Electronic Letters on Computer Vision and Image Analysis 24(2):218-235, 2025

Head and Neck Organ-at-Risk Multi-Modal Semantic Segmentation using 3D-UNet

Takwa Ben Aïcha Gader*, Afef Kacem*, Aziz Ben Amor+, Hamdi Baccar+, Mohamed Nadhir Najjar+

* *University of Tunis, ENSIT - LaTICE, 5 Taha Hussein, Beb Menara, 1008 Tunis, Tunisia*
+ *University of Tunis, ENSIT, 5 Taha Hussein, Beb Menara, 1008 Tunis, Tunisia*

Received 11th of april, 2024; accepted 2nd of december, 2025

Abstract

Head and neck (HaN) cancer is still one of the most common cancers in the world, and radiotherapy is frequently used as the primary treatment. accurate identification of tumors and surrounding organs-at-risk (OARs) is crucial for radiotherapy planning accuracy because it maximizes tumor dose coverage while reducing radiation exposure to healthy tissues. The most common imaging method for treatment planning is computed tomography (CT); however, the limited soft-tissue contrast of CT frequently makes it difficult to accurately identify small or low-contrast OARs. In contrast, magnetic resonance imaging (MRI) offers better vision of soft tissues. In this research, we provide a fully automated multimodal segmentation system for the localization of 30 OARs in the HaN region, integrating co-registered CT and MRI data. Using an early-fusion method to take advantage of complementary anatomical information across modalities and skip connections for effective feature propagation, the model is based on a 3D U-Net architecture tailored for volumetric medical imaging. The model was trained and tested on the publicly accessible HaN-Seg dataset, which contains paired CT-MRI images and expert-annotated OARs masks.

The proposed approach demonstrated significant generalization across anatomically different cases, achieving competitive results with a mean F1 score of 0.778 ± 0.041 and overall accuracies of 94.2 % (training) and 94.1 % (testing). These findings confirm the efficiency of multimodal deep learning in enhancing OARs segmentation accuracy, as well as the system's potential to improve precision, reproducibility, and clinical reliability in head and neck radiotherapy planning.

Key Words:Head and Neck Cancer, Organ-at-Risk Identification, Multi-Modal Semantic Segmentation, Deep Learning, 3D-UNet, Computed Tomography, Magnetic Resonance, Radiation Therapy.

1 Introduction

HAN cancer is a diverse collection of cancers that develop in the upper aerodigestive tract, which includes the lips, oral cavity, pharynx, larynx, salivary glands, nasopharynx, paranasal sinuses, and nasal cavity. Fig. 1 illustrates the anatomical diversity and physical complexity of the principal regions affected by HaN tumors, posing considerable obstacles for radiation planning. The majority of HaN malignancies are squamous cell carcinomas, however additional histological subtypes exist, especially in less common locations such the salivary glands or nasal cavities. Globally, HaN malignancies account for around 4% of all cancer. The oral cavity is the most commonly affected subregion (41%), followed by the pharynx (22%), and the larynx (24%).

Correspondence to: takwa.benaichaa@isima.u-monastir.tn

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2147
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

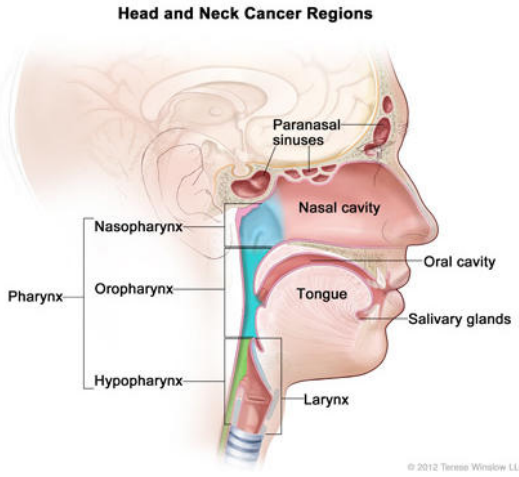


Figure 1: Head and neck cancer regions: labeled anatomical representation of typical tumor sites.

Radiotherapy is a primary treatment option for HaN malignancies. Its success is dependent on accurately delivering high radiation doses to tumor volumes while preserving neighbouring healthy tissues, notably OARs. This necessitates appropriate delineation of both the target volumes and the surrounding OARs, which is often accomplished with CT imaging. Certain OARs, however, are difficult to see on CT scans due to low soft-tissue contrast and are better diagnosed with magnetic resonance imaging (MRI). As a result, integrating CT and MRI modalities has the potential to improve OARs segmentation accuracy.

Despite this potential, multimodal segmentation techniques for HaN cancer are underexplored, in part due to a scarcity of standardized datasets that include both modalities. Most existing research focuses on single-modality segmentation, such as CT or MR. To address this restriction, we offer a fully automated multimodal segmentation approach that uses both CT and MRI scans to identify 30 anatomically unique OARs in the head and neck. The dataset includes coupled CT and MRI images for all participants, making it easier to create a fusion-based deep learning solution. Fig. 2 shows a 3D depiction of segmented OARs using the AAPM naming scheme, highlighting their spatial complexity and the significance of correct delineation for treatment planning.

Recent advances in radiotherapy, combined with simpler surgical techniques and multimodal treatments, have resulted in better patient outcomes. Deep learning is also recognized as an effective method for medical image analysis, notably in segmentation tasks. While CT remains important for dosage estimation because to its physical qualities, MRI provides higher soft-tissue contrast, making its combination useful in OARs delineation. However, few studies have thoroughly evaluated the effectiveness of combining both modes for this purpose.

In this work, we present a fully automated 3D-UNet-based multimodal segmentation system for head-and-neck organs at risk. To improve segmentation accuracy, the model is trained on combined CT and MRI images, taking advantage of their complementary spatial and contrast information. This study's main contribution is the use of a lightweight, reproducible 3D early-fusion architecture that has been customized for paired multimodal radiotherapy data. The approach emphasizes clinical effectiveness, computational efficiency, and reproducibility, laying the framework for future developments in multimodal fusion systems.

The remaining sections of the paper are organized as follows: Section 2 examines related research in the field. Section 3 outlines the suggested methodology, covering preprocessing, model design, and training strategies. Section 4 includes the results of experiments and comparisons to previous methods. Section 5 closes the work and provides recommendations for future research.

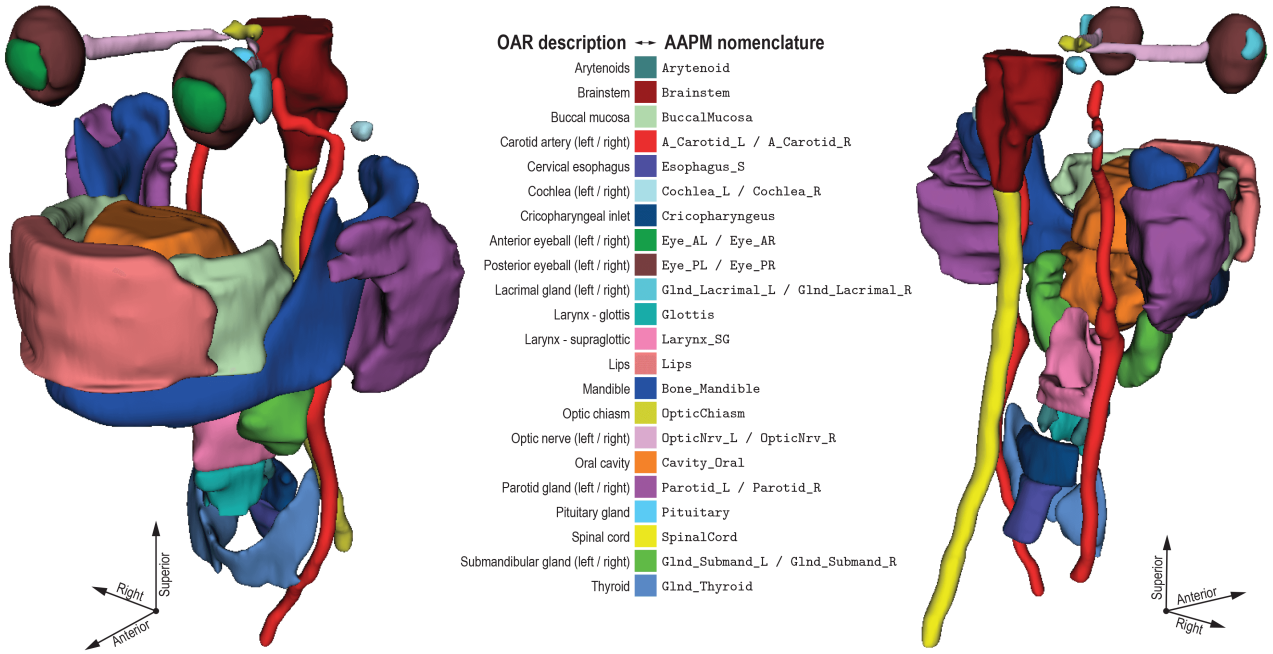


Figure 2: 3D visualization of segmented organs-at-risk (OARs) using CT and MRI modalities. Each structure is labeled and color-coded according to AAPM standards.

2 State of the Art

Medical image segmentation is still an actively explored field, with several techniques created to distinguish anatomical features in different imaging modalities. Even with the advancements, segmenting OARs in HaN radiotherapy is still extremely challenging, and current methods frequently don't meet clinical needs.

As part of the MICCAI 2015 Head and Neck Auto Segmentation Challenge [6] offered a public dataset to evaluate segmentation methods on CT scans. Six teams took part, using a variety of methods such as semi-automatic segmentation, multi-atlas-based segmentation, active appearance models, and statistical shape models. However, none of these techniques obtained results equal to manual delineations conducted by radiologists. The main challenges encountered in HaN OARs segmentation are large organ size variation that directly affects segmentation accuracy, high anatomical variability, low soft-tissue contrast in CT scans, and a limited capacity to encode prior anatomical knowledge into standard shape models.

While CT imaging offers a high contrast between soft tissues and bone, traditional segmentation techniques like region growth, edge detection, and thresholding are limited in their efficiency due to the inherent complexity of the head and neck anatomy. Moreover, the vast number of OARs to be segmented, their varied morphology, and the frequent overlap or proximity between structures make these procedures even more difficult. The obtained segmentation quality was still too low for clinical use, even with the use of sophisticated model-based techniques like 3D level sets [7] and atlas-based methods [8].

In recent years, medical image segmentation, including OARs segmentation from HaN CT scans, has significantly improved due to deep learning, namely, convolutional neural networks (CNNs) [9], [10], [11], [13]. For this goal, a variety of 2D and 3D CNN-based frameworks have been proposed; nevertheless, a significant constraint is the inability to accurately identify small and low-contrast OARs.

AnatomyNet [11], one of the most influential architectures, presented a fully automated 3D residual U-Net made especially for CT segmentation of the head and neck. To improve feature recalibration and attain state-of-the-art performance on large organs, the network integrates squeeze-and-excitation blocks. Its dependence on CT-only data, however, limits its capacity to precisely segment tiny OARs and soft-tissue borders. Furthermore,

the lack of contextual information in single-modality inputs and class imbalance impact its performance.

A variety of multimodal techniques have been presented to take advantage of complementary information from several imaging sources, including CT and MRI, to overcome the drawbacks of single-modality segmentation. Zhang et al. [14] identified three types of deep learning-based fusion approaches: early, late, and hybrid (layer-wise). Early fusion allows the network to learn from all modalities simultaneously from the start of training by combining multimodal inputs at the pixel level. At the decision stage, late fusion, in contrast, combines the feature maps or segmentation predictions produced by processing each modality separately.

Fig. 3 presents a typical late fusion design in which axial, coronal, and sagittal slices are separated into distinct branches of a shared-weight U-Net. An ensemble procedure is then used to aggregate the separate 2D segmentation outputs back into 3D space. This structure creates a consistent global prediction through ensemble integration, but modality- or view-specific information is preserved during independent processing. By combining data representations at various model depths, hybrid fusion designs aim to improve generalization and create deeper cross-modal interactions.

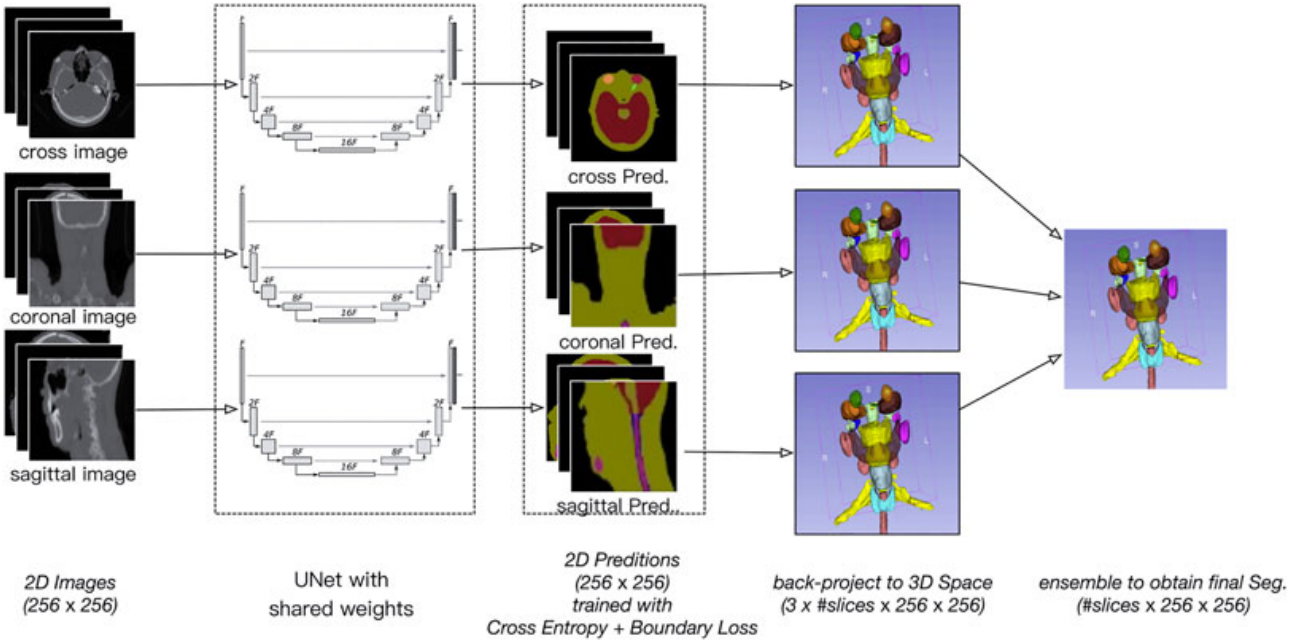


Figure 3: Example of a late fusion multimodal segmentation architecture.

Another major challenge in multimodal segmentation is the missing modality scenario, which occurs when only partial data is available for a specific patient. Even in the absence of one modality, robust models ought to retain a constant level of segmentation quality. Therefore, the question of which fusion approach is the most effective remains open.

Dolz et al. [15] introduced HyperDenseNet, another approach for multimodal fusion. By establishing hyper-dense connection points across feature maps, this 3D CNN design links each layer both within and across modality-specific encoders. The network can exchange data at various depth levels thanks to this connectivity, but it comes with a significant architectural complexity cost. The resulting model's accuracy and clinical application are limited by its high computing resources and memory requirements.

These studies collectively demonstrate a trade-off between architectural intricacy and clinical usability. In clinical workflows, complex multimodal designs like HyperDenseNet frequently lack efficiency and reproducibility despite achieving excellent accuracy. On the other hand, despite their speed and scalability, single-modality frameworks such as AnatomyNet are unable to effectively utilize complementary multimodal information. For real-world radiation planning, this constant trade-off motivates research toward systems that preserve high performance while still being computationally effective and repeatable.

In this work, we propose a multimodal segmentation framework for HaN OARs that uses a unified early-fusion 3D U-Net architecture to directly combine CT and MRI data from the same patient. The network may learn combined spatial and contrast representations from the very first encoding layers according to the proposed model, which combines both modalities at the input level. Our design prioritizes computational economy and consistency while preserving competitive segmentation performance, in contrast to complex architectures like HyperDenseNet. This early-fusion approach is better suited for practical clinical implementation since it minimizes memory consumption and eliminates architectural redundancy.

3 Proposed System

Our system operates by performing early modality fusion, where data from computed tomography (CT) and magnetic resonance imaging (MRI) are merged before inputting them into the segmentation network for the head and neck region. The use of early fusion typically involves combining the complementary information provided by CT and MRI modalities to improve the accuracy and robustness of the segmentation process. CT scans excel in visualizing bones and dense tissues, whereas MRI offers superior contrast for soft tissues.

For segmentation, we employed the U-Net architecture, a widely adopted choice for medical image segmentation tasks due to its ability to effectively capture spatial information through a combination of downsampling and upsampling paths. When employing U-Net for head and neck segmentation, the network is trained on multi-modal data, where CT and MRI images are concatenated or merged at the input layer before being processed by the network. During training, the early fusion U-Net learns to exploit the complementary information from CT and MRI modalities to delineate the boundaries of various structures within the head and neck region more accurately. This integration of information helps the network to better distinguish between different tissues and organs, leading to more precise segmentation results.

3.1 System Overview

The proposed 3D-UNet-based system uses a three-dimensional version of the U-Net architecture, which consists of an encoder-decoder structure with skip connections. This architecture preserves spatial information while capturing contextual dependencies across the entire three-dimensional volume. The encoder extracts features from the input multi-modal imaging data, while the decoder generates segmentation maps at the output. Skip connections enable the integration of both low-level and high-level features, contributing to precise segmentation. The model is trained using a combination of CT and MRI images, leveraging their complementary information to improve segmentation performance. The 3D-UNet architecture comprises several key components:

- The encoder consists of multiple layers of 3D convolutional blocks followed by max-pooling layers. Each convolutional block typically includes 3D convolutional layers with batch normalization and activation functions (e.g., ReLU). These layers extract hierarchical features from the input multi-modal imaging data, gradually reducing spatial dimensions.

- The decoder is symmetrical to the encoder and consists of multiple layers of 3D upsampling blocks followed by 3D convolutional layers. Each upsampling block combines low-resolution feature maps from the encoder with high-resolution feature maps from skip connections. This fusion process helps recover spatial details lost during the downsampling stage.

- The skip connections, alternatively referred to as shortcut connections, which connect corresponding encoder and decoder layers to facilitate information flow across different scales. These connections allow the decoder to access both low-level and high-level features, aiding in precise segmentation by capturing fine details and contextual information simultaneously.

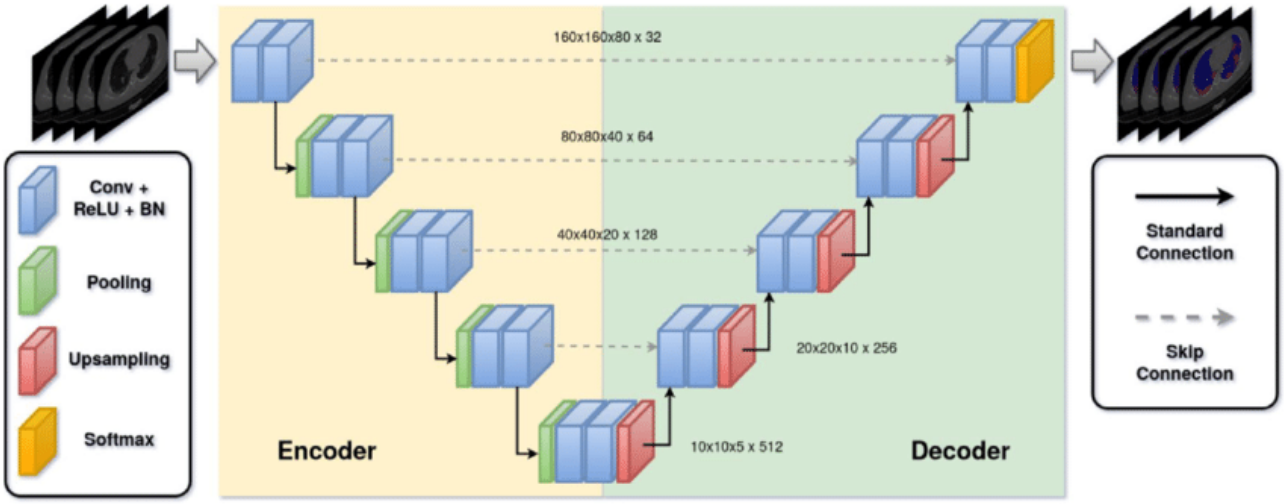


Figure 4: Proposed system general architecture.

– The final layer of the decoder, which typically consists of a 3D convolutional layer with a softmax activation function, produces segmentation probability maps for each class of interest. These probability maps represent the likelihood of each voxel belonging to the target structures, such as organs at risk in the head and neck region.

– The loss function: During training, the model is optimized using an appropriate loss function, such as cross-entropy loss or Dice coefficient loss, to minimize the discrepancy between the predicted segmentation maps and ground truth annotations.

As shown in Fig. 4, the model takes a 3D patch (cuboid) and outputs the segmentation of regions. Skip connections were implemented with concatenation layers. Conv: Convolutional layer; ReLU: Rectified linear unit layer; BN: Batch normalization.

3.2 Rationale for Early Fusion Strategy

The choice of an early-fusion strategy is motivated by the nature of the available data and the intended clinical applicability of the model. Since the HaN-Seg dataset's CT and MRI volumes are already spatially co-registered, the network may jointly use useful information from both modalities from the very first encoding layers by integrating them directly at the input level. While MRI improves soft-tissue contrast and CT offers precise structural and bone delineation, the model can simultaneously learn correlated spatial and intensity patterns.

Early fusion eliminates the need for complicated cross-connection methods or modality-specific encoders, which simplifies the overall design from a structural point of view. During training, this reduction in architectural redundancy significantly decreases memory use and computational cost. Furthermore, the model improves generalization on small and low-contrast OARs, which frequently have restricted boundary visibility and data imbalance, by learning shared low-level representations.

In practice, this design decision achieves a compromise between computational efficiency, interpretability, and performance, which makes it ideal for clinical applications where quick and consistent segmentation is crucial.

4 Experiments and segmentation Results

4.1 Used Dataset

The proposed system was assessed on the publicly available HaN-Seg dataset [16], which was recently released. This database comprises CT and T1-weighted MRI images from 56 patients who underwent image-guided radiotherapy. For each patient, reference segmentation of up to 30 organs at risk (OARs) was obtained by experts who performed pixel-by-pixel image annotations manually. Maintaining the distribution of patient ages and genders, as well as the type of annotation, the patients were randomly divided into a training set (Set 1, 42 cases, 75%) and a test set (Set 2, 14 cases, 25%).

4.2 Data Annotation

Clinical professionals from Slovenia's Institute of Oncology manually annotated segmentation masks for OARs. These annotations were applied to CT scans using pixel-level separation. An experienced radiotherapy technologist annotated about half of the cases, while a panel of qualified oncologists annotated the rest. In certain cases, co-registered MRI scans were used to guide and refine delineations, particularly for soft-tissue features that are less evident in CT alone.

To ensure labeling uniformity and anatomical accuracy, all segments were inspected and manually corrected. The final dataset removed regions considered anatomically unclear or unreliable, such as the pharyngeal constrictor muscles. The ground-truth masks contain up to 30 OARs per patient and conform to the standardized nomenclature specified by the American Association of Physicists in Medicine (AAPM), as depicted in Fig. 2.

Following the annotation and registration stages, all binary masks were combined into a single multi-class label volume per case. This unified representation speeds up the learning process by allowing the network to assess all regions of interest in a single, spatially coherent volume. Fig. 5 shows the final stacked label volume across three standard anatomical planes (sagittal, coronal, and axial).

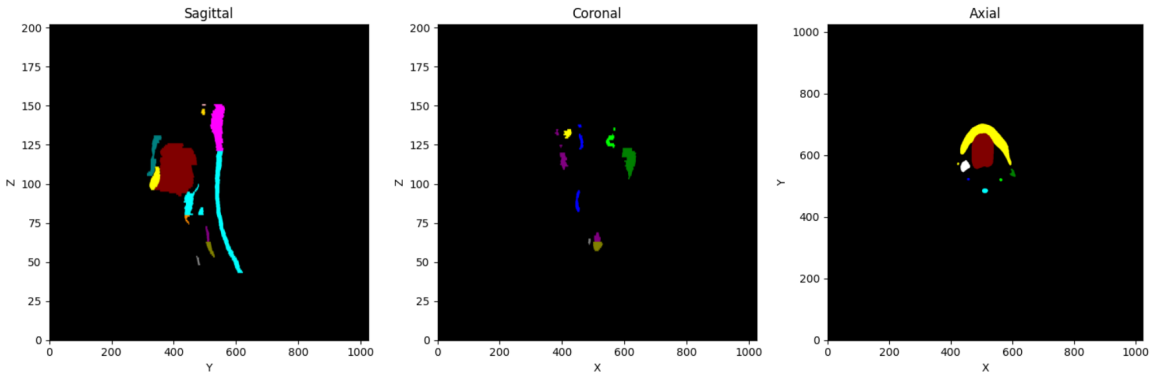


Figure 5: Overlay of binary masks stacked in a multi-class volume

4.3 Data Understanding

We used 3D Slicer as the primary tool to visualize and pre-process our dataset, providing valuable insights into data characteristics and facilitating experimental preparations. We will now proceed with a detailed analysis of our dataset using case number 1 as an example. Figures 6 and 7 respectively display the MRI and CT modalities for all three orientations: sagittal, coronal, and axial. The 3D CT modality images have overall dimensions of (width, height, depth) = (1024, 1024, 202), while the MRI modality images have dimensions of (width, height, depth) = (512, 512, 83).

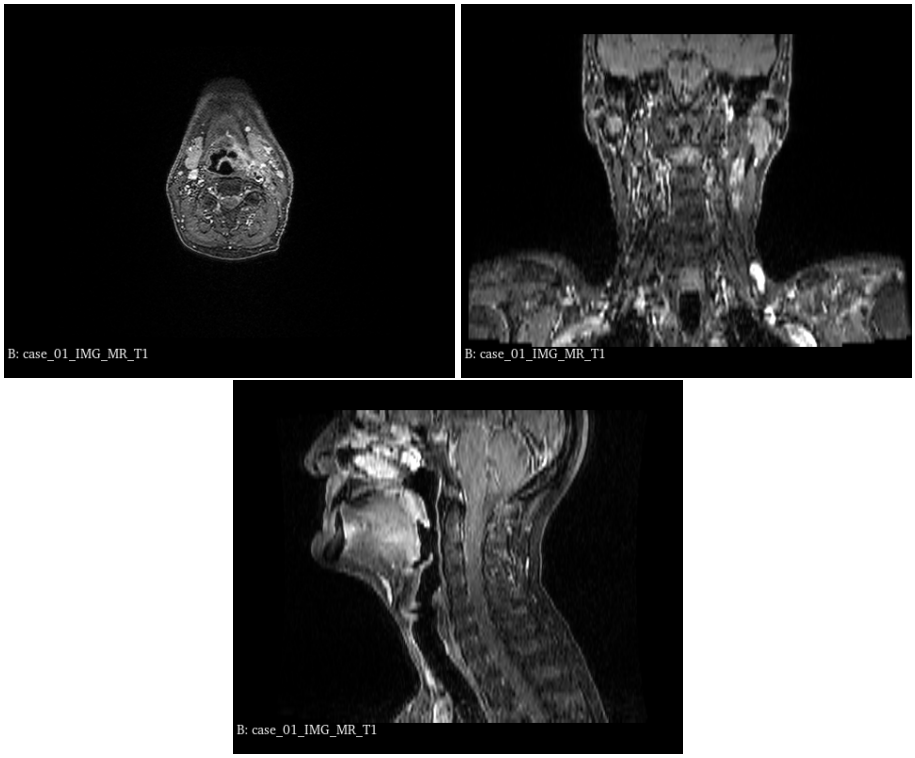


Figure 6: Case 1 MRI image.



Figure 7: Case 1 CT image.

We give visual representations of particular masks for the first case. As illustrated in Fig. 8, these masks will be presented in three different orientations, namely sagittal, coronal, and axial. These masks have been created to accurately depict the ground-truth segmentation of various OARs within the head and neck region.

We then combined the 30 masks into one mask for the first case. This is a crucial step as it allows for the simultaneous visualization of segmentations for all OARs in the three orientations (sagittal, coronal, and axial). By merging masks from all classes into one image, each class is assigned a specific color, which helps understand and interpret the results. This approach offers a clear overview of the segmented anatomical structures in both CT and MRI images, which is essential for assessing the segmentation quality and understanding the spatial relationships between different OARs in the head and neck region. As shown in Fig. 2, a color map is dedicated to the HaN OARs for better distinction and clarity.

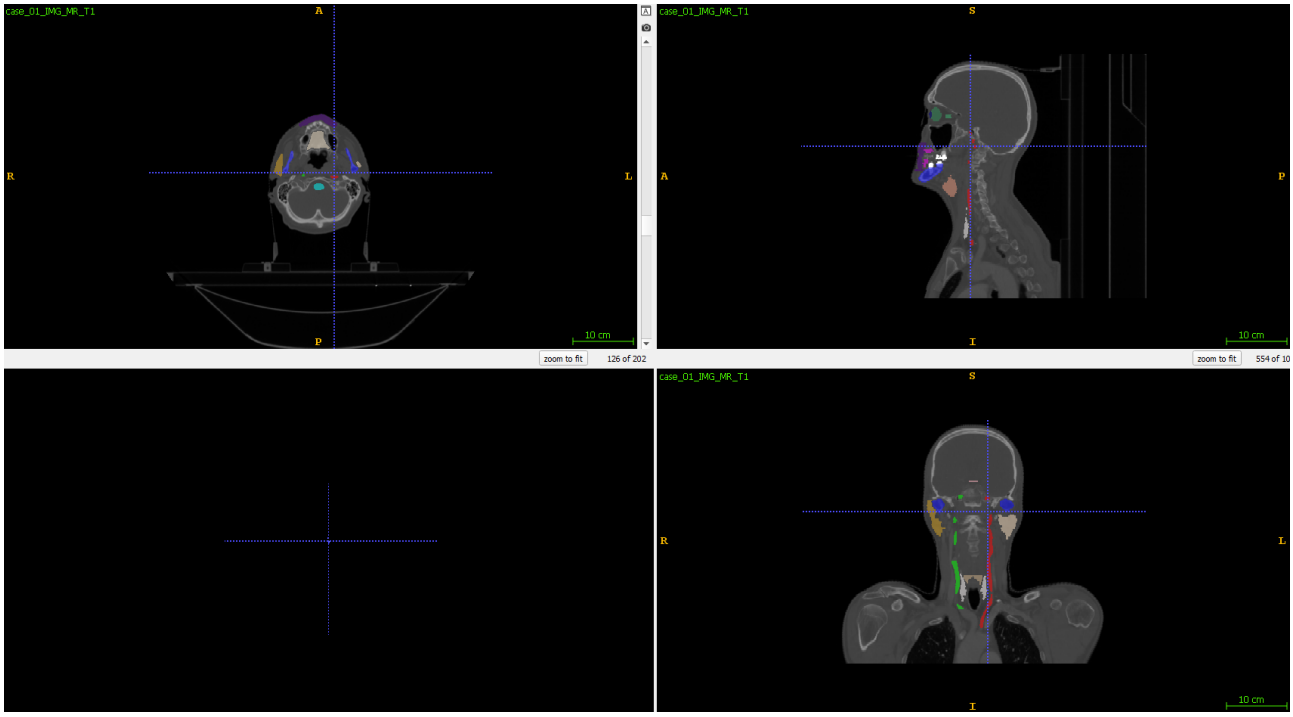


Figure 8: Visualization of case 1 masks.

4.4 Data Preprocessing

During initial data exploration, we faced several challenges that required specific preprocessing steps to ensure data integrity and improve model performance.

- Missing Masks/Classes: We identified and excluded cases with missing masks or classes. This step is crucial to prevent training or testing on incomplete data, which could result in segmentation errors.
- Modalities-Related Issues:
  - Uncentered Masks: Misalignment between CT and MRI volumes was seen in numerous cases, with anatomical features appearing displaced or misaligned across modalities. Fig. 9 demonstrates that the MRI and segmentation masks are not spatially centred with respect to the CT image, resulting in inconsistent overlap between modalities. If left untreated, this spatial disparity can have a major influence on multimodal learning and segmentation performance. To overcome this issue, we performed a centring method that aligned the volumetric centres of both modalities before fusion.
  - Image Intensity Disparities: While CT image intensities were acceptable, MRI Images had very high intensities that needed standardization.
  - Image Size Discrepancy: Images varied in size between $x \times 512 \times 512$ and $x \times 1024 \times 1024$.
- Class Imbalances: The dataset shows a significant difference between large, dominant anatomical structures and smaller, underrepresented organs. Fig. 10 shows that background voxels make up the majority of the image, followed by significant components like the carotid arteries and spinal cord. Smaller organs, such as the submandibular glands, brainstem, and oral cavity, contribute just a small fraction of total voxel volume.



Figure 9: Uncentered Modalities.

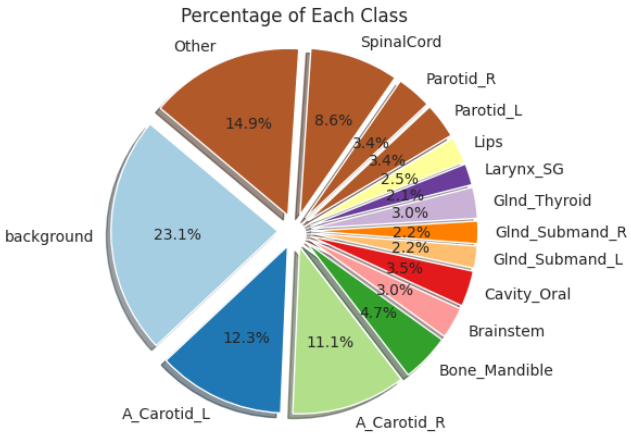


Figure 10: Class distribution across organs and background. The imbalance is evident, with background and large vessels dominating the voxel count, while small organs are underrepresented.

4.4.1 Image Standardization

This is done to guarantee that the input data is as high-quality and uniform as feasible. This includes several tests and adjustments to guarantee that images from various modalities are compatible and ready for analysis. For example, size variations between CT and MRI scans may induce biases into the model.

To deal with this, we carefully cropped the CT images, making sure not to remove areas with critical information. A bounding-box-based cropping method was used to isolate the relevant anatomical region. The minimum and maximum coordinates along each axis were determined for each 3D volume using non-zero voxels from the ground-truth segmentation masks. These coordinates defined a tight spatial bounding box that was applied evenly to both the CT images and their respective masks:

```
cropped_images = large_images[min_x:max_x+1, min_y:max_y+1, min_z:max_z+1]
```

```
cropped_masks  = large_masks[min_x:max_x+1, min_y:max_y+1, min_z:max_z+1]
```

To adjust intensity differences between CT and MRI modalities, MRI images were adjusted using z-score normalization. For each MRI volume, we subtracted the mean and divided by the standard deviation of non-zero voxels:

$I' = \frac{I - \mu}{\sigma}$ where μ is the mean intensity of non-zero voxels, and σ is the corresponding standard deviation.

This standardization assures zero-mean, unit-variance distributions, helping to balance contributions from both modalities and enhancing training stability. It also keeps the network from being influenced by the greater intensity variation commonly found in MRI images.

This strategy retains only the regions containing organs at risk (OARs), reducing unnecessary background and increasing computational performance. It also assures that the spatial dimensions of the dataset are consistent, which is required for multimodal data comparability and fusion. All volumes were then scaled to fit the default spatial resolution of $x, 512, 512$ in both modalities.

4.4.2 Image Registration

To achieve anatomical alignment between modalities, MRI images were resampled to match the voxel spacing, size, and orientation of the matching CT volumes. This procedure involved integrating metadata parameters like origin, spacing, and direction cosine matrices. Specifically, MRI images were resampled to match the voxel spacing, size, and orientation of their corresponding CT volumes. This includes ensuring consistent metadata parameters like *origin*, *spacing*, and *direction cosine matrices*. Furthermore, the centres of each MRI volume were aligned with their CT equivalents by computing the bounding box of non-zero regions in the mask and adjusting both modalities correspondingly. This technique provided an approximate anatomical correlation suitable for early fusion, especially given the very slight misalignment reported in the HaN-Seg dataset, which was obtained using a standardized radiotherapy protocol.

The alignment method was utilized consistently on both the input images and the ground-truth segmentation masks. All important structures were retained after resampling to prevent anatomical loss. After alignment, all binary segmentation masks were combined into a single composite multi-class label volume for each example. This simplification improved training efficiency by allowing the model to learn from a single anatomical representation. As shown in Fig. 11, which displays the overlay of resampled and spatially aligned CT and MRI modalities for three anatomical planes. The three perspectives (sagittal, coronal, and axial) show good alignment of anatomical structures across modalities, allowing voxel-level matching. This phase was critical to ensuring that multimodal fusion could be performed consistently, eliminating mismatches between tissue boundaries and reducing modality-induced artefacts during learning.

Furthermore, MRI image intensities were adjusted to match the range and distribution of CT data. This normalization phase was important to avoid modality-specific biases during model training and to ensure that both input channels contributed equally to the learned feature space.

4.4.3 Dataset Splitting

Dividing the dataset into a training set (70%), a test set (20%), and a validation set (10%) is a standard practice used to evaluate the model's generalization ability to new data. This approach assesses model effectiveness by training it on one data portion (training set) and testing it on a separate portion (test set), simulating its behavior with new observations.

4.4.4 Data Augmentation

During the training phase, data augmentation is applied dynamically while loading batches. This technique adds more diversity to the dataset without any additional resource consumption. Techniques such as rotation,

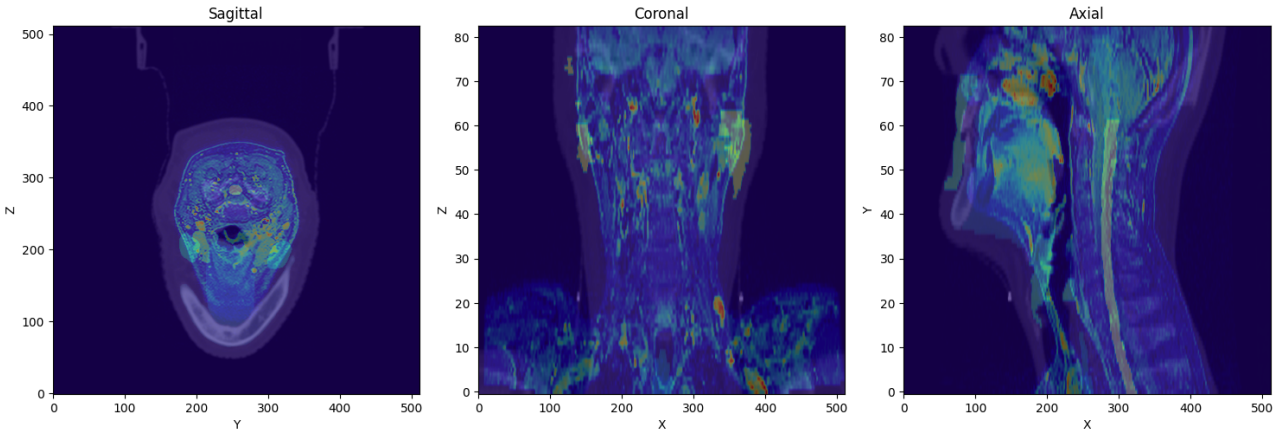


Figure 11: Overlay of resampled and aligned CT and MRI modalities for a representative patient case.

zooming, and lighting variations are used to prepare the model to recognize OARs in different conditions, which helps to improve the model's robustness and reliability.

4.5 Model, training, and validation

We discuss the 3D-UNet model's training procedure in this subsection. The model was trained on the training dataset using the hyperparameters presented in Table 1. As detailed, we used the Adam optimizer for its adaptive gradient descent ability and Categorical Focal Cross-Entropy to reduce class imbalance, particularly in underrepresented organs. Furthermore, early stopping with a 10-epoch patience, combined with a learning rate decay method driven by performance limitations, aided in the prevention of overfitting and convergence.

Table 1: Hyperparameters for Scenario A

Hyperparameter	**Value**
Loss function	Categorical Focal Cross-Entropy
Optimizer	Adam
Early stopping	Enabled (Patience = 10 epochs)
Learning rate decay from 1e-04 to 1e-07	On plateau (Patience = 5 epochs)

Fig. 12 illustrates the model's performance metrics, which show effective and steady training behavior. There is an early rising trend in both training and validation accuracy, converging to 1; after 35 epochs, the training accuracy reached about 94.2%. The cross-entropy curves also show a gradual decrease converging to 0, indicating that the model may successfully lower classification errors with time. The use of categorical focal loss addresses the class imbalance that causes the training loss to maintain relatively higher values even as it steadily declines. This modification improved the learning dynamics of the model while reducing training loss. Additionally, the learning rate decreases from $1e{-}^{04}$ to $1e^{-07}$ across the epochs, showcasing a strategic approach to training optimization. This constant decrease in learning rate enhances the model's generalizability to new data and helps avoid overfitting. The convergence of training and validation metrics indicates that the combination of various training methodologies yields a well-balanced training process overall, ultimately enhancing model performance.

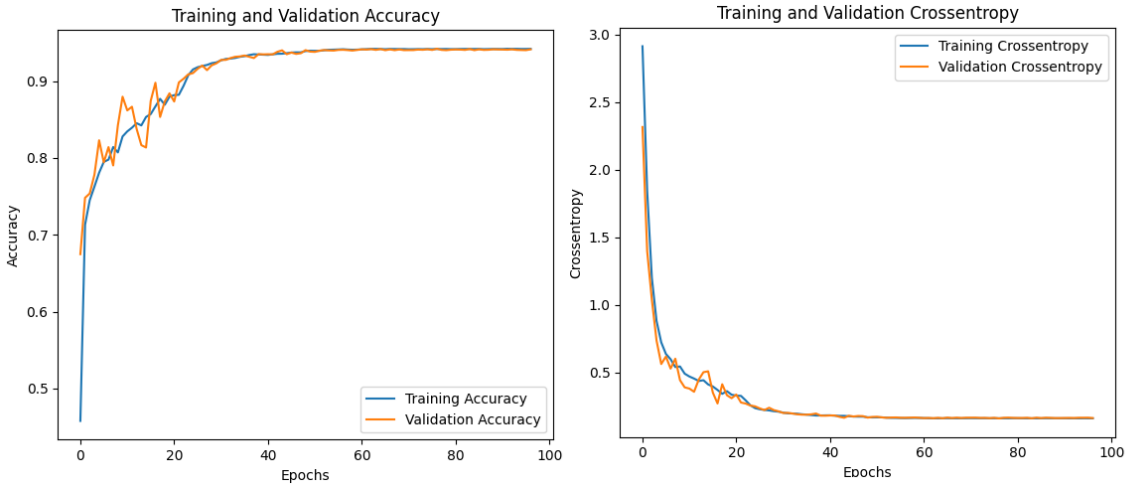


Figure 12: 3D-Unet Model Training History Using categorical cross-entropy.

4.6 Model Testing and Evaluation

We assessed the model's performance on the validation and test sets by calculating the Jaccard and F1 scores. Fig. 13 shows a wide range of F1 and Jaccard scores across different classes.

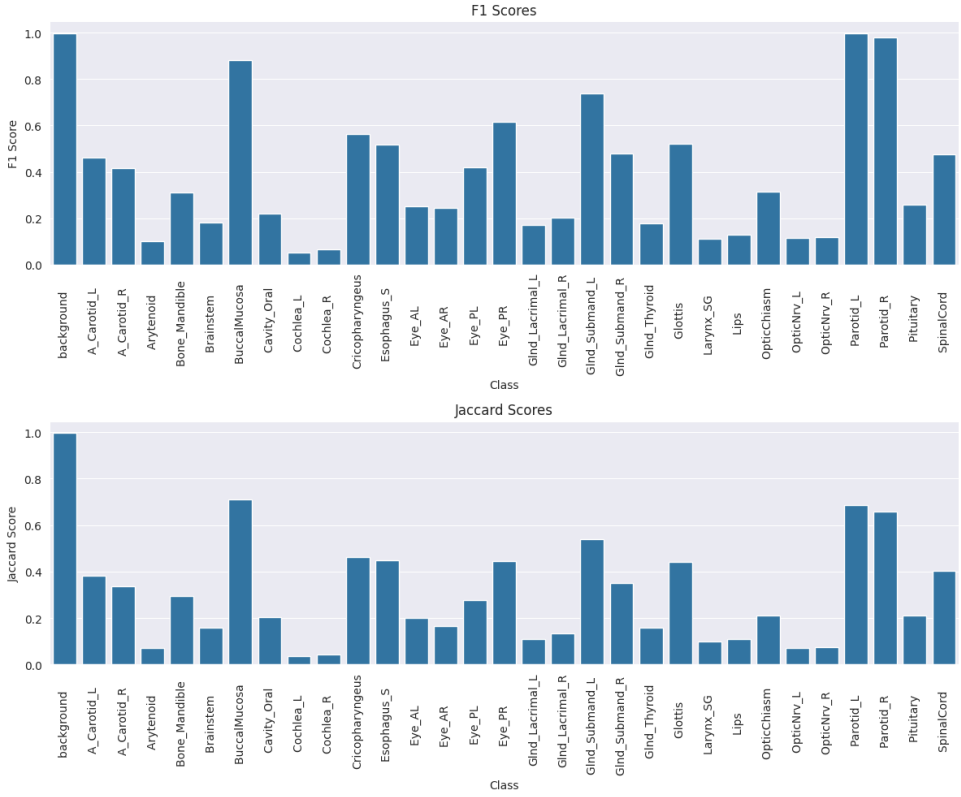


Figure 13: Obtained F1 score and Jaccard index.

– The Jaccard similarity coefficient, also referred to as the Jaccard index or Jaccard similarity, gauges the

similarity between two sets by comparing their intersection to their union. It is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are sets, and $|A|$ denotes the cardinality of set A. The Jaccard coefficient spans from 0 to 1, where 0 indicates no overlap between the sets, and 1 indicates complete overlap.

– The F1 score, a measure of a test's accuracy, takes into account both the precision P and the recall R of the test to calculate the score. It is the harmonic mean of precision and recall, defined as:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

The F1 score achieves its optimal value at 1 and its poorest at 0. It is a useful metric for imbalanced classes because it considers both false positives and false negatives.

In Fig. 13, we can detect:

– **High-Performing Classes**: Classes like Parotid L, Parotid R, BuccalMucosa, and Glnd Submand L have very high F1 scores (> 0.9) and Jaccard indices (> 0.6), indicating excellent performance in segmentation.

– **Medium-Performing Classes**: Classes like Cricopharyngeus, Oesophagus S, Eye PR, and Glottis have moderate F1 scores ($0.4 - 0.6$) and Jaccard indices ($.3 - 0.5$), suggesting acceptable performance.

– **Low-Performing Classes**: Classes like Arytenoid, Bone Mandible, Brainstem, and Cochlea L have low F1 scores (< 0.3) and Jaccard indices (< 0.2), indicating poor segmentation performance.

The segmentation performance of the proposed multimodal 3D-UNet model is presented in Table 2 and compared with the state-of-the-art models.

The proposed 3D-UNet achieved a mean F1 score of 0.778 ± 0.041 across 30 OARs after being trained on combined CT and MRI volumes utilizing an early-fusion approach. The importance of integrating MRI data for better delineation of soft-tissue boundaries and small, low-contrast structures is demonstrated by this result, which outperforms the CT-only baseline models, SAM and the traditional CNNs (VGG16/19).

In comparison to AnatomyNet [11], a CT-only 3D residual U-Net with an average DSC of roughly 0.75 on the MICCAI HaN dataset, our model achieves comparable or slightly greater performance while being computationally lighter and requiring fewer architectural parameters. Furthermore, our early-fusion approach yields comparable segmentation accuracy with significantly lower memory and training costs when compared to HyperDenseNet [15], which uses dense cross-connections across modality-specific encoders.

These results highlight how segmentation quality, efficiency, and clinical usefulness may all be strongly balanced in an early-fusion system that has been rigorously adjusted. The reliability and stability of the proposed method for automatic planning are further demonstrated by the low standard deviation of results across all OARs.

4.7 Segmentation Results

As not all images contain the entire collection of 30 organs at risk (OARs) due to differences in the field of view, we computed the mean assessment metric for each OAR separately, followed by an overall mean to ensure equal class weighting. Our proposed 3D-UNet model showed good segmentation accuracy (94.2% on the training set and 94.1% on the test set), indicating significant potential for assisting planning in head and neck cancer treatment.

Table 2: Comparison of segmentation performance across multimodal and CT-only approaches.

Model	**Imaging Modality**	**Approach**	**Nb of OARs**	**Performance (mean $\pm$ std)**
SAM [17]	CT	Statistical + Post-processing	28	F1 $= 0.648 \pm 0.052$
VGG16/19 [17]	CT	Classic CNN (2D)	28	F1 $= 0.121 \pm 0.166$
AnatomyNet [11]	CT	3D Residual U -Net (CT-only)	9	DSC $\approx$ 0.75 (reported)
HyperDenseNet [15]	CT + MRI	Multi-path dense fusion (3D)	12	DSC $\approx$ 0.78 (reported)
3D-UNet (proposed)	CT + MRI	Early-fusion 3D-UNet (ours)	30	**F1 $= 0.778 \pm 0.041$**

Fig. 14 illustrates how the model performs in various anatomical scenarios. Subfigures (a) and (b) demonstrate the effective segmentation of large, well-defined structures such as the parotid and submandibular glands. These organs are usually easier to segment because of their constant shape, larger size, and high intensity contrast with the surrounding tissues. Subfigure (c) depicts a more problematic scenario involving the cochlea, with more obvious segmentation errors.

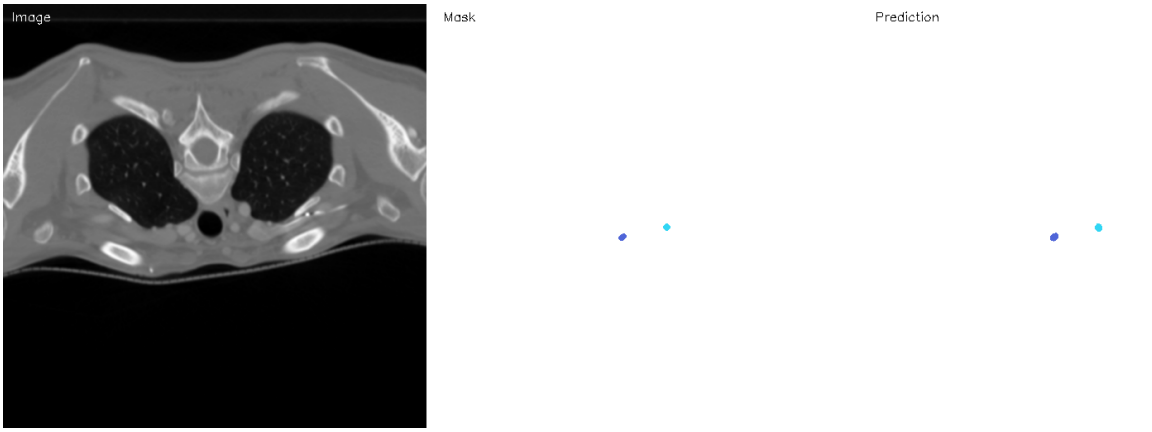
These findings are in line with the quantitative data, which revealed significant variation in segmentation performance among organ classes. While the model performs well on larger, more visible organs, it suffers with smaller, low-contrast parts like the cochlea, brainstem, and arytenoid cartilages. Several factors contribute to this challenge: (1) limited voxel representation rises sensitivity to small errors; (2) weak intensity contrast complicates boundary detection; (3) registration inaccuracies between modalities can skew fine structures; and (4) class imbalance reduces the model's sensitivity to underrepresented organs during training.

Fig. 15 shows a representative case with predicted segmentation findings placed on CT scans across different anatomical planes (axial, sagittal, coronal) and in a volumetric representation. The overlay demonstrates the model's capacity to reliably identify numerous OARs simultaneously, even in anatomically complicated regions of the head and neck. Color-coded contours make it easier to visualize each segmented structure, illustrating the model's ability to preserve organ boundaries and spatial consistency. This visual outcome reinforces the quantitative results, demonstrating the model's ability to handle multimodal data and generate clinically useful predictions.

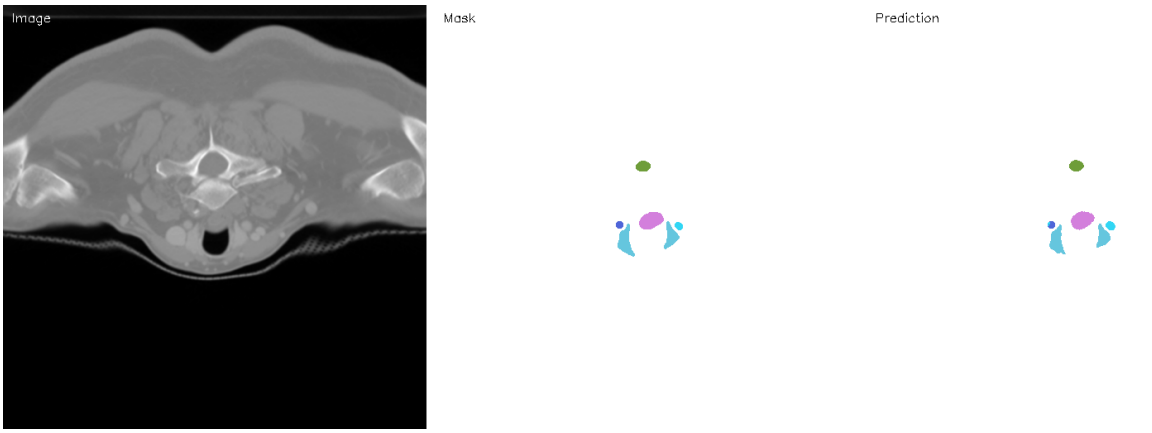
5 Conclusion and Future Work

In this work, a 3D-UNet architecture-based fully automated multimodal segmentation system for head and neck organs-at-risk (OARs) was described. The model uses complementary spatial and contrast information to improve segmentation accuracy by combining co-registered CT and MRI volumes using an early-fusion technique. The proposed approach demonstrated strong anatomical consistency and competitive quantitative results. These findings indicate its potential integration into clinical radiation-planning workflows. It serves as a useful foundation for upcoming studies in multimodal medical imaging segmentation because of its reproducibility and lightweight architecture.

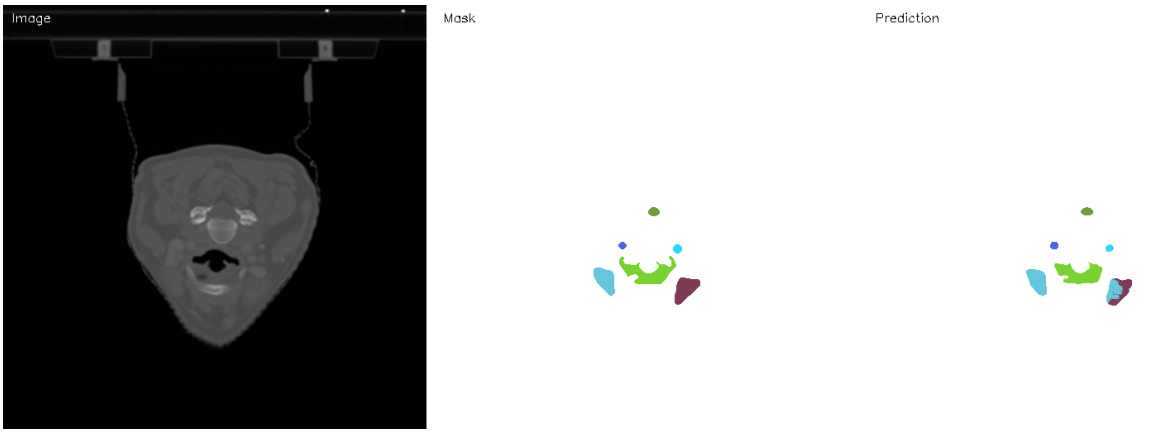
From a medical point of view, the model offers a dependable and computationally effective method for defining OARs, especially in complicated head and neck regions where exact boundary identification is critical for treatment safety. However, because it combines modalities before allowing the network to learn modality-specific abstractions, the early-fusion approach has fundamental constraints. This simplicity renders performance subject to registration issues and may limit the ability to capture the unique features of CT and MRI, such as contrast or resolution variations.



(a) Well-segmented case: Parotid glands and surrounding structures are accurately delineated.



(b) Well-segmented case: Consistent prediction of large organs with good anatomical clarity.



(c) Challenging case: Cochlea and small structures are poorly segmented.

Figure 14: Visual comparison between ground truth and predicted segmentation.

Future research will focus on expanding this work in three directions. First, combining modality-specific encoders with a common decoder may help to improve cross-modal feature learning while maintaining the distinct representational advantages of each modality. Second, by investigating transformer-based architectures, the model will be able to use self-attention processes to capture complex intermodality interactions and long-range spatial dependencies. Finally, combining self-supervised and weakly supervised learning techniques could reduce the need for extensive, finely annotated datasets while improving generalization and robustness in

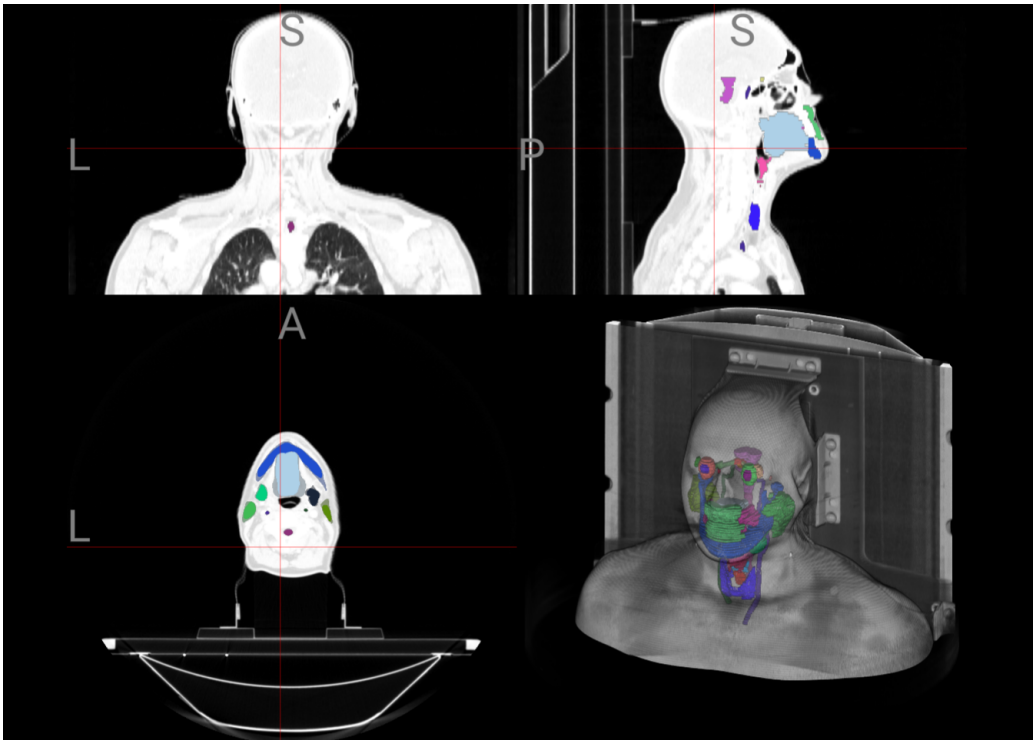


Figure 15: Case Prediction.

actual clinical settings.

This study establishes a reproducible multimodal baseline for head-and-neck OARs segmentation in radiotherapy planning.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, ”On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. https://doi.org/10.1098/rsta.1955.0005

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68-73.

[3] I. S. Jacobs and C. P. Bean, ”Fine particles, thin films, and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350. https://doi.org/10.1016/B978-0-12-575303-6.50013-0

[4] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, ”Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Trans. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982]. https://doi.org/10.1109/TJMJ.1987.4549593

[5] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] Raudaschl, Patrik F., et al., ”Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015.” Medical Physics 44.5 (2017): 2020-2036. https://doi.org/10.1002/mp.12197

[7] Street, Ethan, et al., ”Automated volume analysis of head and neck lesions on CT scans using 3D level set segmentation.” Medical physics 34.11 (2007): 4399-4408. https://doi.org/10.1118/1.2794174

[8] Han, Xiao, et al. "Atlas-based auto-segmentation of head and neck CT images." Medical Image Computing and Computer-Assisted Intervention-MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part II 11. Springer Berlin Heidelberg, 2008.

[9] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks:,"

Med. Phys., vol. 44, no. 2, pp. 547-557, 2017. https://doi.org/10.1002/mp.12045

[10] X. Ren et al., "Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images," Med. Phys., vol. 45, no. 5, pp. 2063-2075, 2018. https://doi.org/10.1002/mp.12837

[11] W. Zhu et al., "AnatomyNet: Deep Learning for Fast and Fully Automated Whole-volume Segmentation of Head and Neck Anatomy," 2018. https://doi.org/10.1002/mp.13300

[12] Yan, Y., et al.: Longitudinal detection of diabetic retinopathy early severity grade changes using deep learning. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) OMIA 2021. LNCS, vol. 12970, pp. 11–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87000-3 2

[13] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," vol. 45, no. October, pp. 4558– 4567, 2018. https://doi.org/10.1002/mp.13147.

[14] Zhang, Y., Sidib´e, D., Morel, O., M´eriaudeau, F.: Deep multimodal fusion for semantic image segmentation: a survey. Image Vis. Comput. 105, 104042 (2021). https://doi.org/10.1016/j.imavis.2020.104042

[15] Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging 38, 1116–1126 (2019). https://doi.org/10.1109/TMI.2018.2878669

[16] Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: HaN-Seg: the head and neck organ-at-risk CT and MRI segmentation dataset. Med. Phys. 50, 1917–1927 (2023). https://doi.org/10.1002/mp.16197

[17] Wei, J., J. Deng, and M. Chao. "SAM-Based Auto-Segmentation of Critical Structures for Head and Neck Cancer Radiotherapy." International Journal of Radiation Oncology, Biology, Physics 120.2 (2024): e113-e114. https://doi.org/10.1016/j.ijrobp.2024.07.2034