: Enabled (patience = 8 epochs, monitoring validation loss). Early stopping triggered at 97th epoch.
- Checkpointing: Best model saved based on lowest validation loss which obtained at 89th epoch

"Fig. 4" illustrates the loss over epochs, where both training and validation loss rapidly decrease and stabilize overfitting close to zero, indicating effective learning and minimal overfitting. "Fig. 5" presents the Dice score progression, demonstrating a rapid increase in segmentation accuracy, exceeding 0.9 and maintaining stability across epochs. The close alignment between training and validation Dice scores suggests strong generalization, reinforcing the model's robustness for 3D lung segmentation tasks.

B. *Test Results*

For testing, the model was evaluated on unseen data and predictions were compared to ground-truth annotations. Performance metrics were computed for each test sample.

"Fig. 6" represents qualitative evaluation of the segmentation performance on test samples. The top row presents an input CT slice in the XZ plane, its corresponding ground truth segmentation mask, and the model's predicted probability map. The probability map shows the confidence of the model in segmenting the target structure, where higher confidence regions are closer to yellow, while uncertain areas remain in shades of green and blue. The bottom row provides additional visualization: the first column overlays the input CT slice with the binary mask to demonstrate the structural alignment, the second column shows the final binary prediction, and the last column overlays the ground truth and predicted segmentation for comparison. The qualitative results indicate that the model captures the primary regions of interest with high accuracy, as the binary prediction closely resembles the ground truth. However, minor boundary discrepancies are visible, likely due to slight uncertainties in the model's decision-making. The final overlay image confirms that most of the segmentation aligns well with the expected output, demonstrating the effectiveness of the proposed model in 3D medical image segmentation.

The Test Loss of 0.0440 indicates that the model maintains a minimal error rate when making predictions, suggesting well-learned feature representations and stable optimization. The Test Dice Score of 0.9432 confirms that the model achieves high segmentation accuracy, with a strong overlap between the predicted and ground truth masks. Additionally, the Test IoU of 0.8926 further supports the model's robustness by indicating a high degree of spatial agreement between the predicted and actual segmentation regions. These metrics collectively highlight the reliability of the model in precisely segmenting lung structures while minimizing false positives and false negatives.

IX. Conclusion and Discussion

This study presents a deep learning framework for lung segmentation, integrating robust preprocessing, dynamic augmentation, and an optimized U-Net model. The model demonstrates high segmentation accuracy, confirmed by Dice and IoU metrics. One of the key improvements in our approach was the incorporation of a more diverse dataset. By expanding the dataset to include varied imaging conditions, anatomical structures, and pathological cases, the model was able to generalize better across different medical scenarios.

For future work, attention mechanisms into the U-Net architecture can be integrated. Attention mechanisms enable the model to focus on the most relevant regions in an image, enhancing its ability to distinguish fine details and structures. Furthermore, increasing the depth of the U-Net by adding more layers and feature channels proved to be beneficial. Deeper architectures allow for richer feature extraction and representation, leading to improved segmentation accuracy. However, this increase in depth also comes with computational challenges, requiring careful optimization of model parameters to balance performance and efficiency.

Acknowledgment

The authors would like to express their sincere gratitude to Associate Professor Ian Postuma, for his valuable guidance and support throughout the project, which was conducted as part of a term project. Our code and model are available at: https://tinyurl.com/53btvjet

References

[1] R. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] K. Kamnitsas, Konstantinos, et al., "Ensembles of multiple models and architectures for robust brain tumour segmentation." International MICCAI Brainlesion Workshop, Springer, Cham, 2017.

[5] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "UNETR++: Delving Into Efficient and Accurate 3D Medical Image Segmentation," IEEE Transactions on Medical Imaging, vol. 43, no. 9, pp. 3377-3390, 2024.

[6] "Lung CT Segmentation Challenge 2017" [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017