Automatic Segmentation of Organ at Risk in Head and Neck Cancer CT Images Using Medical Open Network for Artificial Intelligence (MONAI) with Deep Learning Techniques

.

ABSTRACT

|  |
| --- |
| **Aims:** This study aims to evaluate the deep learning-based Medical Open Network for Artificial Intelegence (MONAI) framework in performing automatic segmentation of Organ at risk (OAR) structures, especially in cases of head and neck cancer.  **Study design:** this research an experimental design with a quantitative approach to predict the segmentation anotated data in head and neck cancer using U-Net model. The study involved training and testin using data from The Cancer Imaging Archive (TCIA) and publicly available dataset, evaluating its performance based on balanced accuracy, precision, recall, and Dice Similarity Coefficient (DSC). The result was analyzed to determine the model accuracy in replicating actual anotated segmentation, providing a data driven assessment of its predictive capability  **Place and Duration of Study:** This research was conducted in Department of Physics Study Program, Udayana University, between January 2024 and June 2025.  **Methodology:** The dataset with 179 subjects consist of two dataset sources, namely the HEAD-NECK-RADIOMICS-HN1 (H&N1) Collection and the Han-Seg collection with CT image sizes of 512 x 512 x 116 - 512 x 512 x 323. The model trains and tests using that annotated dataset, focusing on nine organ labels: brainstem, right submandibular, left submandibular, mandible, optic chiasm, optic nerve, right parotid, left parotid, and spinal cord. The *U-Net* model was implemented using the MONAI framework with preprocessing stages of converting the dataset to a *NIfTI* file, creating a *dataset.json* file, label remapping and applying MONAI transforms, namely deterministic and augmentation. The *U-Net* model evaluation is performed using balanced accuracy, precision, recall, and DSC.  **Results:** The model showed good segmentation performance with an average balanced accuracy value of 0,932 ± 0,043, precision of 0,690 ± 0,189, and recall of 0,863 ± 0,086. The mean value of Dice Similarity Coefficient (DSC) is 0.758 ± 0.132. The highest Dice score obtain in the mandible organ at 0.902, while the lowest value obtain in the optic chiasm at 0.531. The visualization of the results supports the finding that segmentation is more accurate on large and clear organs.  **Conclusion:** The MONAI framework is able to perform automatic OAR segmentation with promising results, especially on large organs. However, performance on small organs is still low due to voxel limitations and label imbalance. |

***Keywords****: Automated Segmentation, MONAI, OAR, Head and Neck Cancer, Deep Learning*

1. INTRODUCTION

Head and neck cancer is among the leading causes of cancer-related deaths, ranking as the sixth most commonly diagnosed cancer globally, with approximately 850,000 new cases and 450,000 deaths annually (Barsouk et al., 2023). This type of cancer includes malignancies of the oral cavity, salivary glands, paranasal sinuses, nasal cavity, nasopharynx, oropharynx, hypopharynx, and larynx (Laura Q.M, Chow, 2012). Radiotherapy is a common treatment modality for head and neck cancer, delivering high doses of radiation. It is widely used due to its ability to conform the dose shape to the actual anatomy while sparing nearby healthy tissues and organs, known as Organs at Risk (OARs) (Lee & Le, 2008)

A large portion of cancer treatments involves radiotherapy, which carries the risk of damaging healthy tissue and causing various complications. Therefore, one of the critical steps in radiotherapy is the segmentation of OARs in medical images. Image segmentation is the process of dividing an image into multiple regions (Isaksson et al., 2023). In radiotherapy, segmentation provides an accurate three-dimensional (3D) spatial description of OARs. Clinically, segmentation is still often perform manually, which is time-consuming (Wong et al., 2020). The complexity of OAR morphology and limitations of imaging tools make manual delineation prone to errors. To minimize post-treatment complications, OARs such as the brainstem, spinal cord, mandible, larynx, pharynx, parotid glands, submandibular glands, nasopharynx, eyes, and optic nerves must be accurately delineated (Harari et al., 2010). Accurate delineation of OARs is essential in treatment planning for head and neck cancer (Sharp et al., 2014).

In recent years, there has been increasing research into applying deep learning for medical image segmentation. One area of focus is the automatic and accurate delineation of OARs using Convolutional Neural Networks (CNNs). CNNs use a multi-layer perceptron architecture where the image passes through a series of computational layers that extract and recognize consistent intensity patterns, ultimately producing predictions. CNNs account for spatial information, allowing nearby pixels to be analyzed together (Ibragimov & Xing, 2017). Deep learning methods, particularly CNN-based ones, have proven to be promising technologies that significantly reduce the time required for medical image segmentation. These methods have demonstrated superior performance compared to manual segmentation (Cardenas et al., 2018)

One example of deep learning in OAR segmentation for head and neck cancer is the study by Siciarz and McCurdy, which employed a U-Net architecture embedded with Inception-ResNet-v2 blocks. Their model achieved a Dice score of 0.82 ± 0.10 across 25 OARs (Siciarz & McCurdy, 2022) Another study trained segmentation models using partially labeled datasets. From 44 independent datasets, the segmentation of 15 OARs in head and neck cancer yielded an average Dice Similarity Coefficient (DSC) of 80.59% (Cubero et al., 2022)

This study utilizes a specialized open-source framework for medical imaging called the MONAI Project (<https://monai.io>). MONAI offers essential components optimized for AI development, including 2D and 3D medical image segmentation. It provides comprehensive transformations for medical imaging, such as I/O, spatial, intensity, cropping/padding, and more. Additionally, MONAI includes pre-trained architectures on medical images that support configurable spatial dimensions for reuse (Cardoso et al., 2021)

Despite ongoing advances, the clinical adoption of automatic segmentation tools remains limited by technical barriers, lack of generalizability, and integration challenges. With the growing number of patients requiring radiotherapy and the complexity of anatomical structures involved, there is an urgent need for robust, open-source, and reproducible solutions that can be readily implemented in clinical workflows. This study addresses that gap by utilizing the MONAI framework to develop and evaluate a deep learning-based segmentation model for OARs in head and neck cancer CT images.

2. material and methods

# Materials

The dataset total of 179 cases, with 125 cases for training, 35 cases for validation, and 19 cases for testing. The datasets were used with the following dataset source, which is The HEAD-NECK-RADIOMICS-HN1 (H&N1) collection and The HaN-Seg collection

1. The HEAD-NECK-RADIOMICS-HN1 (H&N1) collectionwas obtained from The Cancer Imaging Archive (TCIA) under the University of Arkansas for Medical Science (Wee & Dekker, 2019)**.** This dataset from 137 patients of radiotherapy contains CT images with voxel sizes of 512 x 512 x 134 – 512 x 512 x 304, organ contour structure, and radiation therapy structures. A restricted license agreement has been implemented to ensure compliance with regulations and ethical standards set by the dataset provider for the use of this data.
2. The HaN-Seg collection with 42 Patients Head and Neck CT Medical Images with sizes 512 x 512 x 116 – 512 x 512 x 323 and segmentation references that can be accessed in the journal HaN-Seg: The Head and Neck Organ-at-Risk CT & MR Segmentation Dataset (<https://doi.org/10.5281/zenodo.7442914>)

# Methods

**2.2.1. Data Preprocessing**

The data preprocessing pipeline was the application of data transformations using the MONAI Transform framework. Two categories of transformations were implemented i.e deterministic transformations, which were applied consistently across all data groups (training, validation, and test sets), and augmentation transformations, which were exclusively applied to the training data to increase dataset size and variability.

The deterministic transformations included several essential preprocessing steps:

1. Loading the images (*monai.transforms.LoadImaged*),
2. Ensuring the channel dimension is first (*monai.transform.EnsureChannelFirstd*),
3. Reorienting the images to the RAS coordinate system (*monai.transform.Orientationd* with *axcodes="RAS"*),
4. Resampling the voxel spacing to 1.5 × 1.5 × 3 mm using bilinear interpolation for images and nearest-neighbor interpolation for labels (*monai.transform.Spacingd*),
5. Scaling the intensity range to normalize pixel values between 0.0 and 1.0 (*monai.transform.ScaledIntensityRanged*), and
6. Cropping the foreground region while allowing smaller outputs (*monai.transform.CropForegroundd*).

For data augmentation, a series of probabilistic transformations were applied to the training set to enhance model generalizability. These included

1. Random cropping based on positive and negative label sampling (monai.transform.*RandCropByPosNegLabeld*) with a spatial size of 96 × 96 × 96 voxels and four samples per image,
2. Random flipping along spatial axes (monai.transform.*RandFlipd*) with a probability of 0.10,
3. Random rotations by 90 degrees (monai.transform.*RandRotate90d*) applied with a probability of 0.10 and up to three rotation axes, and
4. Random intensity shifting (monai.transform.*RandShiftIntensityd*) with offsets of 0.10 and a probability of 0.50.

Figure (1) shows examples of the datasets after the transformations were applied.

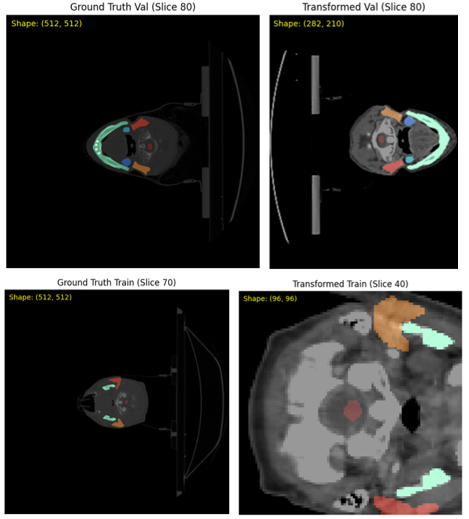
****

Fig. 1. Transformations applied to the validation data (top) and transformations applied to the training data (bottom).

**2.2.2. Setup Model, Optimizer and Loss**

The next stage in this study involved selecting the model architecture, loss function, and optimizer. The model employed was a 3D UNet with the following configuration: *spatial\_dims* set to 3, one input channel (*in\_channels=1*) and ten output channels (*out\_channels=10*), convolutional channel depths of (32, 64, 128, 256, 320), *strides* of (2, 2, 2, 2), two residual units at each level (*num\_res\_units=2*), and batch normalization (*norm='batch'*). The model was executed on the available GPU device (*device*). The training process was configured with a *patch\_size* of (96, 96, 96), a sliding window batch size of 1, training batch size of 1, validation batch size of 1, a maximum of 15,000 iterations, and performance evaluation conducted every 500 iterations.

To improve segmentation accuracy and address class imbalance, a composite loss function, DiceCELoss, was used, combining Dice Loss and Cross Entropy Loss. Dice Loss is effective in measuring the degree of overlap between the predicted output and the ground truth labels, particularly for small anatomical structures, while Cross Entropy Loss penalizes incorrect prediction probabilities. The combination of both loss functions enhances model performance in complex and imbalanced medical imaging data. Model optimization was carried out using the AdamW algorithm, a variant of Adam that explicitly incorporates weight decay regularization, thereby improving resistance to overfitting. The optimizer was configured with a learning rate of 1e-4, commonly used to ensure training stability, and a *weight decay* parameter of 1e-5, which serves as additional regularization.

**2.2.3. Model Evaluation**

Segmentation prediction and evaluation were performed on 10% of the test data (n=19). In evaluating the model involves visualizing the actual segmentation, comparing the predicted segmentation, and evaluating the predicted segmentation on the performance metric. The information in the prediction results, specifically the number of voxels per label, can be classified into four categories: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). For the multiclass label case, these four categories can be visualized using a confusion matrix, as shown in Figure (3).

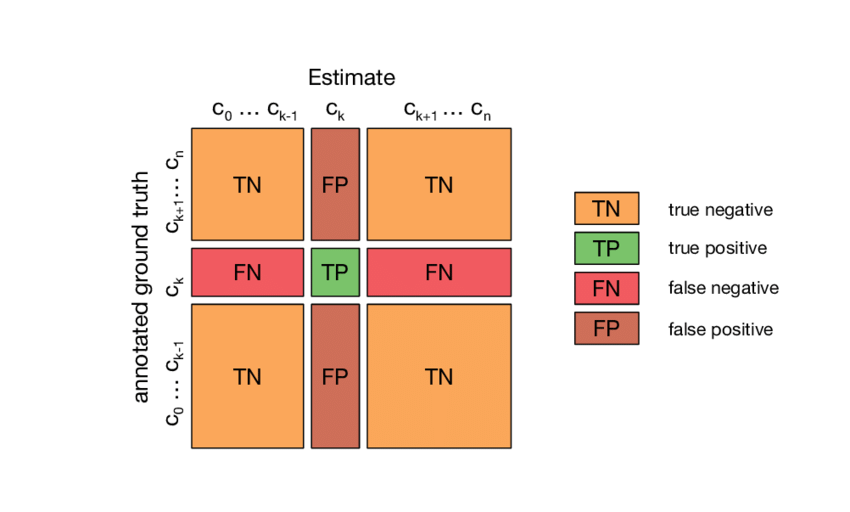


Fig. 2. Multilabel Confusion Matrix,

c0 = label 0,

ck−1 = evaluated label minus 1,

ck = evaluated label,

ck+1 = evaluated label plus 1,

cn = label n

The results from these categories are used to calculate performance metrics such as balanced accuracy, precision, recall, and dice similarity score (DSC). Balanced accuracy is the average of sensitivity (true positive rate) and specificity (true negative rate) and is used when dealing with imbalanced data, i.e., when one target class appears much more frequently than the others (Fulazzaky et al., 2024). It is defined using the following equation:

|  |  |
| --- | --- |
|  | (1) |

Precision effectively describes the purity of positive detections relative to the ground truth. Mathematically, it is defined as follows (Benjdira et al., 2020)

|  |  |
| --- | --- |
|  | (2) |

Recall effectively describes the completeness of positive predictions relative to the ground truth. Mathematically, it is defined as follows (Benjdira et al., 2020):

|  |  |
| --- | --- |
|  | (3) |

The Dice Similarity Coefficient (DSC) is useful in the fields of image segmentation and object detection, as it helps measure the overlap between the predicted data and the ground truth data. The DSC is used to quantify the degree of overlap and similarity between the ground truth segmentation and the predicted segmentation, with values ranging from 0 to 1. Mathematically, it is defined as follows (Benjdira et al., 2020)

|  |  |
| --- | --- |
|  | (4) |

3. results and discussion

The model training was conducted using 70% (n=125) training data and 20% (n=35) validation data. During the training phase, the model trained by using 125 training data as input, and the outputs were compared with the ground truth to calculate the loss. The model then evaluated using 35 validation data every 500 iterations. This process was repeat until the model reached the maximum of 15,000 iterations. The training results are present in the iteration average loss and validation mean dice graphs shown in Figure (4) below. The average loss graph (on the left) shows fluctuations after several iterations. This indicates that there is some variation within the training data, however overall, the loss trend decreases, which suggests that the training proceeded well. Meanwhile, in the validation mean dice graph (on the right), the values show an increase until they stabilize at a dice score of approximately (± 0.71) on the validation data. This stabilization indicates that the model has converged.

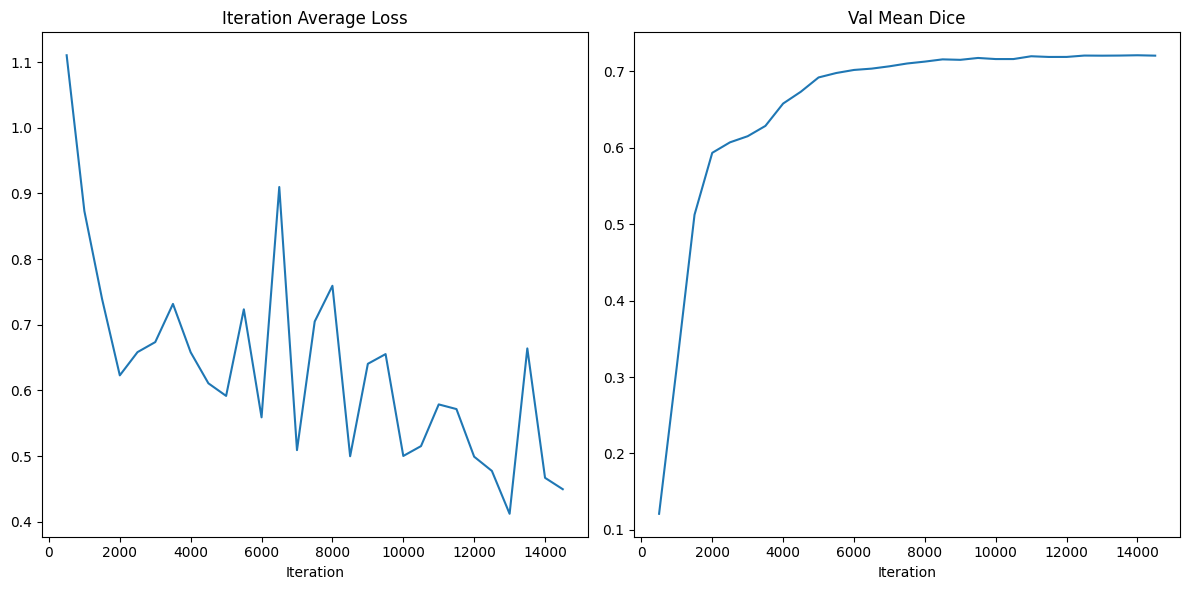


Fig. 3. Iteration average loss (left side), val mean dice (right)

1. Visualization of contours

Subsequently, the model was evaluated using a 10% test set (n=19). First, the model was asses qualitatively through visualization of the ground truth images, transformed images, and predicted contours, as shown in Figure (5). In Figure (5) the ground truth image has a resolution of 512×512 pixels, while the transformed image has a resolution of 321×214 pixels. This difference results from transformations such as resampling and cropping, which were applied to adjust the resolution and focus on the segmentation area. The prediction image displays the deep learning model’s segmentation results for the same organ. Based on the visualization, the predicted area (indicated by the red arrow) demonstrates a reasonably good agreement with the reference labels in the transformed images. Nevertheless, slight differences can be observed along the segmentation boundaries, which are likely caused by the loss of spatial detail due to *down sampling* or by the model’s limitations in precisely detecting the edges of the organ structures.

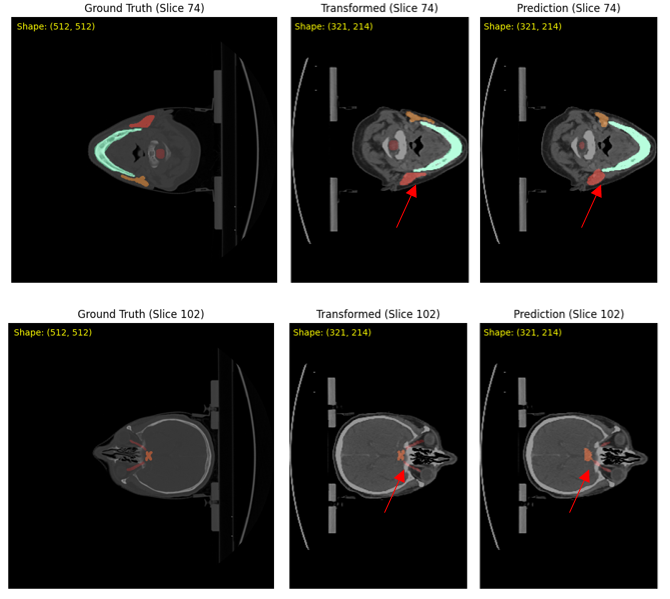


Fig. 4. 2D visualization of CT images in the axial plane

1. *Accuracy, Precision, Recall*

The model appears to lack sufficient information for organs with small anatomical structures compared to larger organs, such as the mandible. Specifically, the voxels representing organs with small anatomy (the optic chiasm and optic nerve) were very limited in the testing data, as shown in the confusion matrix in Figure (6). The number of voxels appears to be imbalanced among the labels, with most voxels dominated by the background. Organs with small anatomical structures, such as the optic chiasm, have a much smaller number of voxels compared to the others. From this confusion matrix, performance metrics like balanced accuracy, precision, and recall can be calculated.

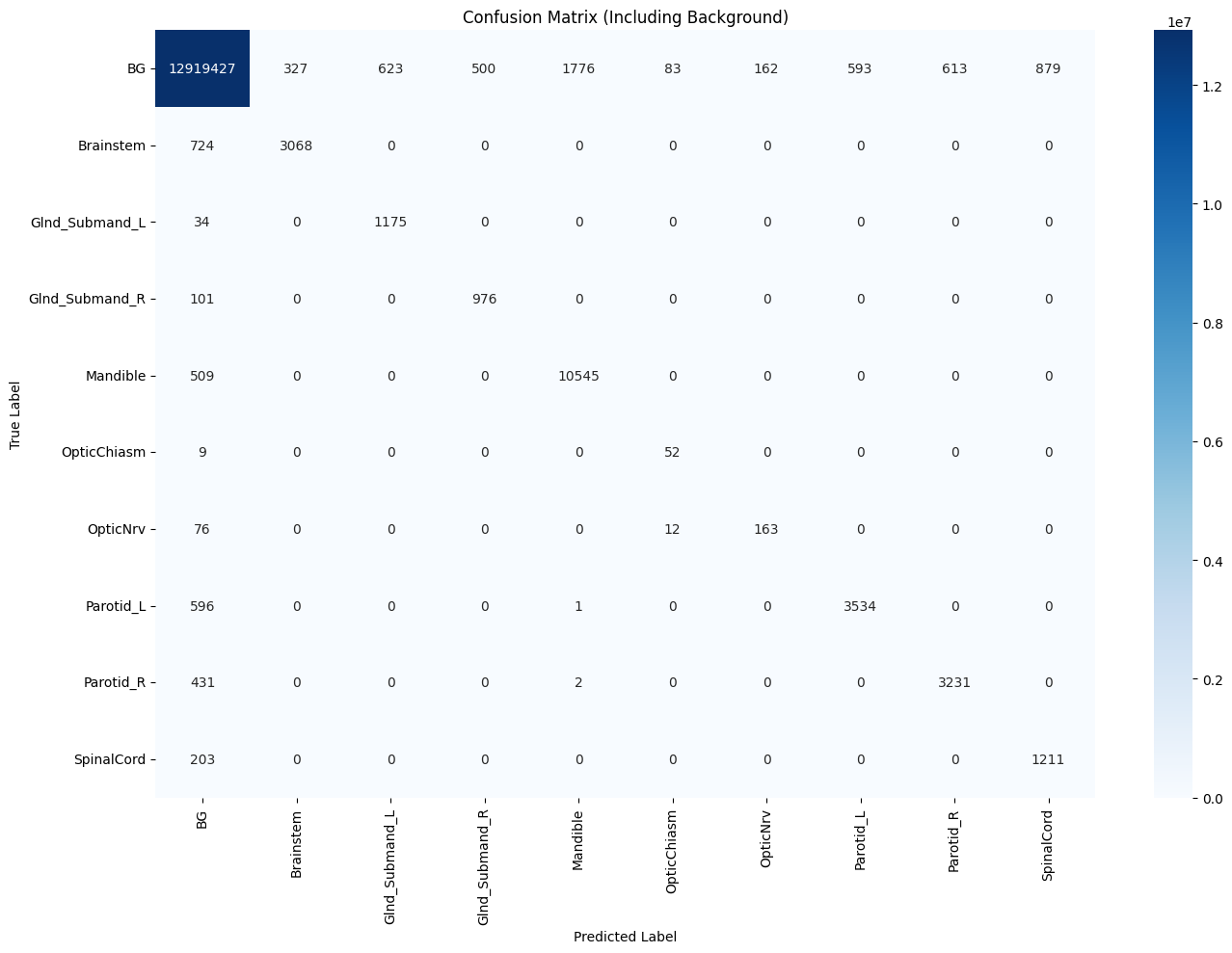


Fig. 5. Confusion Matrix on Test Data (including background)

Table (1) shows the results of balanced accuracy, precision, and recall for all labels or organs at risk (OARs). Based on Table (1), the precision values are noticeably lower compared to recall and balanced accuracy. The lowest precision among all OARs was observed for the optic chiasm at 0.353, while the highest was for the brainstem at 0.904. This indicates that for some labels, such as the optic chiasm, there were a high number of false positives. In other words, the model had difficulty distinguishing between the actual OAR and the predicted OAR.

**Table 1. Balanced Accuracy, Precision, and Recall for each label**

|  |  |  |  |
| --- | --- | --- | --- |
| OARs | Balanced Accuracy | Precision | Recall |
| Brainstem | 0.905 | 0.904 | 0.809 |
| Right submandibular | 0.986 | 0.654 | 0.972 |
| Left submandibular | 0.953 | 0.661 | 0.906 |
| Mandible | 0.977 | 0.856 | 0.954 |
| Optic chiasm | 0.926 | 0.354 | 0.852 |
| Optic nerve | 0.841 | 0.502 | 0.682 |
| Left parotid | 0.928 | 0.856 | 0.855 |
| Right parotid | 0.941 | 0.841 | 0.882 |
| Spinal cord | 0.928 | 0.579 | 0.856 |
| **Mean ± Std** | **0.932 ± 0.043** | **0.690 ± 0.189** | **0.863 ± 0.085** |

1. Dice Similarity Score

The DSC values were calculate using equation (4) separately for each label, the results as shown in Table (2). This calculation aims to evaluate the level of agreement between the automatic segmentation results and the ground truth for each organ or structure. The higher the DSC value, the better the model’s segmentation performance, as it indicates that the model’s predictions have a high degree of overlap with the reference annotations. Table (2) presents the DSC values from the automatic segmentation results for each OAR. The highest DSC was achieve for the mandible (0.902), while the lowest was found for the optic chiasm (0.500), with an average DSC of 0.753 ± 0.141.

**Table 2. The Dice Similarity Coefficient (DSC) for each label**

|  |  |
| --- | --- |
| OARs | DSC |
| Mandible | 0.902 |
| Right parotid | 0.861 |
| Left parotid | 0.856 |
| Brainstem | 0.854 |
| Left Submandibular | 0.782 |
| Right Submandibular | 0.765 |
| Spinal Cord | 0.691 |
| Optic nerve | 0.566 |
| Optic chiasm | 0.500 |
| **Mean DSC ± Std** | **0,753 ± 0,141** |

4. Conclusion

Based on metric analysis and visualization, automatic segmentation using deep learning with the MONAI framework is feasible; the main challenge lies in the imbalance of voxel data dominated by background labels. The model performed well in segmenting large, clearly defined OARs such as the mandible but struggled to identify small and complex structures like the optic chiasm and optic nerve. Evaluation on 19 test cases showed an average balanced accuracy of 0.932 ± 0.043, precision of 0.690 ± 0.189, and recall of 0.863 ± 0.086 across all OAR labels. The overall mean Dice Similarity Coefficient was 0.758 ± 0.132, with the highest score for the mandible (0.902) and the lowest for the optic chiasm (0.531).

Consent

All authors declare that ‘written informed consent was obtained from the patient (or other approved parties) for publication of this case report and accompanying images.

Ethical approval

All authors hereby declare that all experiments have been examined and approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

References

Barsouk, A., Aluru, J. S., Rawla, P., Saginala, K., & Barsouk, A. (2023). Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma. Medical Sciences, 11(2), 42. <https://doi.org/10.3390/medsci11020042>

Benjdira, B., Ammar, A., Koubaa, A., & Ouni, K. (2020). Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks. Applied Sciences (Switzerland), 10(3), 1–24. <https://doi.org/10.3390/app10031092>

Cardenas, C. E., McCarroll, R. E., Court, L. E., Elgohari, B. A., Elhalawani, H., Fuller, C. D., Kamal, M. J., Meheissen, M. A. M., Mohamed, A. S. R., Rao, A., Williams, B., Wong, A., Yang, J., & Aristophanous, M. (2018). Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. International Journal of Radiation Oncology\*Biology\*Physics, 101(2), 468–478. <https://doi.org/10.1016/J.IJROBP.2018.01.114>

Cardoso, M. J., Li, W., Brown, R., Kerfoot, E., Wang, Y., Myronenko, A., Zhu, W., Hashemian, B., Vercauteren, T., & Wang, G. (2021). arXiv : 2211 . 02701v1 [ cs . LG ] 4 Nov 2022 MONAI : An open-source framework for deep learning in healthcare.

Cubero, L., Castelli, J., Simon, A., de Crevoisier, R., Acosta, O., & Pascau, J. (2022). Deep Learning-Based Segmentation of Head and Neck Organs-at-Risk with Clinical Partially Labeled Data. Entropy, 24(11). <https://doi.org/10.3390/e24111661>

Fulazzaky, T., Saefuddin, A., & Soleh, A. M. (2024). Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning : A Comparative Analysis of Balanced Random. 11(4), 969–980. <https://doi.org/10.15294/sji.v11i4.15937>

Harari, P. M., Song, S., & Tomé, W. A. (2010). Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. International Journal of Radiation Oncology, Biology, Physics, 77(3), 950–958. <https://doi.org/10.1016/j.ijrobp.2009.09.062>

Ibragimov, B., & Xing, L. (2017). Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Medical Physics, 44(2), 547–557. <https://doi.org/10.1002/mp.12045>

Isaksson, L. J., Summers, P., Mastroleo, F., Marvaso, G., Corrao, G., Vincini, M. G., Zaffaroni, M., Ceci, F., Petralia, G., Orecchia, R., & Jereczek-Fossa, B. A. (2023). Automatic Segmentation with Deep Learning in Radiotherapy. Cancers, 15(17). <https://doi.org/10.3390/cancers15174389>

Laura Q.M, Chow, M. D. (2012). Head and Neck Cancer. The New England Journal of Medicine, 1, 1257–1264. <https://doi.org/10.1016/B978-1-4377-1604-7.00196-2>

Lee, N. Y., & Le, Q. T. (2008). New Developments in Radiation Therapy for Head and Neck Cancer: Intensity-Modulated Radiation Therapy and Hypoxia Targeting. Seminars in Oncology, 35(3), 236–250. <https://doi.org/10.1053/j.seminoncol.2008.03.003>

Sharp, G., Fritscher, K. D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H., & Yang, J. (2014). Vision 20/20: perspectives on automated image segmentation for radiotherapy. Medical Physics, 41(5), 50902. <https://doi.org/10.1118/1.4871620>

Siciarz, P., & McCurdy, B. (2022). U-net architecture with embedded Inception-ResNet-v2 image encoding modules for automatic segmentation of organs-at-risk in head and neck cancer radiation therapy based on computed tomography scans. Physics in Medicine & Biology, 67(11), 115007. <https://doi.org/10.1088/1361-6560/ac530e>

Wee, L., & Dekker, A. (2019). Data from HEAD-NECK-RADIOMICS-HN1 [Data set]. <https://doi.org/10.7937/tcia.2019.8kap372n>

Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., & Alexander, A. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiotherapy and Oncology, 144, 152–158. <https://doi.org/10.1016/j.radonc.2019.10.019>