SUP'COM

المدرســـة العليـــا للمواصــلات بتونــس
Ecole Supérieure des Communications de Tunis
Higher School of Communications of Tunis

Higher School of Communications of Tunis

Professional Preparation Project

Integration of AI Models into a Medical Diagnosis Application

Prepared by:
Asma Mhatli

Raed Ben Romdhane

Supervisor: Kater Nada Ayari

May 25, 2025

Contents

List of Figures

List of Tables

Introduction

Background

Medical image segmentation is a crucial step in numerous clinical applications, especially in the diagnosis and treatment planning of cancer. One of the significant challenges in lung cancer treatment is the accurate segmentation of tumors from 3D medical imaging data, such as CT scans. Manual segmentation by radiologists is time-consuming and subject to inter-observer variability. Consequently, automated segmentation methods using deep learning have gained attention for their ability to deliver consistent, high-quality results.
Recent developments in deep learning, particularly transformer-based architectures, have made it possible to capture both local and global features within volumetric medical images. UNETR (U-Net with Transformers) has emerged as an effective architecture for 3D medical image segmentation. This project focuses on utilizing UNETR for segmenting lung tumors in 3D CT scans through a region-wise approach, which enhances precision and interpretability.

Problem statement

The primary problem addressed in this project is the automatic segmentation of lung tumors in thoracic CT images. Traditional segmentation approaches using a single model for the entire image may overlook region-specific characteristics due to anatomical variability across the thoracic cavity. These limitations can reduce segmentation accuracy, particularly in challenging zones. Therefore, there is a need for a more structured and region-aware method that preserves both local detail and global coherence in the segmentation task.

Aims and objectives

Aims: To develop a desktop application for 3D lung tumor segmentation using the UNETR deep learning model, integrated with visualization tools for medical image analysis.
Objectives:

- To develop a desktop application that enables:
- Automatic inference using the trained UNETR models,
- Visualization of original CT scans and corresponding 3D tumor segmentations.
- To enhance usability through 2D and 3D views for both the input images and the generated segmentation results.

Chapter 1

State of the Art

Medical image segmentation has rapidly evolved with the emergence of deep learning, particularly CNN-based and more recently Transformer-based models. These advancements have laid the foundation for accurate, automated tumor delineation in volumetric data such as CT and MRI scans (1).

Key Industrial and Open-Source Platforms

- **Arterys:** A commercial platform leveraging 3D ResNet-based CNNs for real-time volumetric lung and cardiac analysis in the cloud. It showcases the integration of scalable AI in clinical radiology workflows.
- **3D Slicer:** A leading open-source platform that supports multiple segmentation architectures including 3D U-Net, V-Net, Attention U-Net, and nnU-Net. Through MONAI, it also enables integration with transformer-based models, facilitating state-of-the-art research experimentation (5).
- **InferVision:** Specializes in chest CT analysis using deep CNNs like V-Net and ResNet-enhanced 3D U-Nets, focused on lung nodule detection and segmentation tasks (5).

Rise of Transformer-Based Segmentation Models

While CNNs dominate current tools, they struggle with capturing long-range spatial dependencies, which are crucial in 3D anatomical structures. Transformers, originally from NLP, have demonstrated superior performance in vision tasks by modeling global context. In medical imaging, they have been successfully adapted to 3D segmentation :

Model	**Encoder Type**	**Strengths**	**Notable Feature**
UNETR	Vision Transformer (ViT)	Global context + spatial precision	Patch-based global transformer encoder
Swin UNETR	Swin Transformer	Local-global fusion, scalable	Shifted window attention
TransBTS	CNN + Transformer (bottleneck)	Efficient, balanced performance	Transformer in bottleneck only
MedT	Gated Axial Attention	Lightweight, anisotropic attention	Gated local-global context
AFTer-UNet	Axial + Fourier Attention	Orientation-aware, fine detail segmentation	Directional encoding + deep UNet decoder
DS-TransUNet	Dual Swin Transformer	High-resolution boundary segmentation	Dense connections + dual attention paths

Table 1.1: Comparison of Transformer-based Medical Image Segmentation Models

Chapter 2

Design and Architecture

The proposed pipeline enhances segmentation accuracy by leveraging anatomical knowledge of the thoracic region. The process begins by dividing the input dataset into three anatomically meaningful subregions:

Region-Based Segmentation Architecture Using UNETR

- Upper thoracic zone (e.g., neck to upper lungs)
- Middle thoracic zone (central lung area)
- Lower thoracic zone (base of the lungs to upper abdomen)

This anatomical division is based on the structural and visual similarities within each region, which helps localize tumors more effectively and simplifies the learning task for the model.
Each of the three subregions (part0, part1, part2) is independently processed using a dedicated UNETR model. Each model is trained specifically on its corresponding anatomical zone to produce a high-resolution 3D segmentation mask tailored to that region. After segmentation, the three regional masks are recombined (superposed) to reconstruct a coherent 3D global segmentation mask representing the full volume of the thorax. This aggregated mask allows:

- Tumor visualization in both 2D and 3D
- Interactive analysis by clinicians
- Support in diagnosis and decision-making.

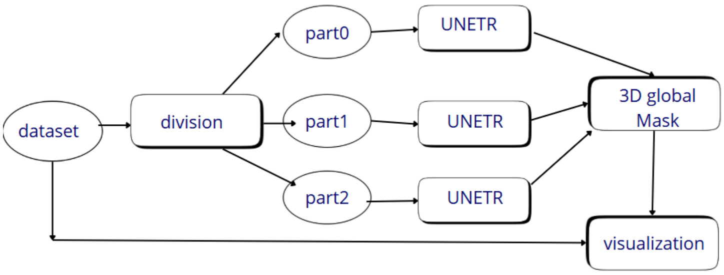

Figure 2.1:
Overall Workflow of the Proposed 3D Segmentation Pipeline

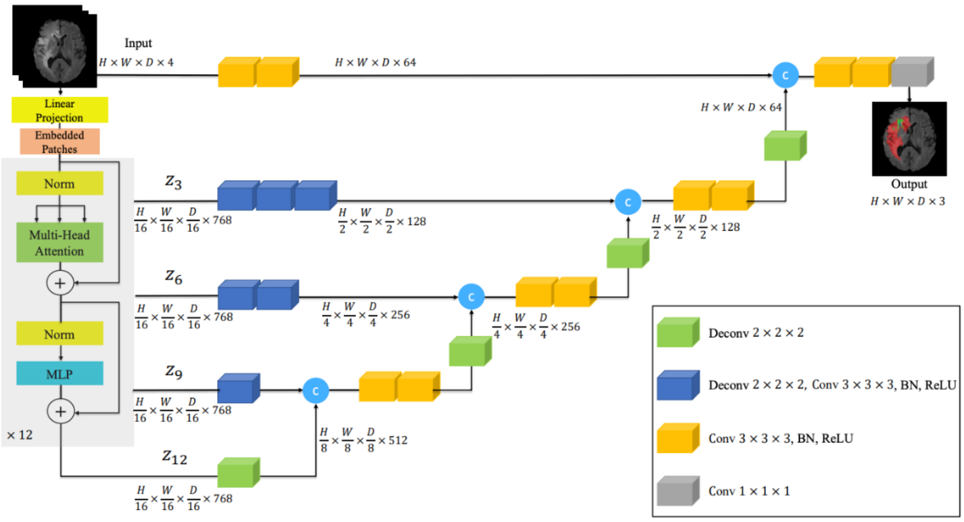
Input
$H \times W \times D \times 4$
$H \times W \times D \times 64$
c
$H \times W \times D \times 64$
Output
$H \times W \times D \times 3$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
Norm
MLP
× 12
z_3
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
c
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
z_6
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
c
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
z_9
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
c
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
🟩 Deconv 2 × 2 × 2
🟦 Deconv 2 × 2 × 2, Conv 3 × 3 × 3, BN, ReLU
🟨 Conv 3 × 3 × 3, BN, ReLU
⬜ Conv 1 × 1 × 1

Figure 2.2:
UNETR Architecture

The UNETR architecture leverages a powerful combination of a transformer-based encoder, skip connections, and a convolutional decoder to perform accurate 3D medical image segmentation. The encoder is built entirely using a Vision Transformer (ViT), where the input volume is first divided into non-overlapping 3D patches and linearly projected into an embedding space. These embedded patches pass through a series of 12 transformer blocks, each composed of multi-head self-attention, layer normalization, and a feed-forward MLP. Importantly, instead of waiting until the final transformer layer, UNETR extracts intermediate features from selected layers—specifically from the 3rd, 6th, 9th, and 12th blocks—creating skip connections that bridge the encoder and decoder. These skip connections carry rich multi-scale semantic information, which are reshaped and progressively integrated into the decoder pathway.(3)

Chapter 3

Environment and Implementation Phases

3.1 Choice and organization of the dataset

We started our work by searching for a suitable dataset. Our application is focused on medical image segmentation, so we decided to use a supervised learning approach. This type of learning requires segmentation masks made manually by radiologists. Because of that, we had to eliminate large datasets like the LIDC-IDRI dataset (which contains 1012 CT scans) in favor of smaller datasets with precise segmentation masks. We chose the *Lung Tumor Segmentation* dataset from Kaggle(4), which contains only 62 CT images. Although this is appropriate for a proof of concept, deploying the solution for real-world use will require a larger dataset.

3.2 Choice of the model

After reviewing several key papers — including *An Image is Worth 16x16 Words*(1), *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*(2), and *UNETR: Transformers for 3D Medical Image Segmentation*(3) , we decided to use UNETR for our project. UNETR fits our needs because it's designed for 3D medical images like CT scans and has proven effective for similar tasks. It is also available in the MONAI framework, making it easy to set up and use, and it offers many customizable settings and parameters that give us flexibility. Overall, UNETR gave us a strong and practical starting point for lung cancer image segmentation.

3.3 Training the UNETR model

Before starting the training process, we uploaded the dataset to Google Drive to minimize data transfer time when loading it into the training environment. We began training the model using the MONAI implementation on Google Colab for the first few epochs. Colab is a powerful tool with a user-friendly interface, which made debugging easier and more efficient. However, its major drawback is the limited GPU time in the free plan. Therefore, for the next step — intensive model training that takes several hours — we switched to Kaggle's platform. Kaggle provides better GPU access with 30 hours per week, compared to Colab's 1.5 hours per day (about 10.5 hours per week).

We continued monitoring the model's performance and adjusted the learning rate when needed, sometimes even going back a few epochs to retrain with new settings. This process was repeated until we reached the best performance possible using the given dataset.

3.4 Application development

The development phase was the simplest. Our application includes a graphical user interface where we integrated the model to ensure a smooth user experience. The only issue we faced was rendering the CT scans in 3D, as the images need some adjustments before being stacked to create a 3D view of the human thorax.

Chapter 4

Model Performance and Evaluation

During the training phase, we tracked the model's performance by logging key metrics. The results across the three parts of the dataset are illustrated in the following figures.

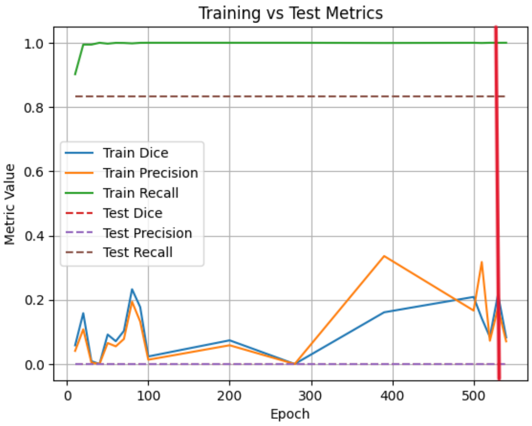

Figure 4.1: Performance metrics for Part0

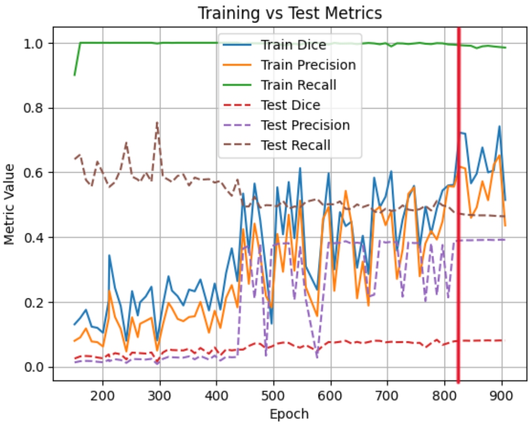

Figure 4.2: Performance metrics for Part1

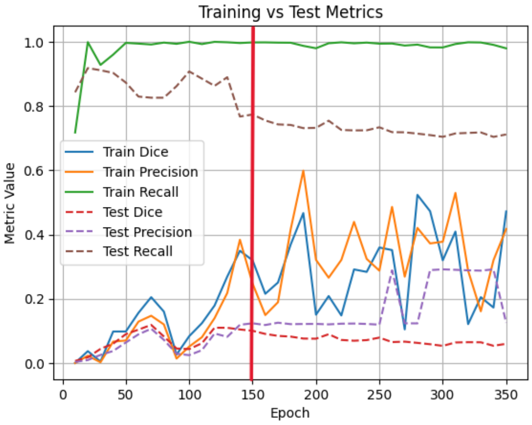

Figure 4.3: Performance metrics for Part2

(The vertical red lines indicate the selected optimal points.)

To select the optimal model parameters, we considered several evaluation metrics: **accuracy**, **recall**, and most importantly, the **Dice coefficient**, defined as:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$
- **Recall (Sensitivity)**: $\frac{TP}{TP+FN}$
- **Dice Coefficient**: $\frac{2TP}{2TP+FP+FN}$

Each metric reflects a different aspect of performance in medical image segmentation:

- **Accuracy** evaluates the overall correctness of predictions, but it can be misleading when the dataset is imbalanced (e.g., tumor regions are sparse).
- **Recall** measures the model's ability to detect all tumor voxels, which is critical in medical diagnostics to avoid missing suspicious regions.
- **Dice coefficient** quantifies the overlap between the predicted and ground truth tumor regions and is widely used in segmentation tasks.

We mainly focused on test results obtained from the **testing dataset**. However, for **Part 0**, which contains no tumors, we referred to results from the **training dataset**. In that case, the only meaningful conclusions are that the recall should be close to 1 (no tumors missed), and the precision should be close to 0 .

Overall, the model's performance peaked at a relatively low level. This is partly due to how the metrics are computed: each 3D scan is evaluated *slice by slice*, and the results are averaged over the full scan depth. If a slice (2D image) contains no tumor, the Dice score for that slice is 0, even if other slices are correctly segmented(No tumor detected). Since lung tumors rarely span the entire thoracic volume, the Dice coefficient and, similarly, the precision cannot reach 1, even with perfect predictions in tumor regions.

We also observed that the model shows relatively low accuracy due to a high number of false positives. It tends to segment slightly more area than the actual tumor and occasionally captures noise. This is an acceptable compromise to ensure **high recall**, meaning most tumor voxels are detected. In medical applications, especially for early-stage lung cancer detection, false positives are less critical than false negatives. It is preferable to detect too much rather than miss a tumor. Radiologists can verify the results, making this a safe approach for clinical use.

We should also highlight that the dataset size was limited: **31 530** grayscale `.npy` files (2D slices of size 256 × 256), grouped into **63** 3D scans. This small dataset size explains, in part, why the model's peak performance remains relatively low.

The following figure illustrates a sample result generated by the model, captured from the desktop application's interface (demonstrating how the segmentation output is visualized in 2D and 3D).

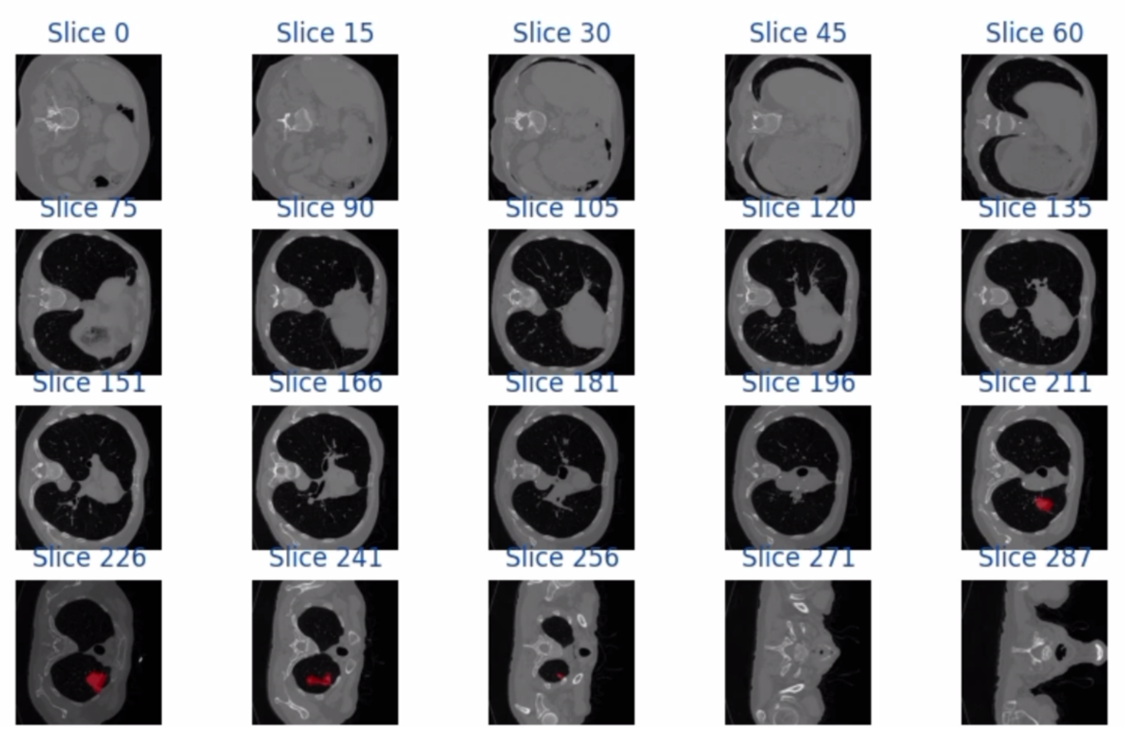

Figure 4.4:
Segmentation Output 2D

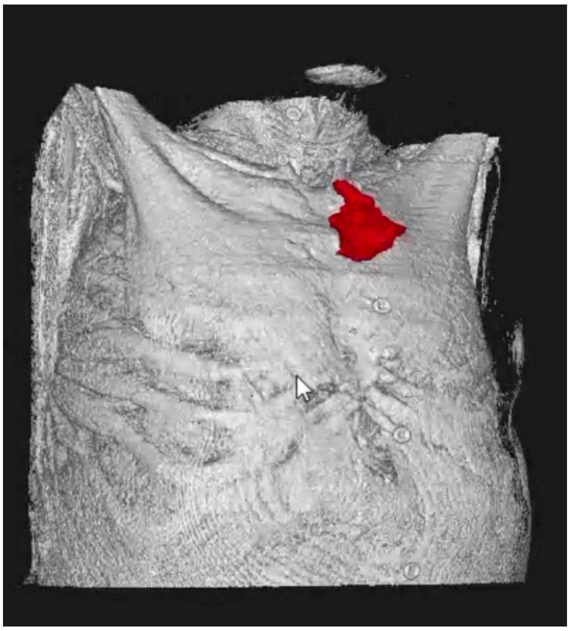

Figure 4.5:
Segmentation Output 3D

Conclusions

This project demonstrated the feasibility of using deep learning techniques, specifically the UNETR model, for 3D medical image segmentation of lung tumors. Despite working with a relatively small dataset, we successfully implemented a supervised learning pipeline that included data preprocessing, model training, and integration into a user-friendly application.

While our current implementation serves as a strong proof of concept, real-world deployment will require larger and more diverse datasets to enhance the model's generalizability. Future improvements could include advanced data augmentation, fine-tuning on more comprehensive datasets, and optimizing 3D visualization techniques for better clinical usability.

Overall, this work lays a solid foundation for further research and development in AI-assisted lung cancer diagnosis.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv preprint arXiv:2010.11929, 2020.

[2] J. Chen, Y. Lu, Q. Yu, T. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*, arXiv preprint arXiv:2102.04306, 2021.

[3] A. Hatamizadeh, H. Yin, J. Kautz, and P. Molchanov, *UNETR: Transformers for 3D Medical Image Segmentation*, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584, 2022.

[4] R. Saeid, *Lung Cancer Segmentation Dataset*, Kaggle, 2022. Available at: https://www.kaggle.com/datasets/rasoulisaeid/lung-cancer-segment

[5] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*, in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 424–432, 2016.

Appendix

Edited and Organized Dataset

The version of the dataset used in this project has been edited and organized for training purposes. It is available on Kaggle at the following link: https://www.kaggle.com/datasets/raedbenromdhane/lung-cancer-data-set

Model Training Code

The training code for the UNETR model used in this work is publicly available on Kaggle: https://www.kaggle.com/code/raedbenromdhane/unter-raed4