Bern University of Applied Sciences

UNIVERSITÄT BERN

Faculty of Medicine
Biomedical Engineering

Master of Science Thesis

Title: Development of a Deep Learning-Based Model for the Detection of Extracapsular Extensions in Head and Neck

by

Léandre Cuenot

of Switzerland

Supervisors
Prof. Dr. Mauricio Reyes and Dr. Daniel Schanne

Institutions
ARTORG Center for Biomedical Research, University of Bern, Medical Image Analysis Group (MIA)

Examiners
Prof. Dr. Mauricio Reyes and Dr. Daniel Schanne

Bern, October 2024

This report is confidential. (Delete this statement in the file "unibe-msc.cls" if the report is not confidential.)

Abstract

The abstract should provide a concise (300-400 word) summary of the motivation, methodology, main results and conclusions. For example:

Osteoporosis is a disease in which the density and quality of bone are reduced. As the bones become more porous and fragile, the risk of fracture is greatly increased. The loss of bone occurs progressively, often there are no symptoms until the first fracture occurs. Nowadays as many women are dying from osteoporosis as from breast cancer. Moreover it has been estimated that yearly costs arising from osteoporotic fractures alone in Europe worth 30 billion Euros.

Percutaneous vertebroplasty is the injection of bone cement into the vertebral body in order to relieve pain and stabilize fractured and/or osteoporotic vertebrae with immediate improvement of the symptoms. Treatment risks and complications include those related to needle placement, infection, bleeding and cement extravazation. The cement can leak into extraosseous tissues, including the epidural or paravertebral venous system eventually ending in pulmonary embolism and death.

The aim of this project was to develop a computational model to simulate the flow of two immiscible fluids through porous trabecular bone in order to predict the three-dimensional spreading patterns developing from the cement injection and minimize the risk of cement extravazation while maximizing the mechanical effect. The computational model estimates region specific porosity and anisotropic permeability from Hounsfield unit values obtained from patient-specific clinical computer tomography data sets. The creeping flow through the porous matrix is governed by a modified version of Darcy's Law, an empirical relation of the pressure gradient to the flow velocity with consideration of the complex rheological properties of bone cement.

To simulate the immiscible two phase fluid flow, i.e. the displacement of a biofluid by a biomaterial, a fluid interface tracking algorithm with mixed boundary representation has been developed. The nonlinear partial differential equation arising from the problem was numerically implemented into the open-source Finite Element framework *libMesh*. The algorithm design allows the incorporation of the developed methods into a larger simulation of vertebral bone augmentation for pre-surgical planning.

First simulation trials showed close agreement with the findings from relevant literature. The computational model demonstrated efficiency and numerical stability. The future model development may incorporate the morphology of the region specific trabecular bone structure improving the models' accuracy or the prediction of the orientation and alignment of fiber-reinforced bone cements in order to increase fracture-resistance.

Acknowledgements

Here you may include acknowledgements.

1. Please sign the following declaration if you did **not** use AI tools (like ChatGPT or DeepL)

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Ich erkläre weiter, dass ich keine unerlaubten Hilfsmittel verwendet habe, namentlich keine weiteren Personen mir beim Verfassen der Arbeit geholfen haben und ich keine Technologien der Künstlichen Intelligenz eingesetzt habe. Mir ist bekannt, dass andernfalls die Arbeit mit der Note 1 bewertet wird bzw. der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

1. Please sign the following declaration if you **did** use AI tools (like ChatGPT or DeepL). The use of AI tools must be permitted by your supervisor.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Als Hilfsmittel habe ich Künstliche Intelligenz verwendet. Sämtliche Elemente, die ich von einer Künstlichen Intelligenz übernommen habe, werden als solche deklariert und es finden sich die genaue Bezeichnung der verwendeten Technologie sowie die Angabe der «Prompts», die ich dafür eingesetzt habe. Mir ist bekannt, dass andernfalls die Arbeit mit der Note 1 bewertet wird bzw. der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Bern, October 27$^{\text{th}}$ 2024

Léandre Cuenot

Contents

Chapter 1

Introduction

Extracapsular extension (ECE) in head and neck cancers refers to the spread of metastatic tumors beyond the lymph node capsule into the surrounding connective tissue. ECE is associated with a poorer prognosis, significantly affecting treatment strategies and reducing overall survival rates in affected patients [4][7]. Accurate early detection of ECE is crucial for optimizing patient management and treatment planning. Proper identification of ECE can guide therapeutic decisions, including the potential benefits of adjunctive treatments such as postoperative concurrent chemoradiotherapy, which may improve outcomes for patients at higher risk due to ECE [6].

Currently, definitive diagnosis of ECE can only be confirmed through postoperative pathology, limiting its clinical utility [2]. In practice, contrast-enhanced computed tomography (CT) is used to detect ECE, relying heavily on physician expertise. Literature indicates that the sensitivity of CT for detecting ECE ranges from 18.8% to 72.2%, with higher sensitivity in more advanced cases. Sensitivity increases from 18.8% in early-stage (grade 1-2) ECE to 72.2% in advanced-stage (grade 4) ECE [8]. This wide range reflects considerable inter-observer variability.

Recent advancements in deep learning have shown significant potential in improving ECE detection compared to manual expertise [2][10][3]. Deep learning algorithms offer more consistent assessments and address the high inter-observer variability observed among human experts. Notably, the HECKTOR challenge, held annually from 2020 to 2022, focused on enhancing segmentation tasks through deep learning. The challenge aimed to develop models for segmenting the primary gross tumor volume (GTVp) and metastatic lymph nodes (GTVn) in the head and neck region. The results were promising, with a substantial majority of participants achieving an aggregate Dice similarity coefficient greater than 0.70 for both GTVp and GTVn, and the top-performing model achieving a Dice score of 0.788, underscoring the efficacy of deep learning approaches in complex segmentation tasks.

In this study, we propose an approach inspired by participants of the HECKTOR challenge, utilizing the same dataset, which integrates positron emission tomography (PET) imaging alongside conventional computed tomography (CT) imaging. PET imaging provides metabolic insights that complement the anatomical information from CT, potentially improving the detection of extracapsular extension (ECE) in head and neck cancers. By leveraging this multimodal approach, we aim to enhance diagnostic accuracy and provide more reliable predictions of ECE, thereby supporting more informed clinical decision-making.

The majority of the top-performing models in this domain have employed ensembles of 3D U-Net architectures. In this work, we aim to assess the robustness of 3D U-Nets under various perturbations that may occur in real-world clinical scenarios. Specifically, we will analyze how different tumor characteristics correlate with the degree to which these perturbations affect segmentation performance.

Additionally, we will link these findings to a relevant clinical application by comparing the model's performance with expert physicians' evaluations. This comparison will explore the correlation between the Dice similarity coefficient, which quantitatively assesses the segmentation accuracy of the deep learning models, and the qualitative assessment provided by physicians. Existing literature has demonstrated a moderate correlation between the Dice similarity coefficient and physician evaluations, with values ranging from 0.36 to 0.5 depending on the anatomical location of the segmented area [5].

Moreover, physicians will also evaluate cases with artificially introduced perturbations to examine whether changes in Dice scores align with variations in their clinical grading. This analysis will provide insights into the relationship between perturbation-induced segmentation errors and the clinical evaluation of ECE.

The primary focus of this study is on the segmentation aspect of the ECE prediction baseline. To enhance the overall predictive model, future work will extend the analysis to the classification component, aiming to improve the model's ability to predict ECE with greater accuracy.

Chapter 2

Methods

2.1 Hecktor dataset

For the segmentation model in this project, we utilized the dataset from the Hecktor Challenge 2022. This dataset consists of both training images with corresponding ground truth labels and test images without labels. For our purposes, we focused solely on the training subset due to the availability of ground truth annotations. This subset comprises 524 imaging cases collected from seven distinct clinical centers, including:

- CHUM: Centre Hospitalier de l'Université de Montréal, Montréal, Canada
- CHUP: Centre Hospitalier Universitaire de Poitiers, France
- CHUS: Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Canada
- CHUV: Centre Hospitalier Universitaire Vaudois, Switzerland
- HGJ: Hôpital Général Juif, Montréal, Canada
- HMR: Hôpital Maisonneuve-Rosemont, Montréal, Canada
- MDA: MD Anderson Cancer Center, Houston, Texas, USA

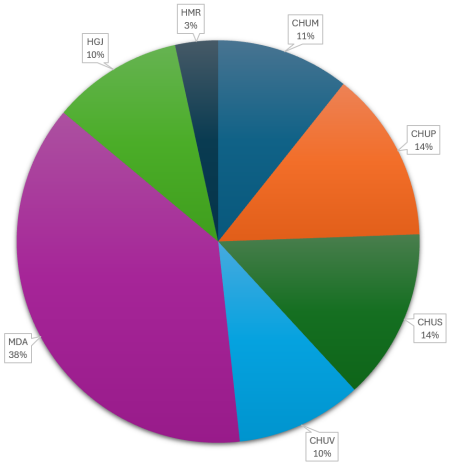
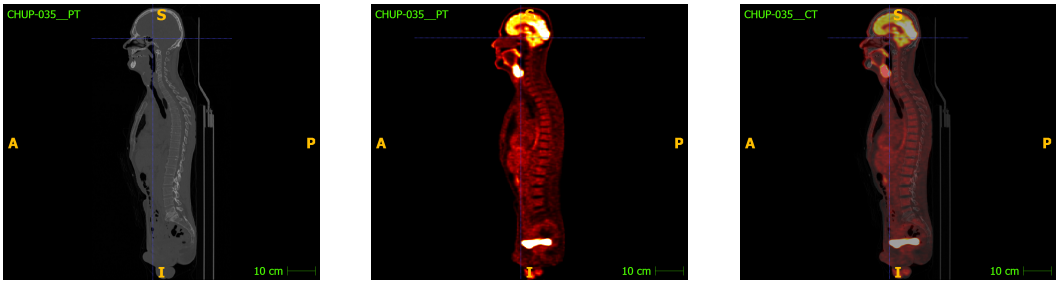


Figure 2.1. Distribution of cases across sites

Each case in the dataset contains two imaging modalities: computed tomography (CT), positron emission tomography (PET), and ground truth labels. The PET images were standardized using the Standardized Uptake Value (SUV). The CT and label images are provided in NIfTI format with a spatial resolution of 524 × 524 pixels in the axial plane, and variable depths across slices. While some CT images focus exclusively on the head and neck region, others encompass the entire body. In contrast, the PET images have a resolution of 128 × 128 pixels in the axial plane, with varying depths similar to the CT images.



(a) CT Scan **(b)** PET Scan **(c)** CT + PET and label

Figure 2.2. An example cases of the CHUP

2.2 Segmentation model

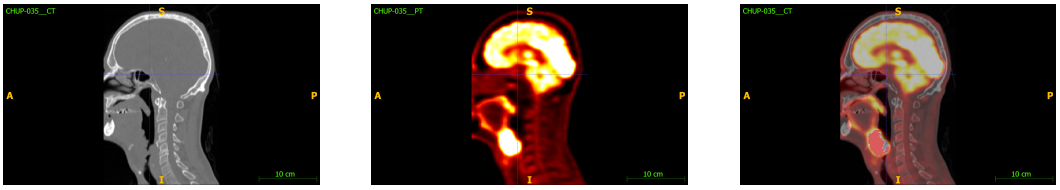
2.2.1 Preprocessing

The initial preprocessing step involved resampling the PET images to achieve uniform dimensions across modalities. Specifically, PET images were resampled to an axial plane resolution of 524×524 pixels. Some labels exhibited minor dimensional discrepancies, occasionally differing by one plane in either width or height. These discrepancies were rectified following the resampling process.

Subsequently, after ensuring that all three imaging modalities (CT, PET, and label) were aligned to the same dimensions, the images were resampled to a common isotropic voxel size of 1×1×1 mm. This resampling was performed to facilitate subsequent cropping of the head and neck region.

For the cropping procedure, the center of the head was identified using the contours of the brain on the PET scan. Based on this central reference point, a subvolume of 200×200×[maximum of 310 pixels] was extracted. This approach minimizes the computational burden associated with background regions devoid of ground truth information. Additionally, the images underwent normalization via z-score clipping to mitigate the effects of outliers.

The processed images were saved in NIfTI format following the above steps. Further preprocessing techniques were applied during the training phase, which will be detailed in subsequent sections. Aspects of the preprocessing procedure were adapted from the methods employed by the winning team of the Hecktor Challenge 2022 [7]



(a) CT Scan **(b)** PET Scan **(c)** CT + PET and label

Figure 2.3. Same example cases after preprocessing

2.2.2 Data augmentation

During the training process, data augmentation was applied to the original images using the Medical Open Network for Artificial Intelligence (MONAI) framework. Spatial augmentations were applied to both imaging modalities, including random flipping, affine transformations (translation, rotation, and scaling). For the CT images, intensity augmentations were also incorporated, such as the addition of Gaussian noise, smoothing, contrast adjustment, and intensity shifting. All augmentations were applied with an occurence probability of 20%.

Following augmentation, the images were randomly cropped into patches of size 192×192×192 voxels. Patches were centered based on labels, with a 10% probability of being centered on background, 45% on primary tumors, and 45% on nodal tumors. In cases where only one tumor type was present, the sampling probability for that tumor was increased to 90%.

For validation, a single patch of equal size was extracted, with a balanced distribution of 33% for background, primary tumors, and nodal tumors. This approach eliminates the need for predictions using sliding windows (e.g., 8 predictions per image), as a single patch is used for each validation image of size 200×200×310.

2.2.3 Model architectures

The model used for segmentation is the Dynamic UNet (DynUNet) from MONAI. It provide several parameter to configure the model. The architecture is represented below :

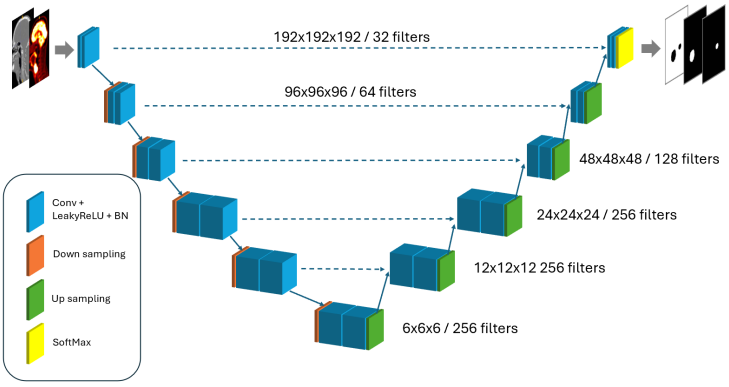


Figure 2.4. Model architecture

The model operates in three dimensions, taking as inputs two modalities: computed tomography (CT) and positron emission tomography (PET). It produces three binary segmentation outputs: background, primary tumors, and nodal tumors.

All convolutional kernels used in the model are of size 3×3×3. The architecture consists of six layers, with downsampling performed consistently by a factor of 2, reducing the input size from 192×192×192 to 6×6×6. Additionally, batch normalization is implemented in conjunction with residual blocks to enhance training stability and model performance. The probabilites are then computed through softmax that incorporate the mutual exclusivity between the 3 segmentation masks.

2.2.4 Training

The training parameters for the model are as follows:

- Optimizer : AdamW
- Learning rate : 1e-4
- Weight decay : 3e-5
- Batch size : 2
- Epochs : 100

A LambdaLR scheduler is employed, utilizing the following decay function:

$$\text{Decay Factor} = \left(1 - \frac{\text{epoch}}{\text{number\_epoch}}\right)^{0.9} \tag{2.1}$$

The loss function combines Dice loss and cross-entropy loss, defined as follows:

$$\text{Dice Loss} = 1 - \frac{2|P \cap Y|}{|P| + |Y|} \tag{2.2}$$

$$\text{Cross-Entropy Loss} = -\left(Y \log(P) + (1 - Y) \log(1 - P)\right) \tag{2.3}$$

where P is the prediction and Y is the ground truth.

Dice loss is particularly advantageous for segmentation tasks involving imbalanced datasets, such as those found in head and neck tumor imaging, where the tumors are often significantly smaller than the surrounding tissues. The Dice coefficient, which Dice loss is based on, focuses on the overlap between the predicted segmentation and the ground truth (2.2), effectively capturing the regions of interest despite their limited size.

In contrast, cross-entropy loss measures the dissimilarity between the predicted probability distribution and the true distribution (2.3). While it is easier to optimize due to its differentiable nature, cross-entropy can be more sensitive to class imbalances. This sensitivity may lead to suboptimal performance when the class of interest (e.g., tumors) constitutes only a small fraction of the total data.

Given these characteristics, a hybrid approach that combines both Dice loss and cross-entropy loss offers a promising solution. This combination leverages the strengths of both methods: Dice loss ensures that the model pays adequate attention to the small tumor regions, while cross-entropy facilitates stable and efficient optimization. Such an approach can yield improved segmentation performance, making it particularly suitable for complex tasks in medical imaging.

To enhance training performance, mixed precision training utilizing PyTorch has been implemented. This approach leverages both single-precision (32-bit) and half-precision (16-bit) floating-point formats during model training. By employing mixed precision, the training speed is significantly increased due to reduced computational overhead, while the memory footprint is minimized. This allows for the efficient use of hardware resources, facilitating larger batch sizes and improved model scalability without compromising numerical stability. The training run on a TODO with around 30Go of memory use.

Prior to training, additional preprocessing steps were applied to the images. Specifically, the CT images were min-max normalized to a range of 0 to 1, ensuring consistent scaling across the dataset. In contrast, PET images were normalized to have a zero mean, adjusting for intensity variations while preserving the relative distribution of values.

These normalization procedures were performed dynamically during the training phase, allowing the images to be stored in their original format in NIfTI files. This preserved the integrity of the original data, enabling direct computation of properties from the segmented regions in subsequent analyses.

To enhance model performance, the training dataset has been partitioned into five folds representing 90% (with 80% for training and 20% for validation) of the initial 524 cases, facilitating a cross-validation approach. This technique allows for the integration of predictions from each fold, thereby improving the consistency and robustness of the model's predictions. By leveraging multiple subsets of the data during training, the model can better generalize to unseen data, ultimately leading to more reliable and stable results. This method not only mitigates overfitting but also enhances the overall accuracy of the predictions.

For the validation process, as outlined in the data augmentation section, a single patch was selected, ensuring equal distribution across the channels. This approach generates a validation set that is representative of the entire image while significantly reducing computational time. The validation set comprised 20

Following model inference, a postprocessing step was applied to the softmax output. This involved converting the probabilistic output into a binary mask by applying a threshold of 0.5, where values greater than or equal to 0.5 were classified as positive for the presence of a tumor, and values below this threshold were classified as negative.

2.2.5 Inference

During the inference phase, the test set comprised 50 cases, constituting approximately 10% of the original dataset. Each case was subjected to 37 inference passes, which included 6 distinct perturbations applied at 3 different levels of severity for each of the two modalities, as well as one baseline inference without any perturbation. Detailed descriptions of these perturbations and their respective parameters are provided in the Analysis section.

The inference results yielded 37 segmented binary masks for each of the 50 cases, representing the model's predictions under different conditions. These binary masks were used for subsequent analysis to assess the consistency and robustness of the model's segmentation performance across different perturbation scenarios.

2.3 Analysis

2.3.1 Robustness

As outlined in the Inference section, six perturbations were applied to the data to assess the robustness of the model against variations that may be encountered in real-world scenarios, such as sensor noise or image artifacts. These perturbations, derived from the torchio library, simulate various types of distortions and alterations that may impact model performance. The six perturbations are described below :

- **Noise**: Random variations in pixel intensity, simulating the effect of sensor or environmental noise during image acquisition.
- **Motion**: Simulates patient movement during the scan, leading to blurring or streaking in the images.
- **Blur**: Reduces image sharpness, simulating out-of-focus images or lower resolution captures.
- **Spike**: Alternating bright and dark lines in the image, simulating sensor errors or environmental interference during the scan.
- **Bias**: Intensity shifts or gradients across the image, simulating uneven lighting or sensor sensitivity changes.
- **Ghosting**: Artifacts that simulate image echoes or multiple exposures due to acquisition errors, causing repeated structures in the image.

Here is a visual representation of the perturbations applied :

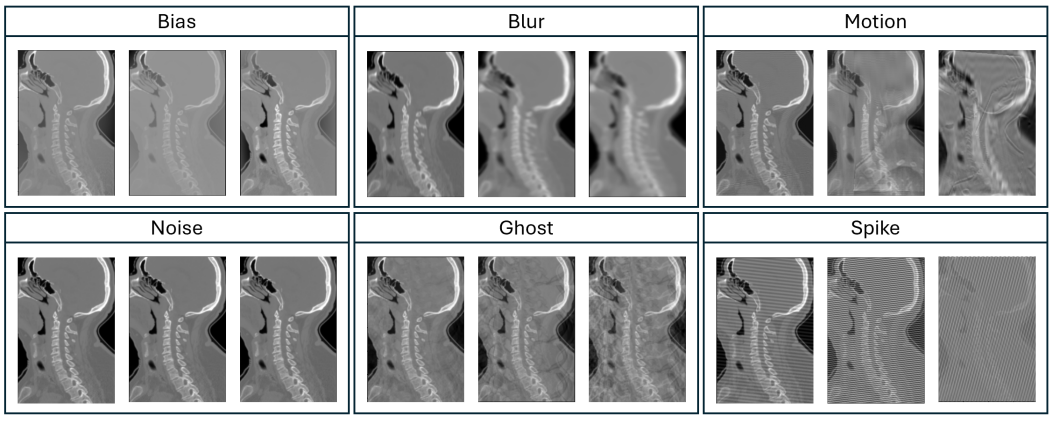


Figure 2.5. Perturbations

Each perturbation was applied independently to either the CT or PET modality at three varying degrees of severity: low, medium, and high. These severity levels correspond to increasing intensities of the perturbation, where low severity represents minimal distortion and high severity introduces significant alterations to the image. The figure above illustrates the perturbations at each severity level for the CT modality.

Note that the noise perturbations may not be accurately represented here; however, more detailed visualizations and discussion of the noise effects on the images are provided in the Results section.

To assess the impact of perturbations on segmentation performance, we utilized evaluation metrics from the Pymia library. The metrics employed are as follows:

- **Dice coefficient**: A similarity index that measures the overlap between predicted and ground truth segmentation.
- **Hausdorff distance**: A metric that evaluates the maximum boundary distance between predicted and ground truth segmentation contours.
- **Jaccard index**: A measure of overlap between predicted and true segmentation, similar to the Dice coefficient but more sensitive to boundary errors.
- **Sensitivity**: Measures the proportion of true positives correctly identified by the model.
- **Specificity**: Measures the proportion of true negatives correctly identified by the model.
- **Accuracy**: Evaluates the overall performance of the model by considering both true positives and true negatives.

To analyze the impact of each perturbation, the delta for each evaluation metric was computed by calculating the difference between the baseline (unperturbed) metric and the corresponding metric obtained after applying the perturbation. This delta value quantifies the performance change induced by each perturbation, providing insight into how robust the model is under varying conditions.

For each of the six perturbations, delta values were computed for all three severity levels, separately for both the CT and PET modalities. Boxplots were generated to visualize the distribution of these delta values across all test cases. The boxplots represent the central tendency (median) and variability (interquartile range) of the delta values for each metric, with whiskers extending to the most extreme non-outlier values. Outliers were identified to highlight cases where perturbations had an unusually large effect on model performance.

This analysis enables a detailed comparison of the model's robustness to different perturbations and degrees of severity for both imaging modalities. Further interpretation of these results is discussed in the Results section.

In addition to the computation of the previously described metrics, we computed properties of the segmented area to investigate potential correlations between the change in Dice coefficient induced by perturbations and these properties. For example, we examined whether the volume of the segmented area correlates with greater changes in perturbed images due to motion. The properties are as follows:

- **Volume**: Total number of voxels belonging to the region of interest, computed as the sum of all 'True' values in the ground truth boolean mask.
- **Surface Area**: The area of the surface boundary of the region, computed using compactness and surface area algorithms on the labeled region.
- **Compactness**: A shape descriptor that relates the surface area to the volume, indicating how compact or spread out the region is. (Maximum value of 1 corresponds to a perfect sphere.)
- **Distance from Center**: Euclidean distance from the center of mass (centroid) of the region to the center of the entire image.
- **Boundary Length**: Length of the boundary of the region, calculated using the Sobel operator on the ground truth mask.
- **CT Intensity Variability**: The standard deviation of the intensity values in the CT image for the region of interest, measuring intensity variation within the region.
- **PET Intensity Variability**: The standard deviation of intensity values in the PET image for the region of interest, representing variability in PET signal.
- **CT F/B Contrast**: Contrast between the mean intensity in the foreground (region of interest) and the background in the CT image, calculated as the difference between the mean intensities.
- **PET F/B Contrast**: Contrast between the mean intensity in the foreground and background in the PET image, computed similarly to the CT contrast.
- **SUVmax**: The maximum standardized uptake value (SUV) in the PET image within the region, indicating the highest metabolic activity.
- **CT Number (HU)**: The mean Hounsfield unit (HU) value within the region of interest in the CT image.
- **PET Number**: The mean standardized uptake value (SUV) within the region of interest in the PET image.
- **Regions**: The number of regions in the labeled ground truth, representing distinct anatomical regions.
- **Entropy**: A measure of the randomness or complexity of the region, reflecting the degree of heterogeneity in the image data.

Scatter plots of each perturbation at varying degrees were generated against each of the properties listed above. The goal is to determine whether perturbations have a more significant effect on specific cases, such as certain anatomical regions or volume sizes. Statistical correlations were computed using Pearson, Spearman and Kendall methods to assess these relationships.

2.3.2 Clinical evaluation

To assess the relevance of geometrical metrics used in deep learning, we conducted a clinical evaluation of 50 test cases in collaboration with a physician specializing in head and neck radiation oncology. The evaluation utilized a 5-point Likert scale for grading. The representations of the 5-point grades are provided below:

- ☆: Unable to utilize the prediction; necessitates complete reconstruction.
- ☆☆: Able to utilize the prediction but requires significant modifications.
- ☆☆☆: Able to utilize the prediction with some modifications.
- ☆☆☆☆: Able to utilize the prediction with minor modifications.
- ☆☆☆☆☆: Able to utilize the prediction in its original form.

The objective of this evaluation is to provide a comprehensive assessment of the model's performance and to quantify the correlation between the geometrical metrics and physician expertise. The evaluation was conducted specifically for each label, including Primary and Nodal tumors. Subsequently, scatter plots were generated to visualize the relationships between the various geometrical metrics, including the Dice score, Hausdorff distance, and Jaccard index. The correlations between these metrics and the physician evaluations were computed using Pearson correlation coefficients.

The 50 cases were categorized into two groups: one consisting of baseline images (without perturbations), representing 32 images, and the other comprising perturbed images influenced by the three most significant perturbations. For each perturbation and modality, three cases were selected, resulting in a total of six cases for each of the three perturbations, accounting for the final 18 cases.

This approach allows us to analyze the impact of perturbations on real-case evaluations and to assess whether there is a correlation with the effects represented by the delta Dice score.

This analysis, enables us to evaluate the relevance of the geometrical metrics in predicting clinical outcomes. The findings will contribute to understanding the robustness of the model, its practical applicability in clinical settings, and the significance of geometrical metrics in enhancing model performance and decision-making in radiotherapy. Limitations of this study include potential biases arising from the subjective nature of physician evaluations, as the assessment was conducted by a single observer.

Chapter 3

Results

3.1 Robustness

3.1.1 Perturbations effects

The following graphs illustrates the impact of perturbations on segmentation performance. The columns correspond to the baseline and the six different perturbations (2.3.1). The rows depict the delta values of various segmentation metrics (2.3.1), indicating the extent of variation in segmentation performance due to the perturbations. The baseline, having no delta, represents the original values for each metric. Each of the three boxplots corresponds to one of three levels of perturbation severity, providing a visual representation of how segmentation performance is affected at increasing degrees of perturbation.

The graph below represents the scenario in which perturbations are applied to nodal tumor labels in the CT modality.

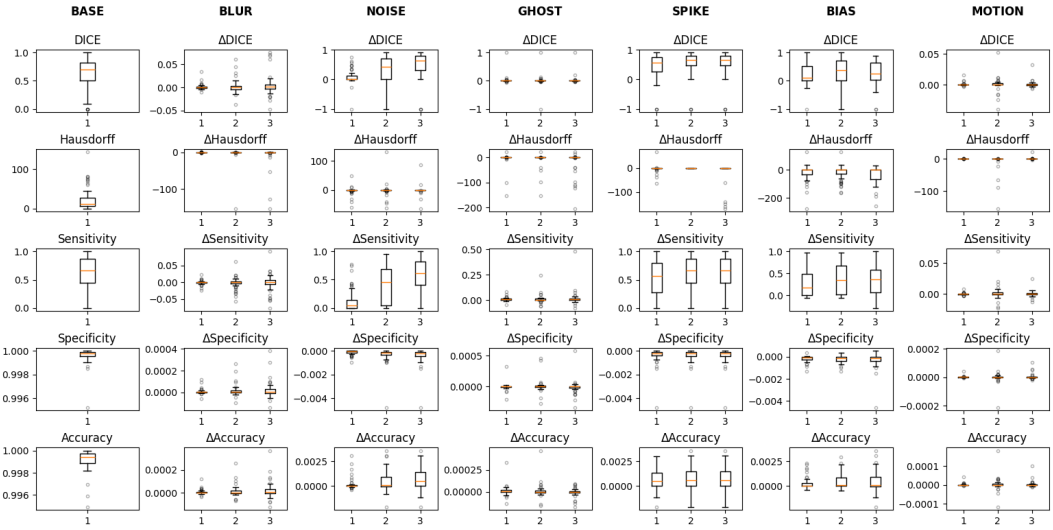


Figure 3.1. Perturbations applied to nodal tumor labels in the CT modality

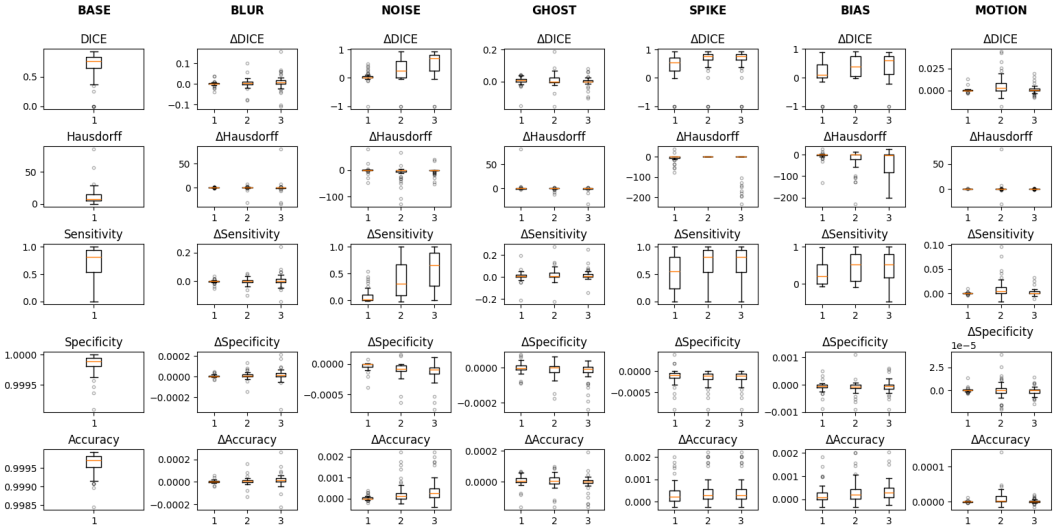


Figure 3.2. Perturbations applied to primary tumor labels in the CT modality

The graph illustrates significant variations in the Dice coefficient under spike, bias, and noise perturbations, underscoring their substantial impact on segmentation performance across both labels. Notably, the Hausdorff distance demonstrates that bias perturbation exerts the most pronounced negative effect, with the boxplot range extending up to 200. Sensitivity, specificity, and accuracy metrics exhibit trends consistent with the Dice coefficient, revealing markedly larger effects under spike, bias, and noise perturbations. In contrast, the model demonstrates relative robustness against motion, blur, and ghost perturbations, as indicated by minimal performance variability across all metrics.

Moreover, noise perturbation exhibits a less pronounced effect at the initial degree of severity compared to higher levels, while spike and bias perturbations show significant impact even at the first degree.

Interestingly, in certain cases, the Dice coefficient delta is negative, indicating an improvement in segmentation performance under perturbation.This behavior suggests that certain perturbations may occasionally enhance segmentation accuracy, depending on the model and context.

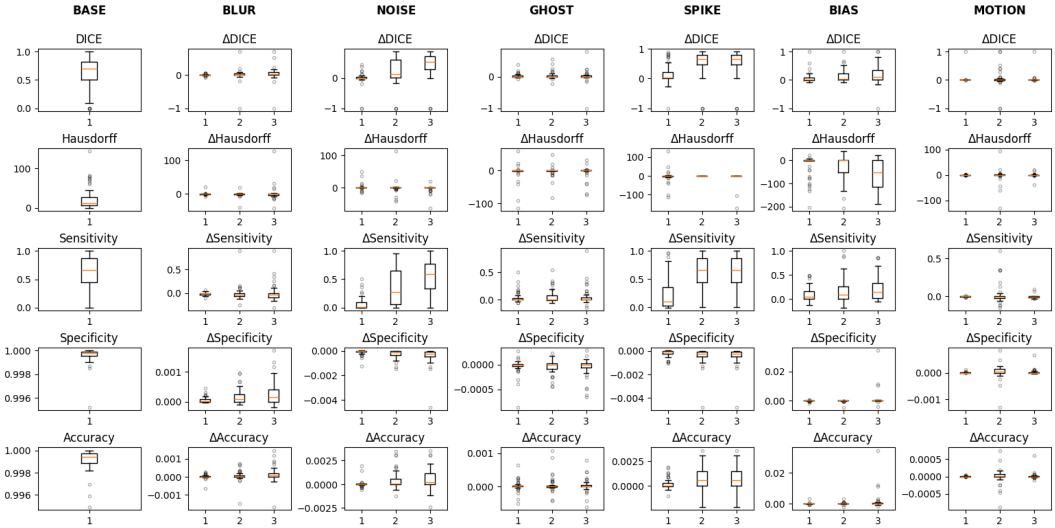


Figure 3.3. Perturbations applied to primary tumor labels in the PET modality

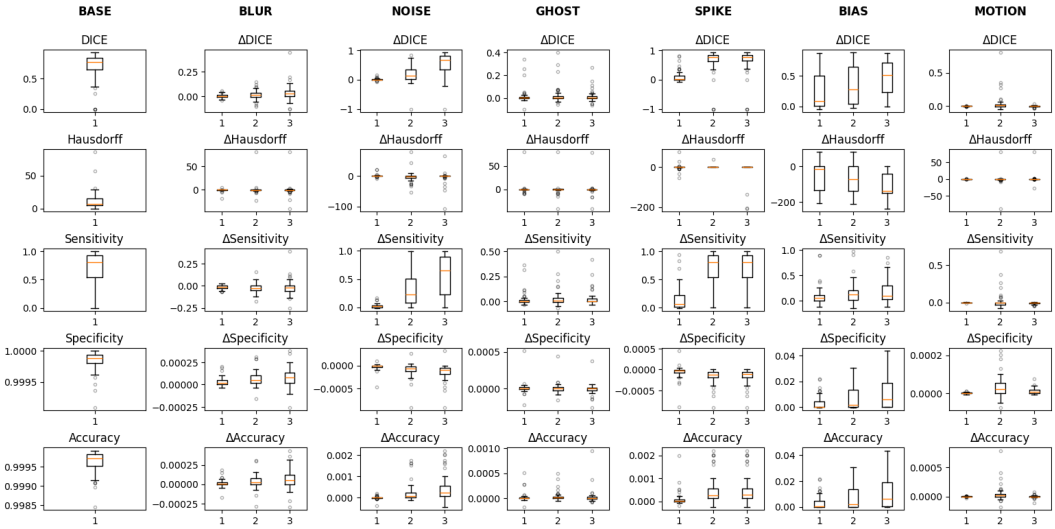


Figure 3.4. Perturbations applied to primary tumor labels in the PET modality

The two graphs demonstrate the effect of perturbations on PET images across both labels. Similar to the behavior observed in CT images, spike, bias, and noise perturbations have a pronounced impact, while blur, ghost, and motion perturbations exhibit a more robust and stable performance. However, in the case of PET images, the first degree of noise and spike perturbations shows less variation in the Dice coefficient, sensitivity, specificity, and accuracy, indicating greater robustness compared to CT images. In contrast, bias perturbation continues to exert a substantial effect even at the first degree of severity, particularly in the Hausdorff distance, where a significant impact is still observed.

3.1.2 Correlation with properties

The following graphs present the correlation between the 13 previously discussed properties of the segmentation area (ref). Similar to the earlier graph, the figures are organized into 7 columns, with the baseline in the first column and the 6 perturbations in the subsequent columns. The baseline scatter plot illustrates the correlation between the original Dice coefficient and the segmentation properties, establishing a reference for the unperturbed performance. The columns corresponding to each perturbation display the correlation between the delta Dice (change in Dice coefficient) and the segmentation properties, highlighting whether significant segmentation degradation is associated with specific properties.

Each scatter plot includes data for all three degrees of perturbation, represented by distinct markers, allowing for visual differentiation of the effects of increasing perturbation severity. Also, the pearson correlation coefficient is visible for each degree. The graphs presented here represent cases where perturbations have been applied to the nodal label in the CT modality. The remaining three category (Primary in CT, Nodal and Primary in PET) will be provided in the appendix. A more comprehensive correlation graph, encompassing all four categories, will follow to offer a broader view of the relationships between the segmentation properties and the applied perturbations.

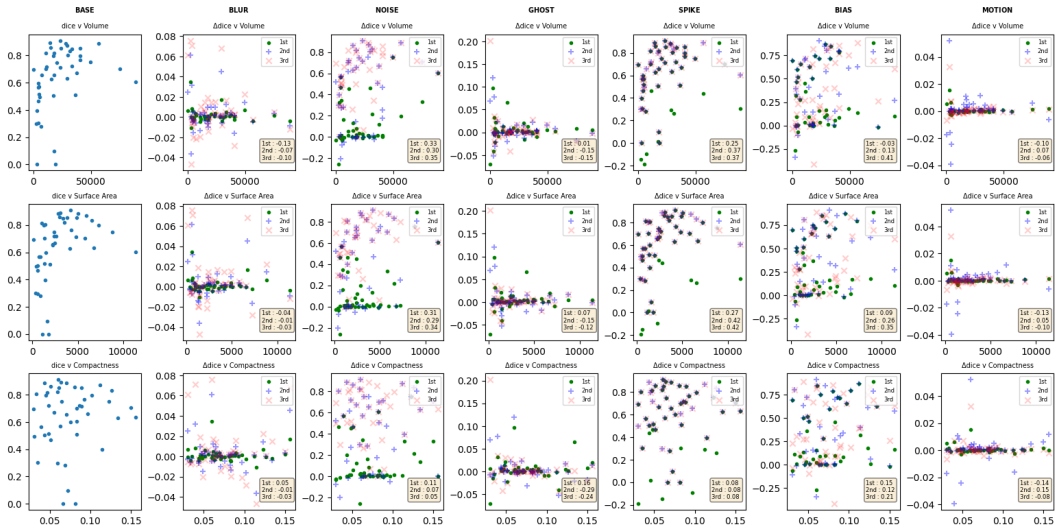


Figure 3.5. Correlation with perturbations applied to nodal tumor labels in the CT modality

On the left, the Dice coefficient exhibits a crescent-shaped distribution rather than a clear linear correlation for the three properties related to tumor size and shape. The correlation is consistently higher for perturbations that exert a greater impact, as anticipated. In contrast, blur, motion, and ghost perturbations present a significantly flatter scatter plot with considerably smaller variations, reflecting the robustness observed in earlier analyses.

Among the perturbations, spike and noise demonstrate the strongest correlation with volume and surface area, with correlation coefficients ranging from 0.30 to 0.42, the highest being associated with surface area. In the case of bias, the correlation coefficient increases across the degrees of perturbation, showing a very low correlation at the first degree. By the third degree, the correlation reaches levels comparable to those of other perturbations. Notably, bias exhibits the highest correlation with compactness, although this correlation is not statistically significant, peaking at 0.21.

Interestingly, ghost perturbation reveals a prominent negative correlation, suggesting an improvement in segmentation performance, with the strongest correlation observed with the compactness property.

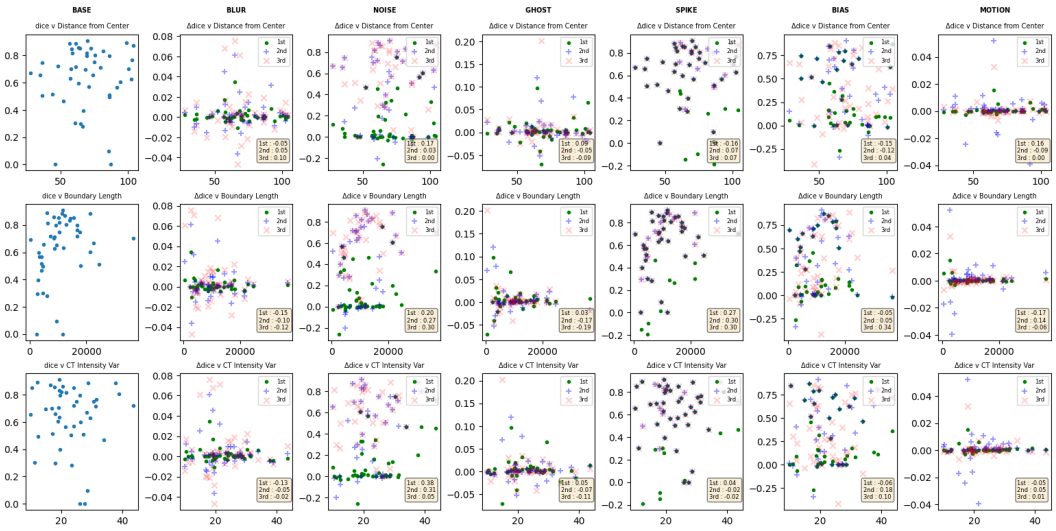


Figure 3.6. Perturbations applied to primary tumor labels in the CT modality

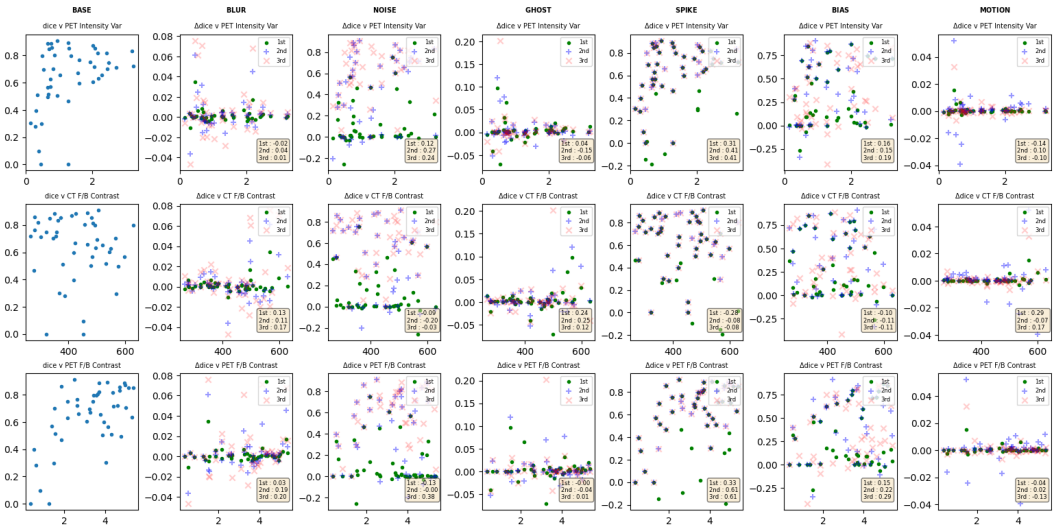


Figure 3.7. Perturbations applied to primary tumor labels in the CT modality

The distance from the center did not demonstrate a significant correlation with any perturbation, with the highest coefficient reaching only 0.17. In contrast, for the boundary length, spike and noise perturbations exhibited correlation coefficients ranging from 0.20 to 0.30, while bias showed a correlation of 0.34 at the third degree of perturbation.

Regarding intensity properties, in CT variations, the only perturbation that displayed a significant correlation was noise, with coefficients of 0.38 and 0.31, but only at the first

two degrees of severity. For PET variations, the correlation with noise was lower; however, spike demonstrated a stronger correlation, achieving a score of up to 0.41.

In terms of contrast, PET generally exhibited greater correlation, particularly with spike, which reached a maximum of 0.61. The third degree of noise showed a correlation of 0.38, while bias presented a more moderate correlation, peaking at 0.29.

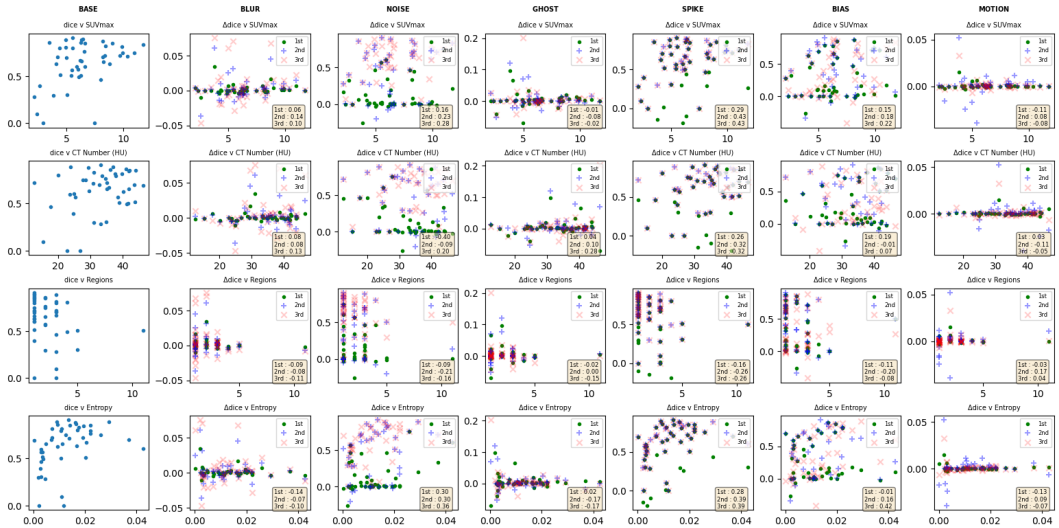


Figure 3.8. Perturbations applied to primary tumor labels in the CT modality

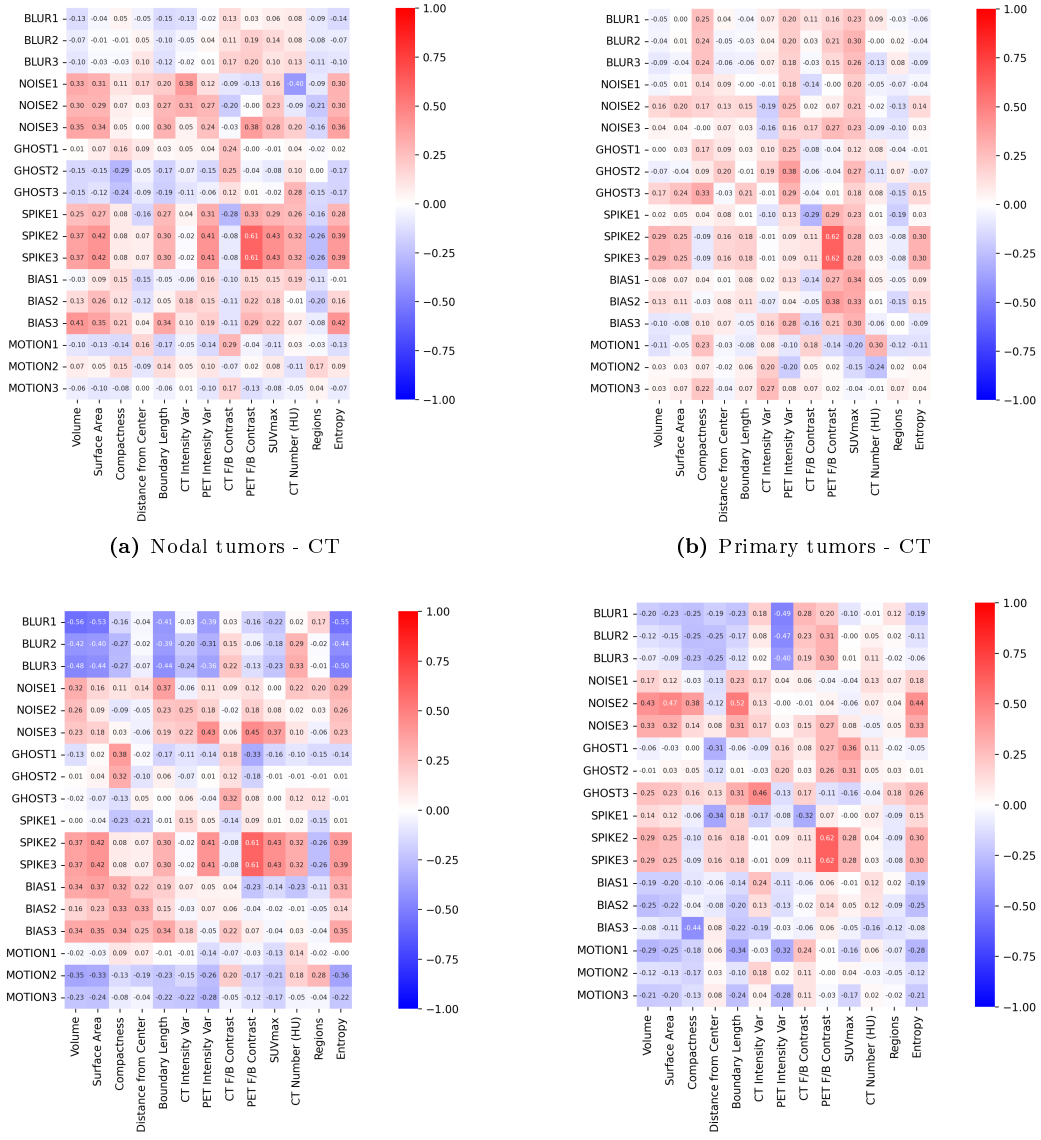
For the last four properties analyzed, the SUVmax demonstrated a moderate correlation with spike perturbations, ranging from 0.29 to 0.43 across the degrees of severity. Noise and bias also exhibited correlations, albeit at more moderate levels, around 0.20. In the case of CT number, spike perturbations again displayed the strongest correlation, peaking at 0.32, while a surprising correlation of 0.28 was observed with ghost perturbations at the third degree. Notably, noise exhibited irregular behavior, with correlation coefficients fluctuating between -0.40 and 0.20.

The regions showed inconsistent behavior across properties, except for spike, which revealed a stable negative correlation ranging from -0.16 to -0.26. As expected, entropy demonstrated stronger correlations with effective perturbations, particularly spike, which reached values of up to 0.39, and bias at the third degree, which peaked at 0.42. Noise showed a more consistent behavior across degrees, with correlation coefficients ranging from 0.30 to 0.36.

Overall, among the 13 properties examined, the strongest correlations were observed with perturbations that most significantly affected segmentation performance. In contrast, blur, motion, and ghost perturbations exhibited only minor correlations due to their limited variability. For spike, bias, and noise, the correlation coefficients generally clustered around 0.30 to 0.40, with occasional peaks reaching up to 0.60.

In the following section, a more comprehensive correlation graphic will present the Pearson correlation coefficients of the various properties across each perturbation and degree, as well as for the four categories of images. This analysis will encompass both labels and modalities, providing a detailed overview of the relationships between segmentation properties and the applied perturbations.

In those graphics, the y axis represent the perturbations with each time the 3 degrees of severity and on x axis the 13 properties analysed. The correlation color scale is represented on the right.



(a) Nodal tumors - CT

(b) Primary tumors - CT

(c) Nodal tumors - PET

(d) Primary tumors - PET

Figure 3.9. Correlation graph for CT and PET modalities

Starting with the nodal label in the CT modality, we observe the same general trends identified in the previous section. The highest correlations are consistently seen with spike, bias, and noise perturbations, with most properties showing strong correlations with these perturbations. Notable exceptions include compactness, distance from center, CT variations, CT contrast, and regions, which do not exhibit significant correlations.

Spike emerges as the perturbation with the highest correlation scores, followed by noise and then bias. In contrast, ghost perturbations generally show an improvement in segmentation performance across most properties, with the strongest negative correlation observed for compactness. The only exception is CT contrast, which displays a positive correlation with ghost perturbations.

Blur and motion perturbations exhibit similar behavior, showing low correlation coefficients, consistent with previous analyses. Properties such as volume, surface area, boundary length, PET variations, SUVmax, and entropy display comparable responses to the perturbations. These properties generally show moderate positive correlations with spike, bias, and noise, while ghost perturbations produce negative correlations. Motion and blur have a minimal effect, with low correlation coefficients.

Interestingly, an increase in the number of regions appears to improve segmentation performance across all perturbations.

3.2 Clinical evaluation

Chapter 4

Discussion and Conclusions

4.1 Discussion

Interpret your results in the context of past and current studies and literature on the same topic. Attempt to explain inconsistencies or contrasting opinion. Highlight the novelty of your work. Objectively discuss the limitations.

4.2 Conclusions

Formulate clear conclusions which are supported by your research results.

Chapter 5

Outlook

Provide a vision of possible future work to continue and extend your thesis research.

Bibliography

[1] V. Andrearczyk, V. Oreiller, and et al. Overview of the hecktor challenge at miccai 2022: Automatic head and neck tumor segmentation and outcome prediction in pet/ct. Lecture Notes in Computer Science, 13626(2):1–30, March 2023.

[2] B. H. Kann, S. Aneja, G. V. Loganadane, J. R. Kelly, S. M. Smith, R. H. Decker, J. B. Yu, H. S. Park, W. G. Yarbrough, A. Malhotra, B. A. Burtness, and Z. A. Husain. Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. Scientific Reports, 8(5):14036, Sept. 2018.

[3] B. H. Kann and J. L. et al. Screening for extranodal extension in hpv-associated oropharyngeal carcinoma: evaluation of a ct-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. Lancet Digital Health, 5(8):e360–e369, 2023.

[4] H. R. Kelly and H. D. Curtin. Squamous cell carcinoma of the head and neck: Imaging evaluation of regional lymph nodes and implications for management. Seminars in Ultrasound, CT, and MR, 38(1):466–478, Oct. 2017.

[5] F. Kofler, I. Ezhov, F. Isensee, and B. M. et al. Estimates of human expert perception for cnn training beyond rolling the dice coefficient. Melba, 2(9):27–71, May 2023.

[6] V. Krstevska. Evolution of treatment and high-risk features in resectable locally advanced head and neck squamous cell carcinoma with special reference to extracapsular extension of nodal disease. Journal of BUON, 20(6):943–953, July 2015.

[7] A. Myronenko, M. M. R. Siddiquee, D. Yang, Y. He, and D. Xu. Automated head and neck tumor segmentation from 3d pet/ct: Hecktor 2022 challenge report. In Proceedings of the HECKTOR 2022 Challenge, pages 31–37. Lecture Notes in Computer Science (LNCS, Volume 13626), 2023.

[8] R. S. Prabhu and K. R. M. et al. Accuracy of computed tomography for predicting pathologic nodal extracapsular extension in patients with head-and-neck cancer undergoing initial surgical resection. International Journal of Radiation Oncology Biology Physics, 88(3):122–129, Jan. 2014.

[9] T. V. Thomas, M. R. Kanakamedala, E. Bhanat, A. Abraham, E. Mundra, A. A. Albert, S. Giri, R. Bhandari, and S. Vijayakumar. Predictors of extracapsular extension in patients with squamous cell carcinoma of the head and neck and outcome analysis. Cureus, 13(4):e16680, July 2021.

[10] Y. S. e. a. Yoshiko Ariji. Ct evaluation of extranodal extension of cervical lymph node metastases in patients with oral squamous cell carcinoma using deep learning classification. Oral Radiology, 36(7):148–155, Apr. 2020.

Appendices

Appendix A

Vector and Tensor Mathematics

A.1 Introduction

...

A.2 Variable Types

...

Appendix B

Another Appendix

B.1 Section 1

...

B.2 Section 2

...