Bern University of Applied Sciences

Faculty of Medicine
Biomedical Engineering

UNIVERSITÄT BERN

Master of Science Thesis

Title: Development, Robustness Analysis and Clinical Evaluation of a Deep Learning-based Segmentation Model for Head and Neck Cancer Patients

by

Léandre Cuenot

of Switzerland

Supervisors
Prof. Dr. Mauricio Reyes

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

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

Bern, October 2024

Abstract

Accurate detection of extracapsular extension (ECE) is crucial for improving prognosis and guiding treatment decisions in head and neck cancers. Traditional methods relying on physician expertise to interpret contrast-enhanced CT images often lead to significant variability. This study employs a multimodal approach combining FDG-PET and CT imaging, leveraging deep learning to enhance ECE detection, focusing on the segmentation step.

We implemented a 3D U-Net architecture to assess segmentation performance. Our objectives were to evaluate accuracy with the Dice similarity coefficient and compare the model's performance against expert assessments. We utilized the HECKTOR Challenge 2022 dataset, with additional preprocessing and data augmentation. The model generated binary outputs for background, primary tumors, and nodal tumors, employing a hybrid loss function that combined Dice and cross-entropy losses. A five-fold cross-validation strategy enhanced model generalization, and a test set of 50 cases was evaluated under various perturbations to assess robustness.

We evaluated the impact of perturbations on segmentation performance across imaging modalities and tumor labels, focusing on key metrics. The results showed that the model was robust to ghosting, blurring, and motion artifacts, with minimal variation in these metrics even under severe conditions. However, the model demonstrated reduced resilience to noise, particularly spike noise and bias, revealing areas for improvement. These trends were consistent across both modalities and tumor labels.

Clinical evaluations indicated strong inter-observer agreement, with a correlation of approximately 0.65 for both labels in cases without perturbations. Most cases received positive ratings, reflecting a general trend of agreement among observers. Significant correlations between the Dice coefficient and clinical evaluations, especially for primary tumors, suggest that higher Dice scores align with more accurate assessments. Notably, these correlations remained stable or improved under various perturbations, underscoring the relevance of the metrics even in challenging conditions.

Limitations include the absence of MRI data, important for diagnosing primary tumors, and reliance on the Hecktor dataset, which lacks consistent patient-specific details. Clinician disagreement with the ground truth could also introduce bias in the Dice coefficient. Despite these issues, the model performed well with strong metric scores, supporting its clinical relevance. Future work should incorporate additional imaging modalities like MRI, refine data augmentation for improved robustness, and develop classification models to better predict ECE for enhanced diagnostic accuracy and patient care.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Mauricio Reyes, for his invaluable guidance throughout my thesis. His dedication and support have been truly appreciated, and his exceptional leadership created a positive and productive environment within the team. I am also deeply thankful for the opportunities he provided during my research, including the project that formed the basis of this thesis.

My sincere thanks also go to my co-supervisor, Dr. med. Daniel Hendrick Schanne, for enabling this collaborative project and for his dedicated involvement in its progress.

I am further grateful to Prof. Dr. med. Olgun Eliçin for his expert input and engagement in this research.

Finally, I would like to thank the entire MIA team for their continuous support, valuable advice, and kindness, all of which contributed to making this experience truly remarkable.

„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 25$^{\text{th}}$ 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 [5][11]. 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 [7].

Currently, definitive diagnosis of ECE can only be confirmed through postoperative pathology, limiting its clinical utility [3]. 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 [10]. 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 [3][14][4]. Deep learning algorithms offer more consistent assessments and address the high inter-observer variability observed among human experts. Notably, the HECKTOR challenge [1][2], 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 hypothesize that performance levels of a deep learning based segmentation model for head and neck cancer patients is sufficient for clinical use. We propose an approach inspired by participants of the HECKTOR challenge, utilizing the same dataset, which integrates Fluorodeoxyglucose positron emission tomography (FDG-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 [2][12][13][8]. 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 [6].

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

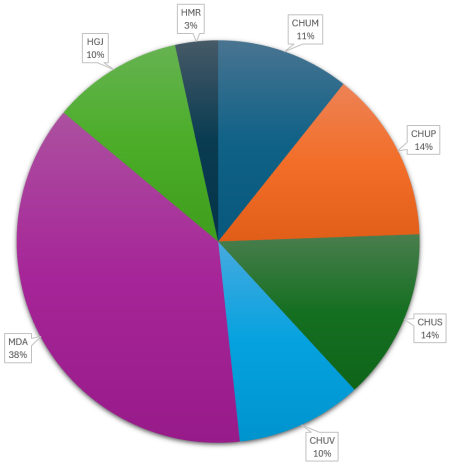
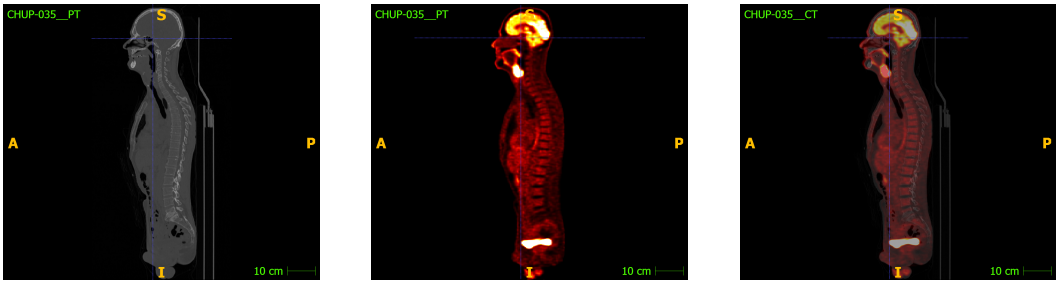


Figure 2.1. Distribution of cases across centers

Each case in the dataset contains two imaging modalities: computed tomography (CT), Fluorodeoxyglucose positron emission tomography (FDG-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 case of the CHUP

2.2 Segmentation model

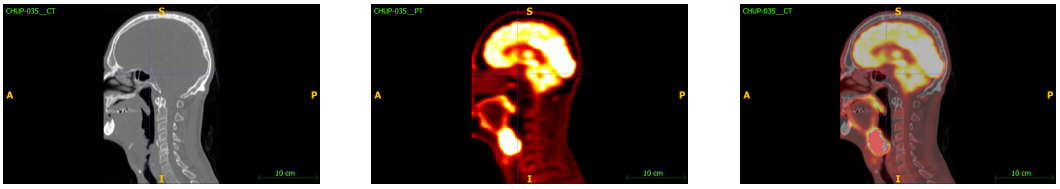
2.2.1 Preprocessing

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 and further data augmentation were adapted from the methods employed by the winning team of the Hecktor Challenge 2022 [9]



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

Figure 2.3. Same example case 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 $192 \times 192 \times 192$ 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 $200 \times 200 \times 310$.

2.2.3 Model architectures

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

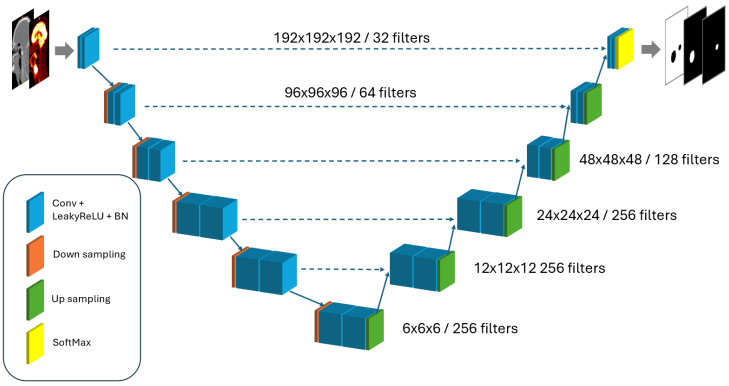


Figure 2.4. Model architecture

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 $3 \times 3 \times 3$. The architecture consists of six layers, with downsampling performed consistently by a factor of 2, reducing the input size from $192 \times 192 \times 192$ to $6 \times 6 \times 6$. Additionally, batch normalization is implemented in conjunction with residual blocks to enhance training stability and model performance. The probabilities are then computed through softmax, which incorporates the mutual exclusivity between the three 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. Note that the background was included solely during the training process and was not utilized in the monitoring of loss functions.

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 process utilized approximately 30 GB of memory.

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% of the training data, which was passed through the model to estimate the probability of tumor presence.

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.

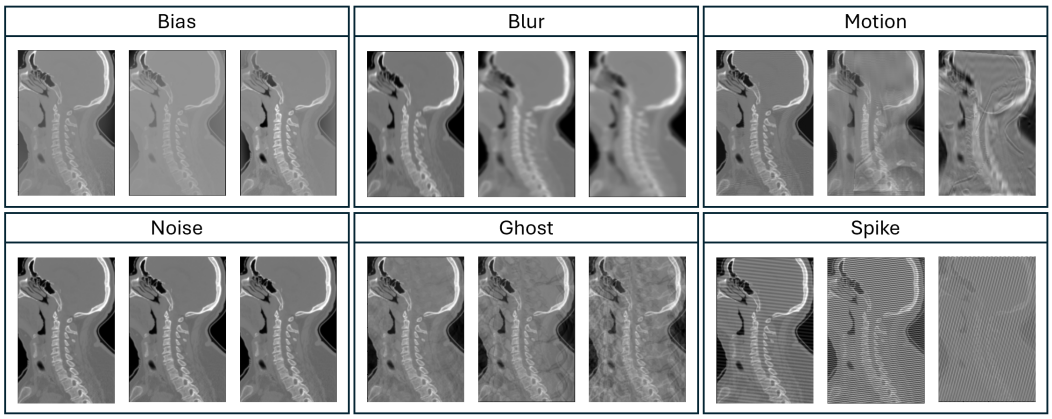


Figure 2.5. Visual representation of the perturbations applied

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. For detailed values of each perturbation at varying degrees, refer to the appendix (B).

Note: the noise perturbations may not be accurately represented here; A better visual representation of noise is provided in appendix (B)

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.
- **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 the 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.
- **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 method 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 two physicians specializing in 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.

This analysis was performed for both physicians' results. A correlation between the evaluations of the two physicians was also processed in order to compare with the Dice result. Both physicians have evaluated the cases under the same conditions to avoid any bias.

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.

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 and primary tumor labels in the CT modality.

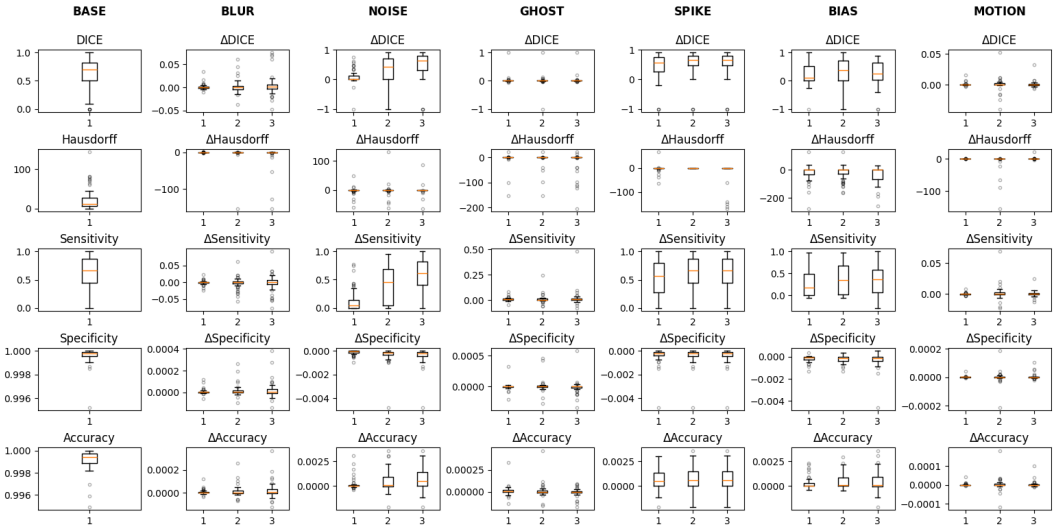


Figure 3.1. Perturbations applied to nodal tumor labels in the CT modality

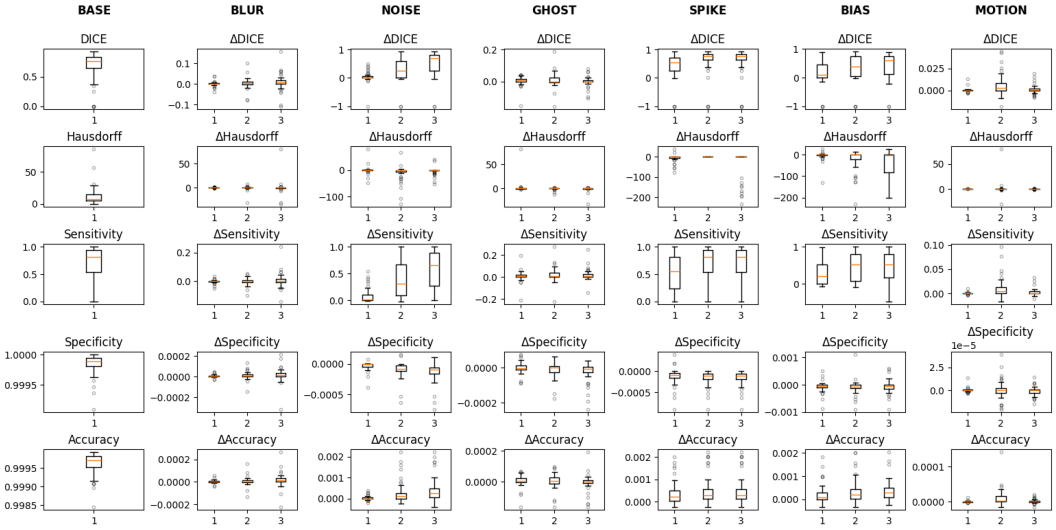


Figure 3.2. Perturbations applied to primary tumor labels in the CT modality

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.

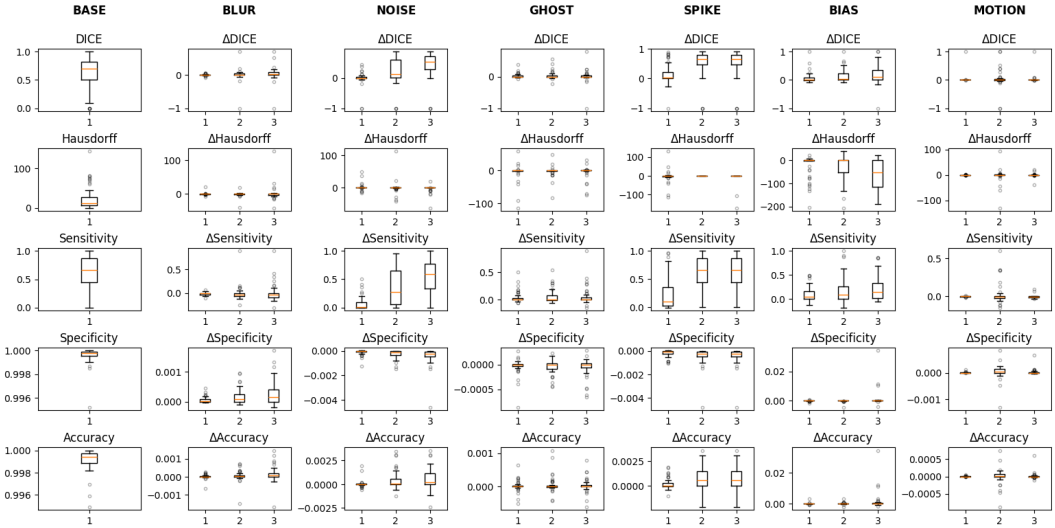


Figure 3.3. Perturbations applied to primary tumor labels in the PET modality

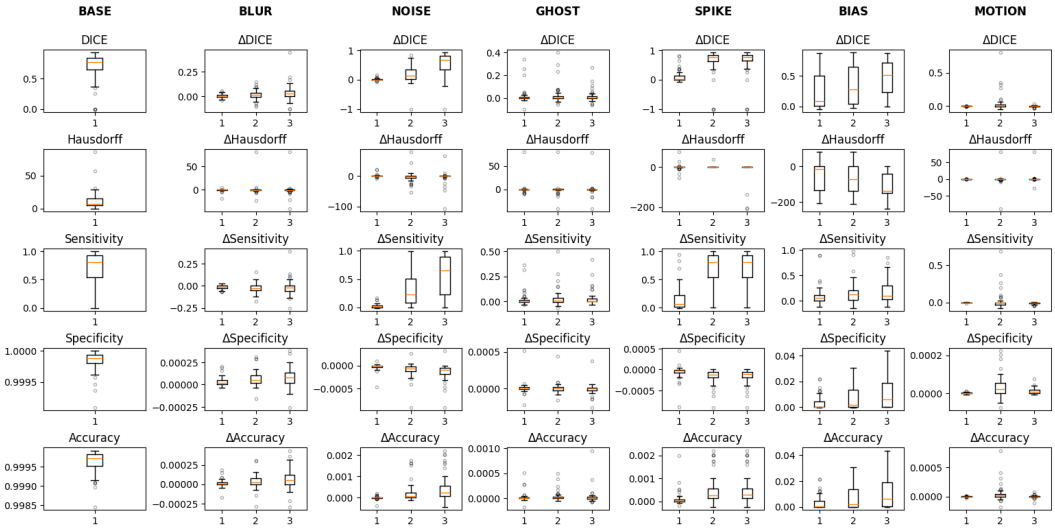


Figure 3.4. Perturbations applied to primary tumor labels in the PET modality

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 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 (2.3.1). 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) are provided in the appendix (A). A more comprehensive correlation graph, encompassing all four categories, will follow to offer a broader view of the relationships.

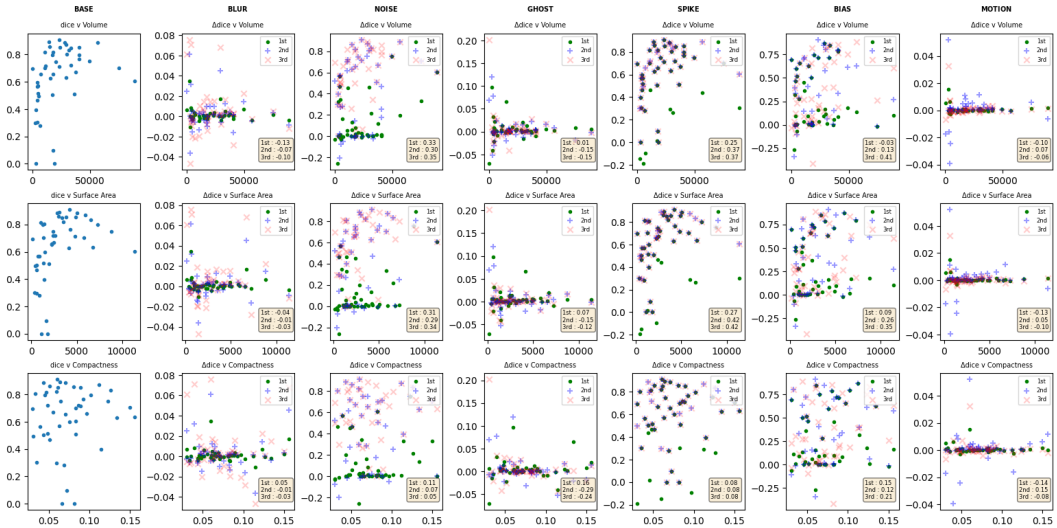


Figure 3.5. Correlation with perturbations applied to nodal tumor labels in the CT modality

On the left, the Dice coefficient exhibits a crescent-shaped distribution for the three properties. 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. Motion and ghost exhibit the highest variations in properties at lower values, whereas Blur demonstrates a more evenly distributed effect across the property range. The analysis demonstrates that the marker positions for spike events remain generally consistent across both second and third degrees. Furthermore, these markers are positioned in close proximity to the original baseline distribution, indicating a complete removal of prediction (dice = delta dice). The effect is also observed for bias and noise, though it is less pronounced. 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 as the properties values decrease, with the strongest correlation observed with the compactness property.

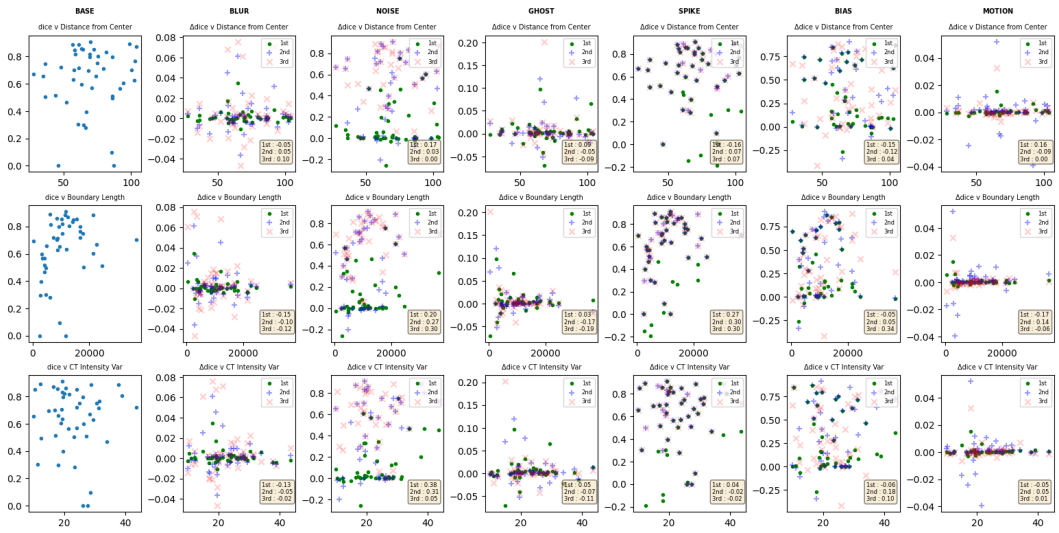


Figure 3.6. Perturbations applied to primary tumor labels in the CT modality

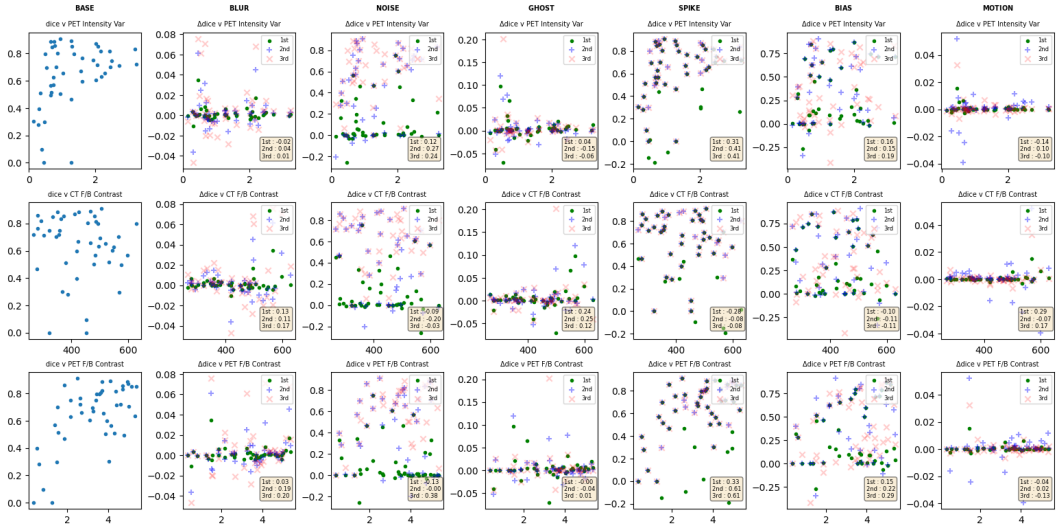


Figure 3.7. Perturbations applied to primary tumor labels in the CT modality

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.

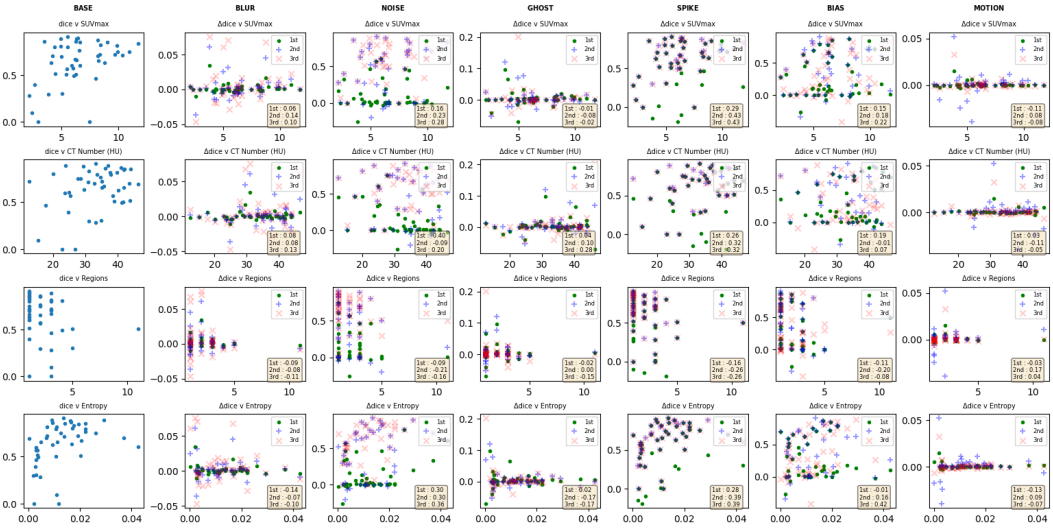
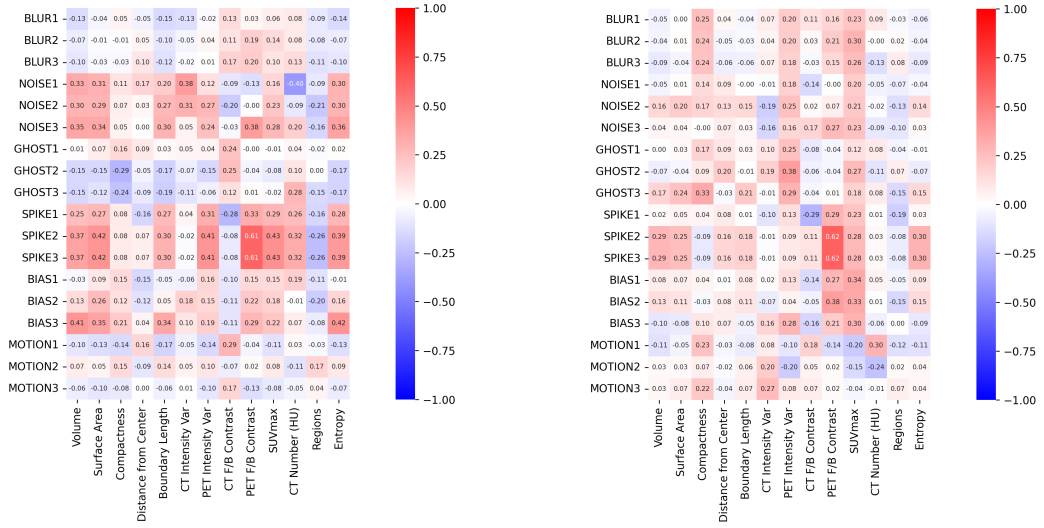


Figure 3.8. Perturbations applied to primary tumor labels in the CT modality

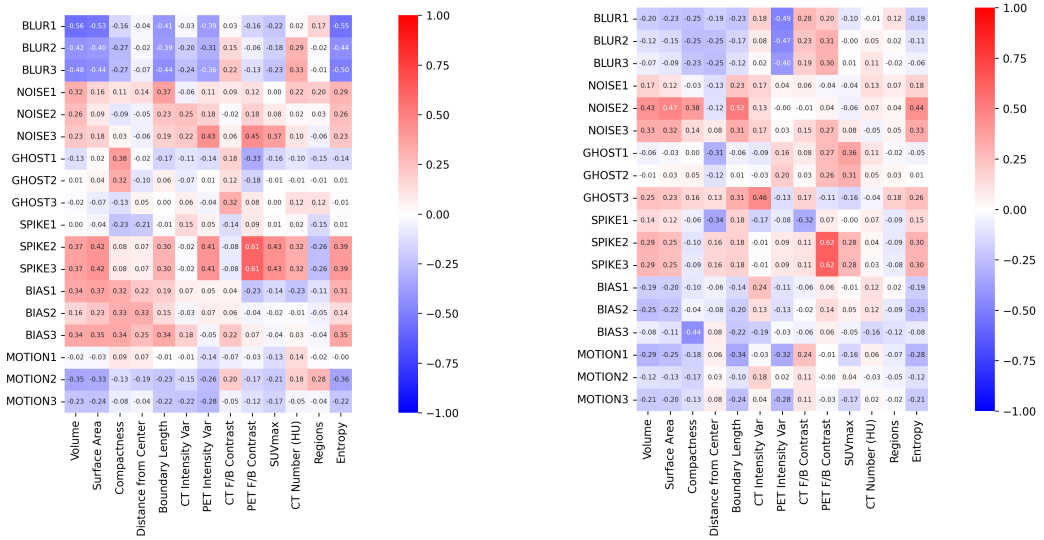
For the last four properties analyzed, the SUVmax demonstrated a moderate correlation with spike perturbations, ranging from 0.29 to 0.43 across the degrees of severity. Noise and bias also exhibited correlations, albeit at more moderate levels, around 0.20. In the case of CT number, spike perturbations 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 general graphic present the Pearson correlation coefficients for various properties across each perturbation and degree, as well as for the four image categories. This analysis covers both labels and modalities, providing a detailed overview of the relationships between segmentation properties and applied perturbations. The y-axis in these graphics shows the delta Dice scores caused by perturbations, with rows corresponding to the three severity levels, while the x-axis represents the 13 analyzed properties. A color scale on the right illustrates the strength and direction of the correlations.



(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 (a), 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 as most properties values decrease, 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.

In the primary tumor on the CT modality (b), a noticeable difference was observed between the two labels, particularly in the effect of compactness, which had a greater impact with lower segmentation degradation. Volume, surface area, boundary length, and entropy exhibited similar effects, primarily influencing spike perturbations, though with reduced correlation in the presence of bias and noise.

For primary tumors, SUVmax showed an effect across all perturbations, with a negative correlation under motion, while CT number had a reduced influence. Regions continued to improve segmentation but to a lesser extent compared to nodal tumors. Variability in PET data correlated with blur, noise, and ghost perturbations, whereas CT data showed more correlation with motion. Contrast behavior for both modalities was similar to nodal tumors, though with weaker correlation for CT under motion perturbations.

For the nodal region in the PET modality (c), the behavior of volume, surface area, boundary length, and entropy remains consistent with prior observations but demonstrates notable differences. These properties exhibit a strong positive correlation with spike, bias, and noise perturbations, and a strong negative correlation with blur and motion, with no significant effect observed for ghost perturbations.

SUVmax follows a similar trend to these four properties, though with a reduced correlation in the presence of bias perturbations. Compactness showed a positive correlation with bias, similar to distance from the center, while also presenting a negative correlation with blur.

Both intensity variabilities demonstrated negative correlations with blur and motion, and positive correlations with noise. PET variability additionally showed a stronger positive correlation with spike perturbations compared to CT.

For contrast, CT exhibited a positive correlation with blur and ghost perturbations, while PET showed a stronger correlation with spike and noise, accompanied by a slight negative correlation with blur and motion. The regional segmentation improvements observed previously were no longer general across all perturbations, instead showing a concentrated effect on spike perturbations.

For primary tumors in the PET modality (d), the four properties volume, surface area, boundary length, and entropy exhibit similar behavior as in prior observations. However, unlike in nodal tumors, the overall correlation is reduced, with a notable shift in bias correlation from positive to negative.

SUVmax is primarily correlated with spike and noise perturbations. The CT number shows significantly less impact compared to nodal tumors, with consistently low correlation coefficients, a pattern also observed for the regions.

Compactness demonstrates a negative correlation with blur and motion, with a slight positive correlation observed with noise.

For intensity variabilities, PET shows a strong negative correlation with blur and a less pronounced negative correlation with motion. In contrast, CT exhibits a positive correlation with noise.

Regarding contrast, PET shows correlations with spike, ghost, blur, and the third degree of noise, while CT primarily correlates with blur and motion.

Upon conducting a comprehensive comparison of the four graphs, it becomes evident that for the CT modality, the nodal tumor label exhibits a strong correlation with spike, noise, and bias, while the correlation for primary tumors is more distributed across various image properties, demonstrating specific and notable correlations.

In contrast, the PET modality reveals a higher prevalence of negative correlations, particularly with respect to motion and blur for both nodal and primary tumor labels. For nodal tumors in PET, positive correlations are similarly concentrated around spike, noise, and bias, akin to the CT modality. However, a distinguishing feature for nodal tumors in PET is the shift of bias towards negative correlation, and no clear resemblance with the CT modality is observed. Additionally, correlations in PET for nodal tumors are noted with spike, noise, and some degree of correlation with ghost artifacts. We can see that in the four graphs the second and third degree have mainly the same values like seen before indicating the complete removal of the prediction. Furthermore, across all degrees of severity for each perturbation, the correlation behavior is mostly not linear with increasing severity levels.

3.2 Clinical evaluation

The following graphs illustrate the distribution of the 5-star clinical evaluations for both observers. The left graph represents the original cases, comprising 32 evaluations, while the right graph shows the perturbed cases, with a total of 18 evaluations.

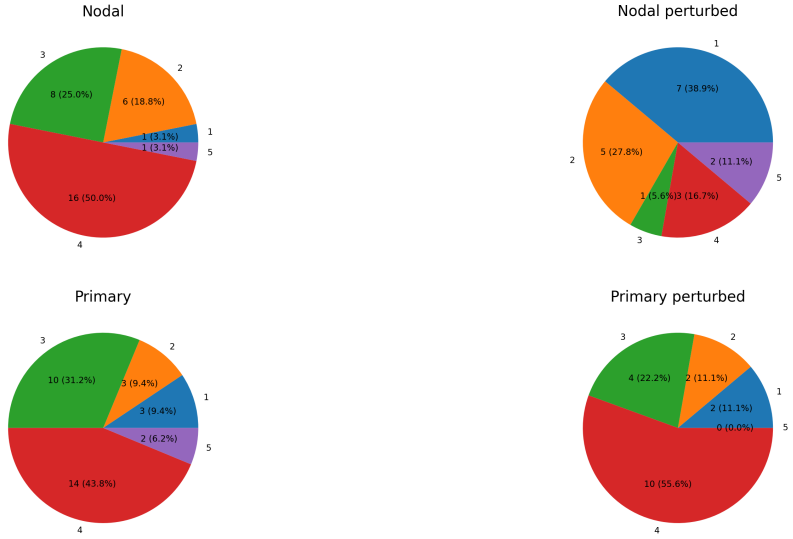


Figure 3.10. Clinical evaluation repartition observer 1

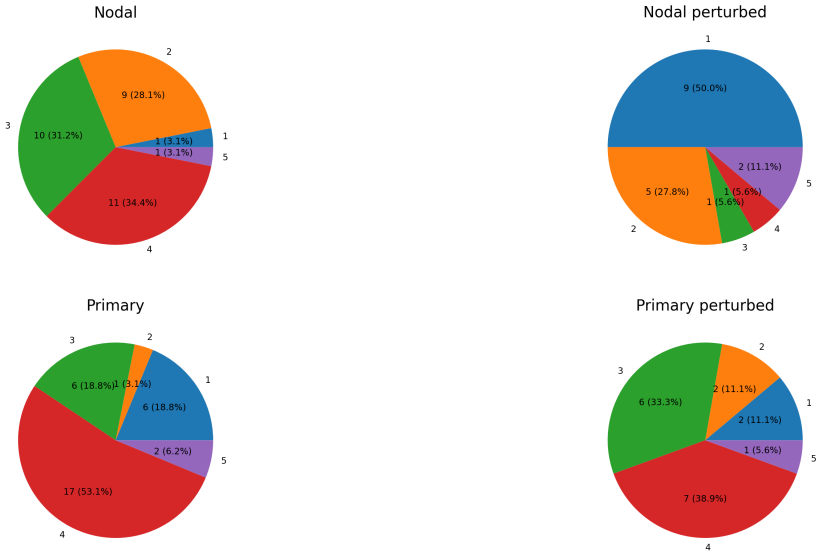


Figure 3.11. Clinical evaluation repartition observer 2

For the first observer, the analysis of the original cases reveals that approximately 75% of the evaluations fall between 3 and 4 stars for both channels. Notably, the nodal channel exhibits a slightly higher incidence of 2-star evaluations, alongside a reduction in the extreme values of 1 and 5 stars compared to the primary channel. In contrast, the perturbation analysis indicates that the nodal channel was significantly more affected by the applied perturbations in the selected cases with majority of 1 and 2 stars. Conversely, the distribution of evaluations for the primary channel remains consistent with that of the original cases but with some changes including more 4 stars suprisingly and not any 5 stars. Comparing with the second observer, for the original cases, the repartition for the nodal channel is more equal accross 2, 3 and 4 stars with approximately 30% each with 2 and 3 stars more present. However, for the primary tumors, the 4-stars percentage is higher than before with also a 1-star grade more prevalent. For the perturbed cases, the same behavior observed in nodal label with even worst impact in this evaluation. For the primary, the only major change is that 3-star are a bit more present at expense of the 4-stars.

In the subsequent graphs, we will further investigate whether these changes are also reflected in the DICE coefficient measurements, which would be indicated with a significant correlation between the DICE coefficient and the evaluation scores.

This graph that follows represent the corresponding 5 categories of dice coefficient for the same cases. With each category having a range of 0.2 from 0 to 1. Which would arbitrarly represent the gradings in likert scale.

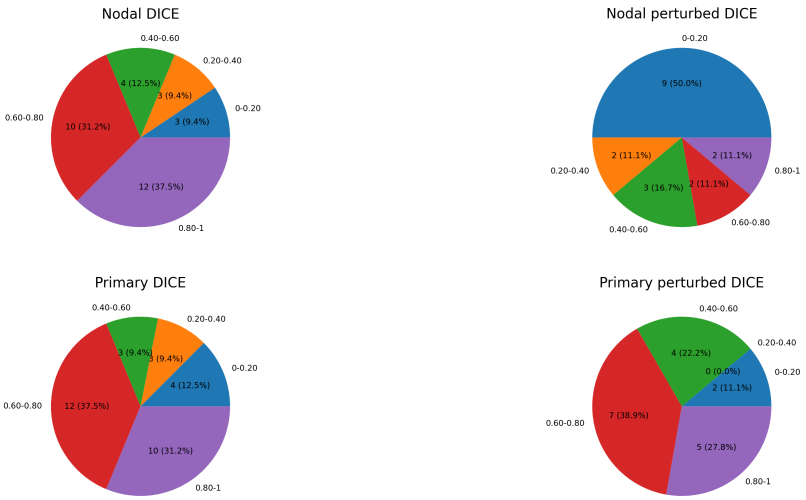


Figure 3.12. Clinical evaluation repartition

For the original labels, approximately 70% of the scores are above 0.60, with a similar distribution for scores below this threshold. In contrast, the perturbed labels show a notable decline, particularly for the nodal labels, where 50% of the scores fall between 0 and 0.2, and over 75% of the scores are below 0.6. The primary labels, however, maintain a similar distribution to the original, except that no scores are observed between 0.2 and 0.4.

Both grading systems and Dice coefficients follow a similar pattern, with a marked decline in nodal label performance under perturbation. Scores above 0.6 generally align with higher grading levels (3- and 4-star ratings).

The following graphs illustrate the correlation among three metrics: the DICE coefficient, Hausdorff distance, and Jaccard index, in relation to clinical evaluation gradings. The correlation analysis is conducted across both channels for the original 32 cases. The Likert gradings are plotted on the x-axis, while the metrics are represented on the y-axis. In each plot, the correlation trend line is depicted in red, with the Pearson correlation coefficient displayed in the top right corner.

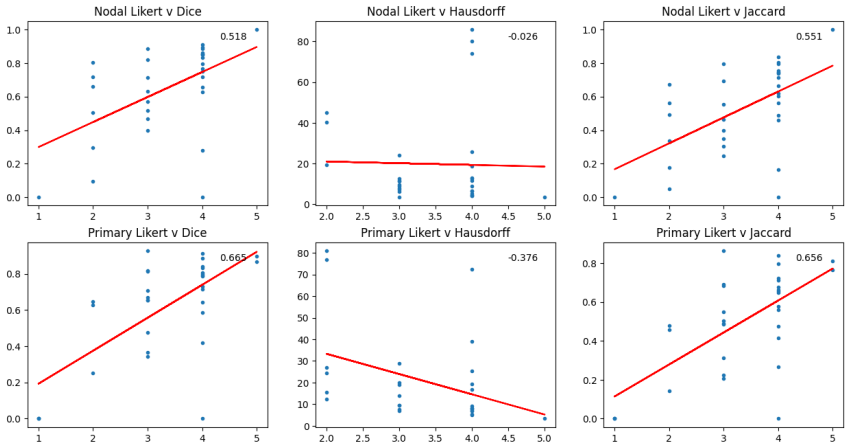


Figure 3.13. Observer 1 clinical evaluation correlation with DICE on original cases

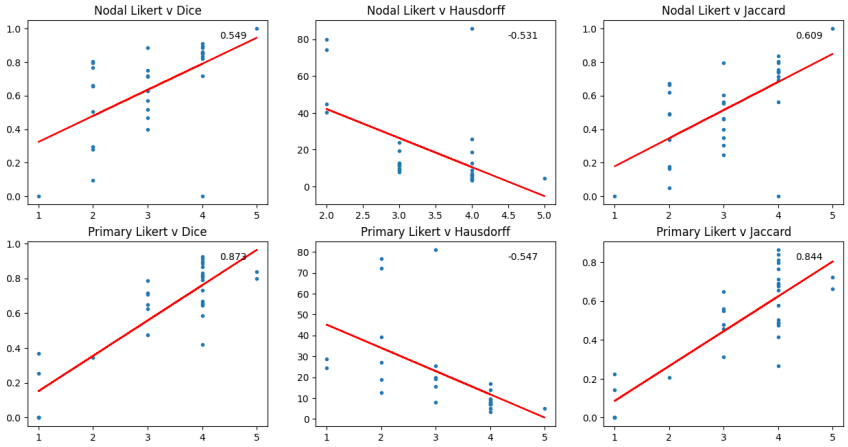


Figure 3.14. Observer 2 clinical evaluation correlation with DICE on original cases

For the first observer, the Jaccard index and DICE coefficient demonstrate similar behavior, with highly aligned trend lines and strong Pearson correlation coefficients. In contrast, the Hausdorff distance shows a weaker correlation, particularly for cases without ground truth data, which were excluded due to the sensitivity of this metric in such instances. Among the two most reliable metrics, the primary label exhibits stronger correlations, with values of 0.665 and 0.656, compared to the nodal channel, which yields lower correlations of 0.518 and 0.551.

Additionally, there is a tendency for the range of DICE coefficients to contract at higher grading levels for nodal label. Interestingly, some cases rated with four stars exhibit low DICE coefficients. These cases were reviewed with clinical experts and identified as particularly challenging, where the image alone was insufficient for accurate evaluation without supplementary information.

In comparison to the second observer, the nodal label shows similar behavior, with comparable point distributions, reflected by close correlation coefficients. However, a notable difference arises in the Hausdorff distance, which here demonstrates a more representative and improved correlation. For the primary label, the second observer achieves significantly higher correlations with the DICE coefficient, around 0.85 compared to 0.65 for the first observer's DICE and Jaccard indices, and -0.55 compared to -0.37 for the Hausdorff distance. Additionally, the second observer shows less variability in the DICE range for higher gradings.

The following graph exhibit the correlation between the two physicians for the original cases.

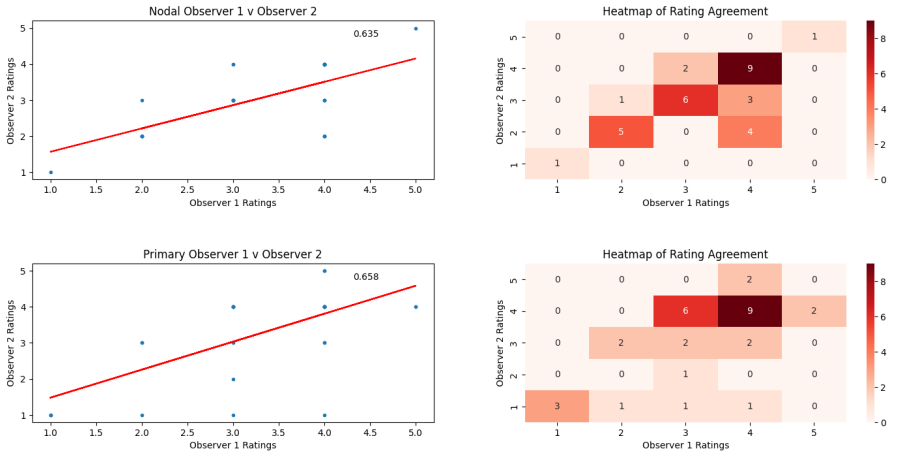


Figure 3.15. Phyisician correlation in original cases

For the nodal label, the majority of grades (22 out of 32) lie on the diagonal, indicating a strong level of agreement between observers. However, the second observer tends to assign slightly lower grades compared to the first observer. The correlation between the two observers for the nodal label is 0.635, which, while higher than the correlation between the DICE coefficient and the physician's assessment (0.518 and 0.549), does not represent a significant increase.

For the primary label, the inter-observer agreement is more dispersed, which accounts for the observed difference in correlation coefficients between the first and second observers (0.665 vs. 0.873). The correlation for the primary label (0.658) is similar to that for the nodal labels (0.635), as well as to the correlation between the first observer's assessment and the DICE coefficient (0.665). However, this correlation is lower when compared to the second observer's correlation with the DICE coefficient (0.873).

Here is the same graphs and analysis but for the 18 perturbed cases cases.

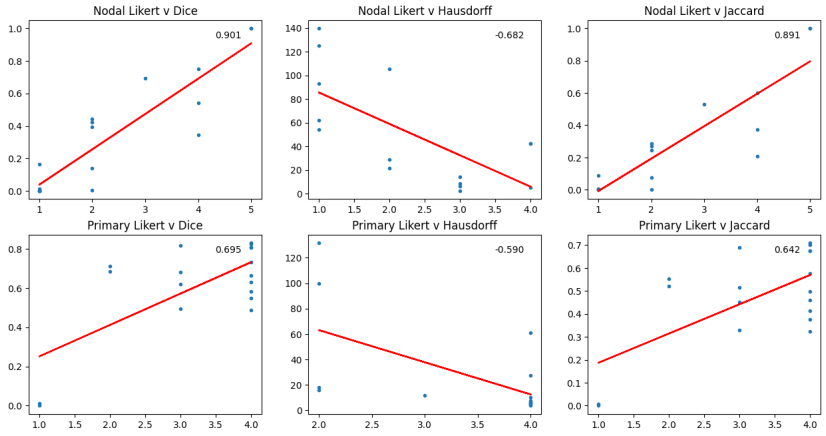


Figure 3.16. Observer 1 clinical evaluation correlation with DICE on perturbed cases

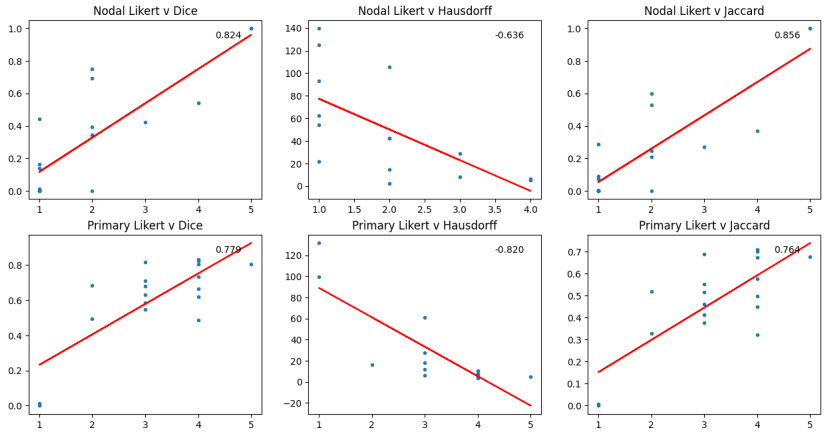


Figure 3.17. Observer 2 clinical evaluation correlation with DICE on perturbed cases

The Hausdorff distance appears to be more representative, showing higher correlation values. However, the DICE and Jaccard indices remain more consistent across different cases and are less affected by the absence of ground truth in one label.

For the first observer, the peak correlation is 0.901 for the nodal label. In contrast, for the second observer, the difference between nodal and primary labels is less pronounced, with correlations of 0.824 and 0.778, respectively, compared to 0.901 and 0.695 for the first observer. Overall, for both observers, the correlation values are higher compared to the original cases. In contrast to the original cases, the point distribution similarity between the two observers is less pronounced.

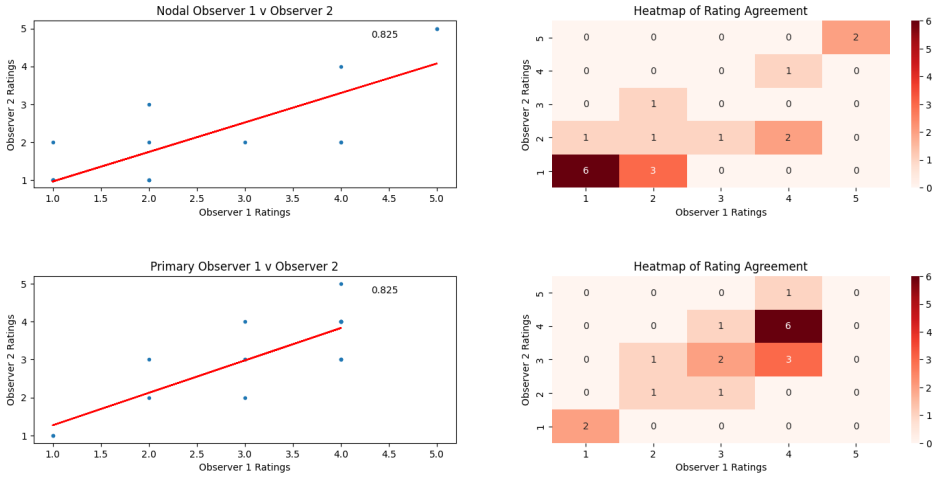


Figure 3.18. Phyisician correlation in perturbed cases

The correlation for both channels is observed to be identical, with values of 0.825, closely matching the correlation seen with the DICE coefficient. As illustrated in the pie chart, the nodal label predominantly consists of lower grades. In contrast, the primary label exhibits a wider distribution of grades, with a higher proportion of elevated scores.

For the nodal label, Observer 2 tends to assign lower grades for the same cases compared to Observer 1. However, for the primary label, the grades are more evenly distributed between the two observers.

Compared to the original cases graph, the correlation is significantly higher with 0.825 against 0.635 for nodal and 0.825 against 0.658 for the primary tumors.

Chapter 4

Discussion and Conclusions

4.1 Discussion

4.1.1 Robustness

The results of the perturbation analysis demonstrate that the model exhibits robustness to approximately half of the applied perturbations, including ghosting, blurring, and motion artifacts with The Delta Dice coefficient exhibited a maximum variation of approximately 0.1 for the most severe cases assessed. During model training, various intensity augmentations were employed, such as Gaussian noise, smoothing, intensity shifts, and contrast adjustments. These augmentations were expected to enhance the model's performance under these perturbations. However, it is important to note that these augmentations were applied exclusively to the CT modality, while the PET modality also demonstrated resilience to these perturbations. This observed robustness of the PET modality to perturbations such as blur, motion, and ghosting could be attributed to its lower structural complexity compared to CT. PET imaging primarily provides functional information in the form of Standardized Uptake Value (SUV) heatmaps, which are inherently less reliant on fine anatomical details. As a result, PET images may exhibit a natural resistance to perturbations that distort or duplicate structural features, such as blur and ghosting, which typically affect high-resolution modalities like CT. Moreover, the lack of intricate anatomical structures in PET likely contributes to its relative insensitivity to motion artifacts, as these perturbations do not significantly compromise the functional information provided by the heatmaps.

Interestingly, despite the inclusion of noise augmentation during training, the model showed lower robustness to noise compared to the aforementioned perturbations. The model remained relatively robust at lower levels of noise severity, but performance degraded as the noise severity increased. On the other hand, perturbations such as spike noise and bias were particularly problematic, causing significant degradation in performance even at the lowest levels of severity with the Delta Dice coefficient approached values as high as 1 or near it.

Consultation with a clinician revealed that the severity of the applied perturbations was higher than what is typically encountered in clinical practice. According to the clinician, such extreme perturbations would likely warrant image re-acquisition. This insight suggests that the model's demonstrated robustness to motion, blur, and ghosting artifacts is likely to hold in real-world scenarios, where such perturbations occur at lower levels of severity. Consequently, the model's performance under these conditions may be considered sufficient for practical clinical applications. Nevertheless, to improve the model's robustness to these more severe perturbations, future training could incorporate augmentations involving bias, spike , and higher levels of noise severity. This approach could potentially enhance the model's resilience to these challenging conditions.

The analysis of spatial properties such as volume, surface area, boundary length, and entropy reveals an average correlation of approximately 0.3 with perturbations including spike, bias, and noise in nodal tumors. This observation suggests that larger and more complex regions are more susceptible to these perturbations. Conversely, a negative correlation was identified in PET imaging for both channels concerning motion and blur, with correlation values reaching as low as -0.56 for blur. Furthermore, the bias exhibited a negative correlation for the primary label in PET imaging, indicating that smaller and less complex primary tumors are disproportionately affected by these perturbations compared to nodal tumors.

Additionally, when perturbations are introduced in CT imaging, particularly for primary tumors, the variability and maximum standardized uptake value (SUVmax) in PET demonstrate stronger correlations than observed in scenarios where perturbations are applied directly to PET. This finding suggests an inherent relationship between the metabolic activity of tumors (from PET SUVmax) and their sensitivity to CT-based segmentation perturbations, highlighting how tumors with greater metabolic intensity tend to experience more segmentation deviations when the CT is altered. In contrast, the intensity properties of CT do not yield significantly higher correlations with PET when subjected to perturbations compared to the effects observed when perturbations are directly applied to PET.

Moreover, focusing on PET variability, we observed a prominent negative correlation of approximately -0.45 for primary tumors under PET perturbations, indicating that the impact of blur is more pronounced in less variable signals. This makes sense because uniform metabolic activity often correlates with a simpler, smoother anatomical structure. As a result, these tumors are less sensitive to the smoothing effect of blurring, since their boundaries are easier to detect. Overall, examining the specific effects of perturbations rather than solely focusing on spatial properties indicates that the most impactful perturbations—specifically bias, noise, and spike—tend to correlate with higher values. In the context of PET imaging, the perturbations of blur and motion exhibit negative correlations across both channels, while bias correlates negatively only for the primary channel. This suggests that generally, the lower the values of properties such as size, intensity, variability, and complexity, the more significantly the segmentation is affected by these perturbations. The observed similarity between the spike degrees 2 and 3 across all four graphs can be attributed to the fact that, at these levels of degradation, the dice coefficient frequently approaches zero. This extreme reduction in dice value results in nearly identical delta values for these degrees, leading to a similar pattern of correlation propagation.

4.1.2 Clinical evaluation

As shown in the pie chart, the majority of the scores are above 0.6, indicating strong performance from a metric perspective. When comparing these results with real-world gradings, we observe that most cases fall within the 3- and 4-star categories, suggesting promising clinical utility.

To further validate these findings, we examined the correlation to confirm that these scores accurately represent the underlying cases. For the original cases, the correlation was found to be in the high-moderate range, with a stronger correlation for primary tumors for both observer. This is expected, as primary tumors tend to be singular and centrally located in the images, while nodal labels can be multiple and distributed across both sides, introducing more variables that can affect the model's and physician's judgment.

Following perturbation, we observed high correlations (0.901 and 0.824) for nodal labels, which may seem significant. However, this result should be interpreted with caution, as there were several 1-star evaluations (7 and 9 out of 18 cases) where the model failed to segment any regions. These outliers are clustered closely in the graph, artificially inflating the correlation. In contrast, for primary tumors, this bias was much less pronounced, with only two instances of 1-star evaluations. The correlations for primary tumors after perturbation were a bit higher than for the original cases, indicating that the metrics remained consistent and relevant to real-world clinical performance, even under perturbation.

Also, something to take in account and seen with clinicians, they doesn't agree everytime with the groundtruth which create bias with dice which is directly base on the ground truth. In ideal world the ground truth and the evaluation could be performed by the same person. Also, in clinical practice they always use MRI for the diagnostic especially for primary tumors which was not able in this dataset.

These results demonstrate a higher correlation compared to the existing literature [6]. The observed correlation indicates that the Dice coefficient is relevant for real-world applications, although there is still room for improvement in its predictive accuracy.

4.1.3 Limitations

As previously discussed, the absence of MRI data represents a significant limitation, as MRI is currently the gold standard for the diagnosis of primary tumors in the head and neck region. Its omission restricts the model's applicability in clinical practice, where MRI plays a crucial role in providing detailed anatomical and soft tissue contrast essential for accurate diagnosis.

Furthermore, the evaluation process is limited by the lack of patient-specific information, as the Hecktor dataset provides such information, including HPV status, for only a subset of patients (primarily for two out of seven sites). This lack of comprehensive data may affect the accuracy and relevance of the clinical assessments.

Additionally, clinician disagreement with the ground truth in certain cases introduces a potential bias against the Dice coefficient, as it is directly dependent on the ground truth. This discrepancy can undermine the validity of the Dice score as a reliable metric in such instances.

4.2 Conclusions

The robustness analysis of the model has shown its capacity to handle common perturbations such as ghosting, blurring, and motion artifacts with minimal performance degradation, particularly in PET imaging. The PET modality's resilience is likely due to its reliance on functional data rather than fine anatomical details, making it less vulnerable to structural distortions. However, despite noise augmentation during training, the model exhibited sensitivity to noise, especially spike and bias perturbations, which resulted in significant performance drops even at low levels of severity. This suggests a need for further training augmentations that address these specific challenges.

Correlation analysis of spatial properties demonstrated that larger and more complex nodal tumors were more susceptible to noise and bias, while smaller and less complex primary tumors were disproportionately affected by perturbations, particularly in PET. Interestingly, CT-based perturbations had a stronger effect on PET segmentation accuracy, especially for tumors with higher metabolic activity. This intermodal sensitivity points to the potential need for more integrated augmentation strategies.

From a clinical perspective, the model's performance is promising, with most cases achieving strong metric scores. However, some discrepancies between the model's segmentation and clinical ground truth were noted, highlighting a challenge in aligning automated model outputs with real-world clinical evaluations, particularly for primary tumors where MRI is typically preferred but unavailable in this dataset. The clinician's disagreement with ground truth in some instances also underscores the limitations of relying solely on Dice scores for evaluation. Nonetheless, the high correlations observed, even under perturbations, suggest that the model maintains clinical relevance.

Chapter 5

Outlook

Integrating additional imaging modalities, such as Magnetic Resonance Imaging (MRI), into the segmentation pipeline could significantly enhance model performance, particularly in identifying necrotic tissues. The incorporation of contrast-enhanced Computed Tomography (CT) scans may further aid in distinguishing between lymph nodes and surrounding vascular structures, thereby improving the accuracy of segmentation outcomes.

In light of the findings from our analysis, there is an opportunity to retrain the segmentation model by leveraging insights gained throughout this study. This may involve refining data augmentation strategies to include a broader range of perturbations, such as biases, spikes, and various forms of noise, which can enhance the model's robustness. Additionally, a comprehensive analysis of the segmentation performance with the newly introduced modalities, alongside the enhanced CT images, would provide valuable insights into how these modifications impact the segmentation accuracy and the correlations between the Dice coefficient and clinical assessments.

With the segmentation model achieved, a subsequent classification model could be developed to evaluate the likelihood of extracapsular extension (ECE). This model would facilitate further analysis of its confidence levels in detecting ECE, providing clinicians with a valuable tool for improving diagnostic precision and patient management strategies.

Bibliography

[1] V. Andrearczyk, V. Oreiller, S. Boughdad, C. C. L. Rest, H. Elhalawani, M. Jreige, J. O. Prior, M. Vallières, D. Visvikis, M. Hatt, and A. Depeursinge. Overview of the hecktor challenge at miccai 2021: Automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, pages 1–37, 2022.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] R. H. e. a. Mohamed A. Naser, Lisanne V. van Dijk. Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality pet/ct images. In Head and Neck Tumor Segmentation, volume 12603, pages 85–98, 2021.

[9] 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.

[10] 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.

[11] 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.

[12] S. F. e. a. Thomas Weissmann, Yixing Huang. Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. Frontiers in Oncology, 13:1115258, 2023.

[13] L. H. e. a. Yiling Wang, Elia Lombardo. Comparison of deep learning networks for fully automated head and neck tumor delineation on multi-centric pet/ct images. Radiation Oncology, 19:3, 2024.

[14] 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

Correlation with properties

A.1 Properties correlation with perturbation on primary label in CT

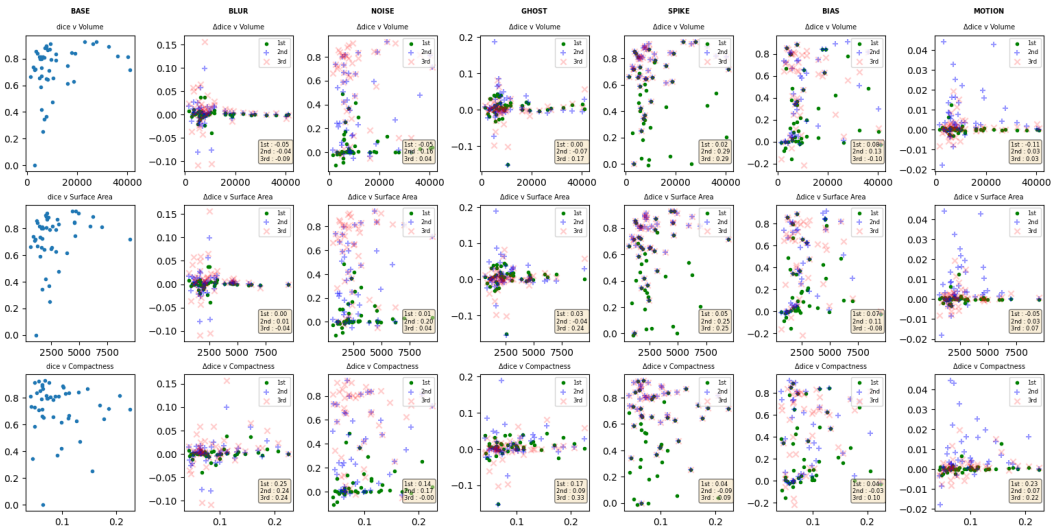


Figure A.1. Correlation with Volume, Surface Area and Compactness

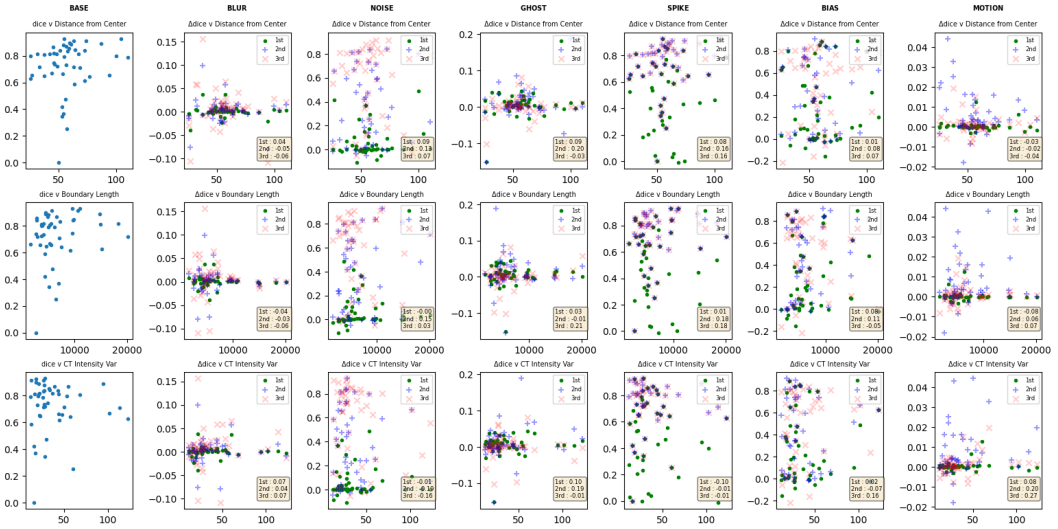


Figure A.2. Correlation with Distance from Center, Boundary Length and CT Intensity Var

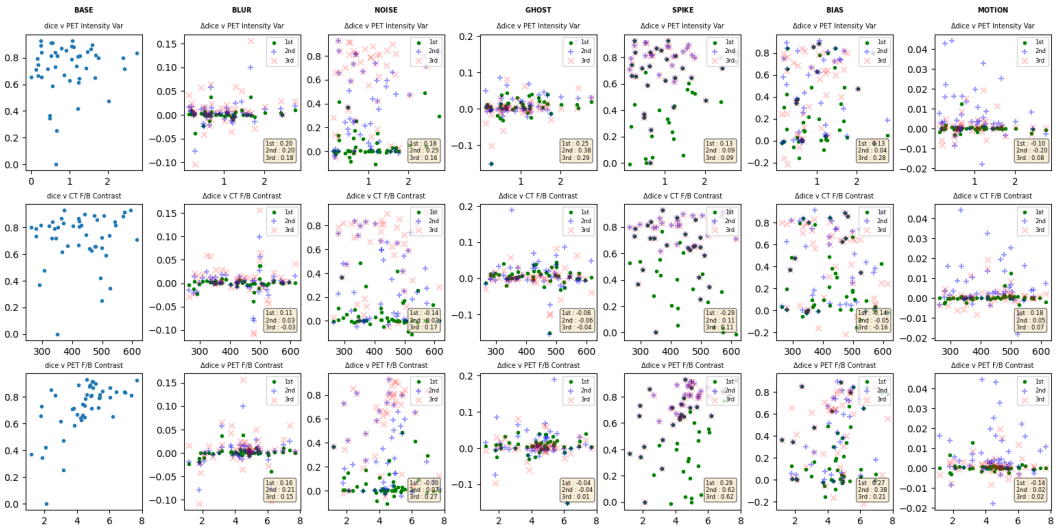


Figure A.3. Correlation with PET Intensity Var, CT contrast and PET contrast

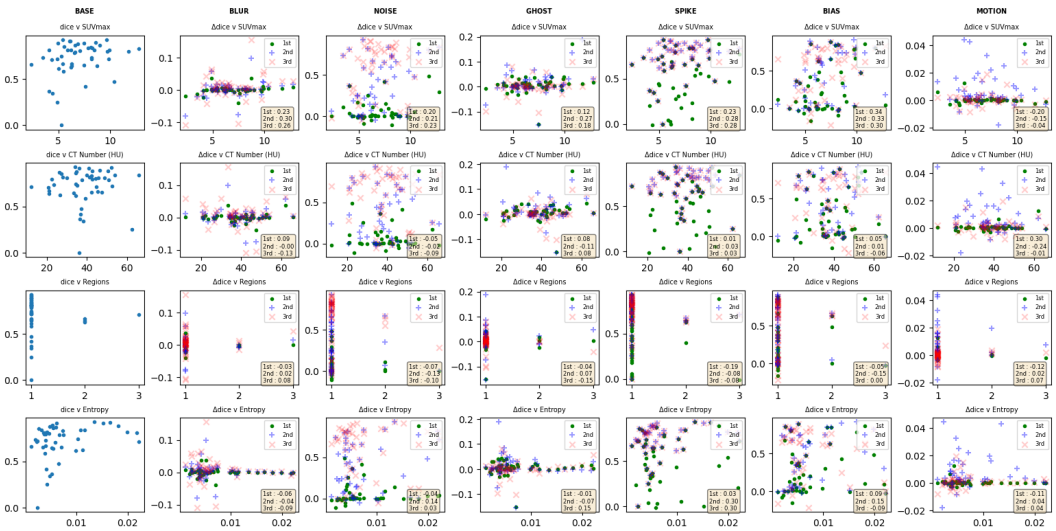


Figure A.4. Correlation with SUVmax, CT number, Regions and Entropy

A.2 Properties correlation with perturbation on nodal label in PET

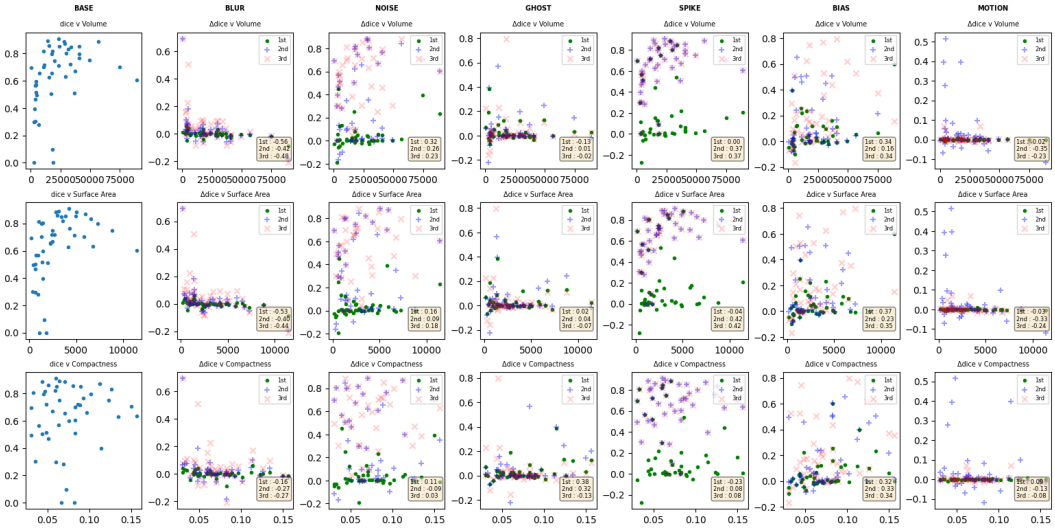


Figure A.5. Correlation with Volume, Surface Area and Compactness

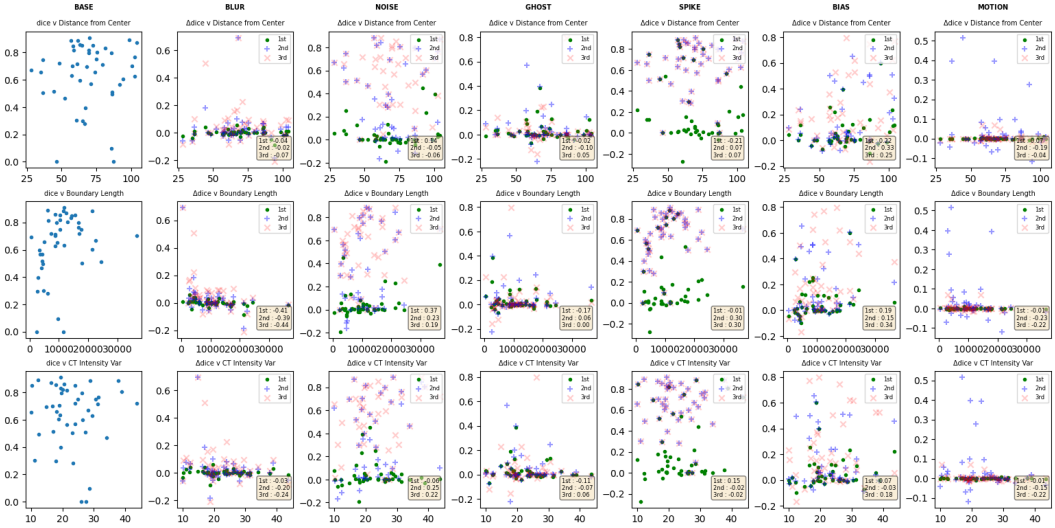


Figure A.6. Correlation with Distance from Center, Boundary Length and CT Intensity Var

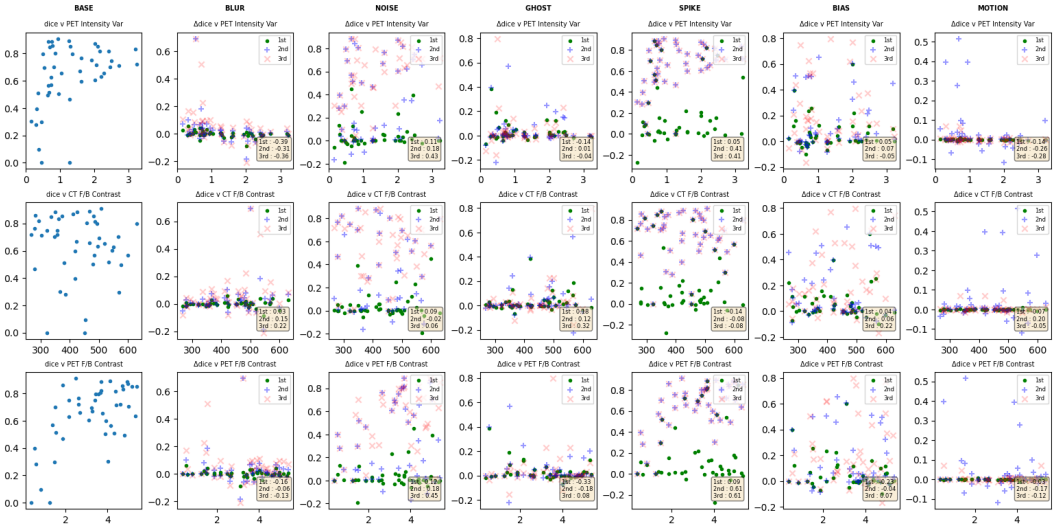


Figure A.7. Correlation with PET Intensity Var, CT contrast and PET contrast

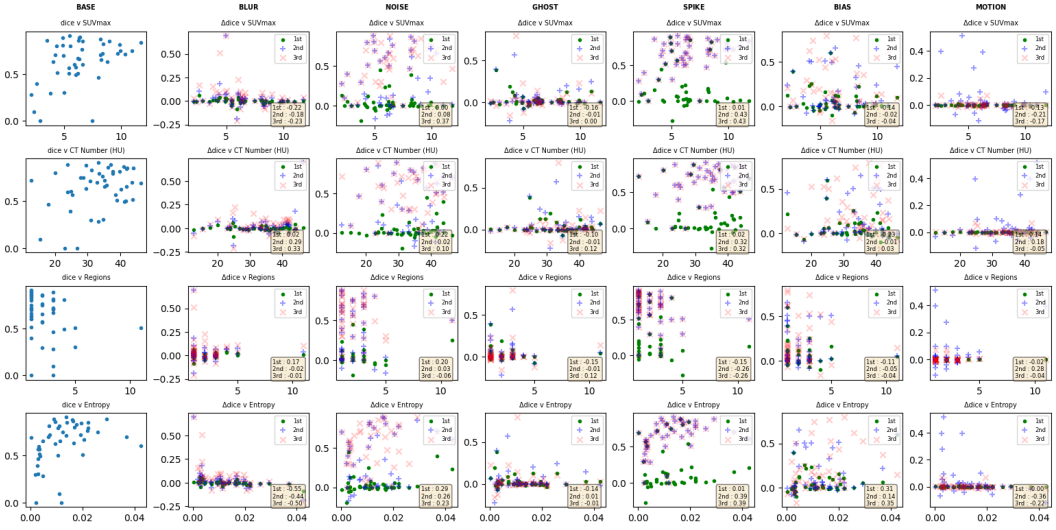


Figure A.8. Correlation with SUVmax, CT number, Regions and Entropy

A.3 Properties correlation with perturbation on primary label in PET

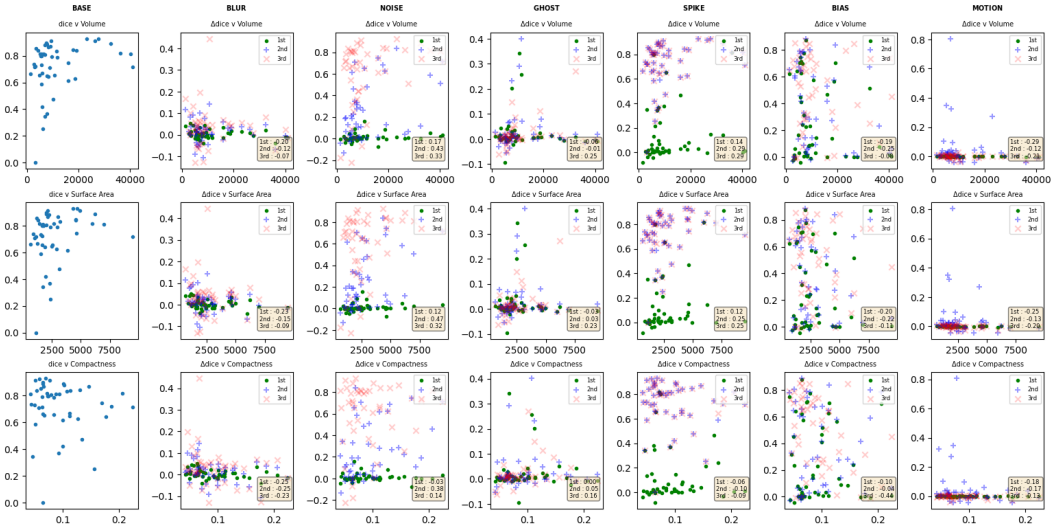


Figure A.9. Correlation with Volume, Surface Area and Compactness

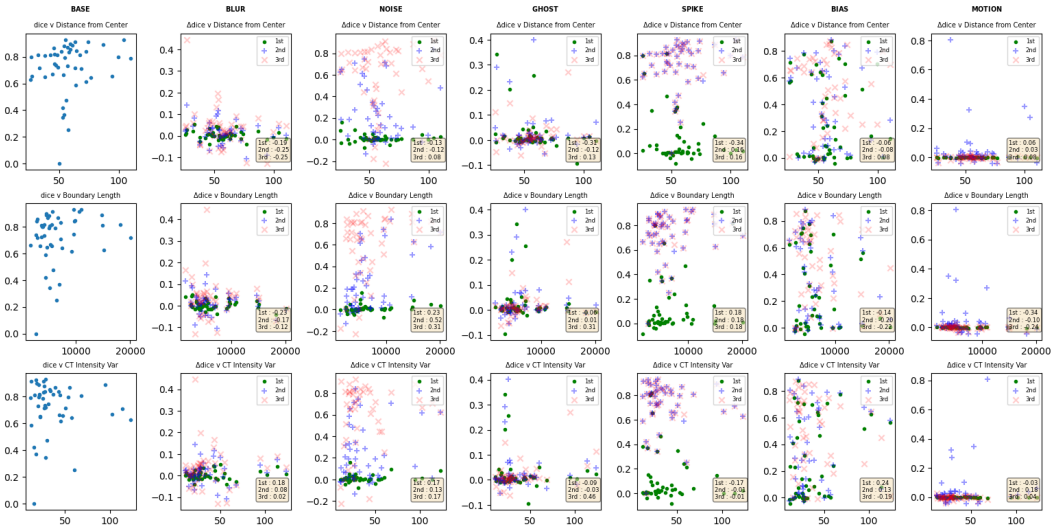


Figure A.10. Correlation with Distance from Center, Boundary Length and CT Intensity Var

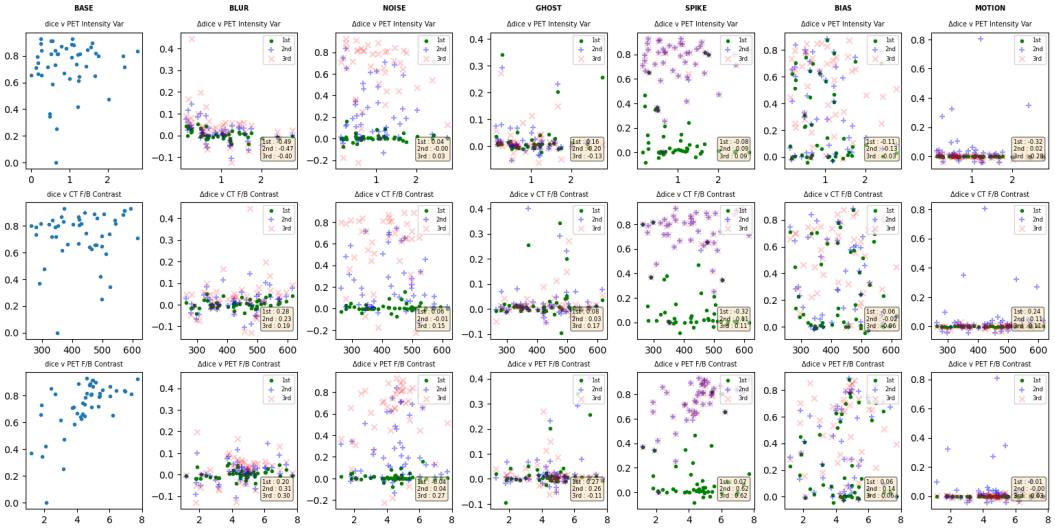


Figure A.11. Correlation with PET Intensity Var, CT contrast and PET contrast

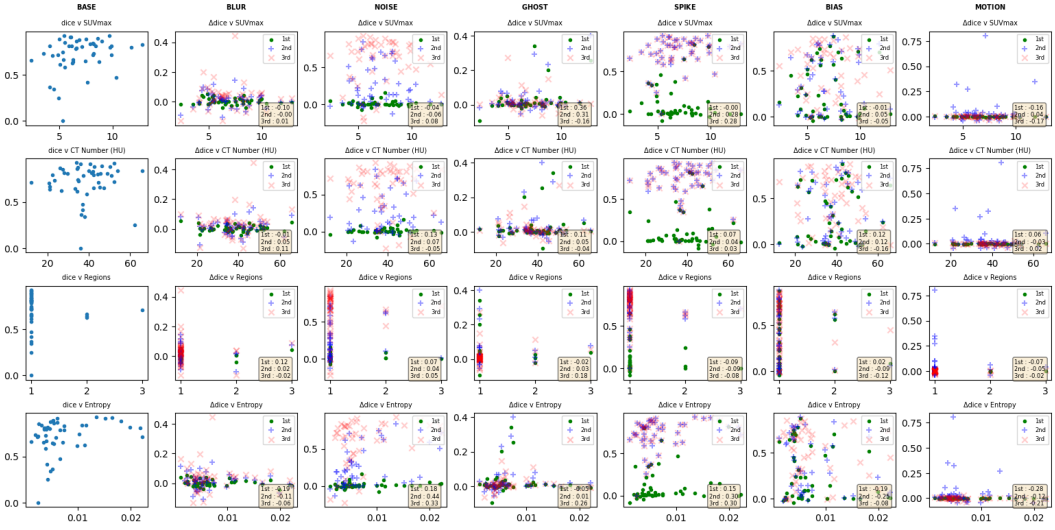


Figure A.12. Correlation with SUVmax, CT number, Regions and Entropy

Appendix B

Perturbations

B.1 Values of the perturbations

Perturbation Type	Degree 1	Degree 2	Degree 3
Motion	Degrees: 10, Translation: 10, Num Transforms: 2	Degrees: 20, Translation: 20, Num Transforms: 3	Degrees: 30, Translation: 30, Num Transforms: 4
Ghosting	Num Ghosts: (4, 10), Axes: (0, 1, 2), Intensity: (0.5, 1), Restore: 0.02	Num Ghosts: (6, 15), Axes: (0, 1, 2), Intensity: (0.7, 1.5), Restore: 0.04	Num Ghosts: (8, 20), Axes: (0, 1, 2), Intensity: (1, 2), Restore: 0.06
Spike	Num Spikes: 1, Intensity: (1, 3)	Num Spikes: 2, Intensity: (2, 5)	Num Spikes: 3, Intensity: (3, 7)
Bias Field	Coefficients: 0.5, Order: 3	Coefficients: 0.75, Order: 3	Coefficients: 1.0, Order: 4
Noise	Mean: 0, Std: (0, 0.25)	Mean: 0, Std: (0.1, 0.5)	Mean: 0, Std: (0.2, 0.75)
Blur	Std: (0, 2)	Std: (1, 3)	Std: (2, 4)

Figure B.1. Values of the perturbations

B.2 Representation of noise perturbation

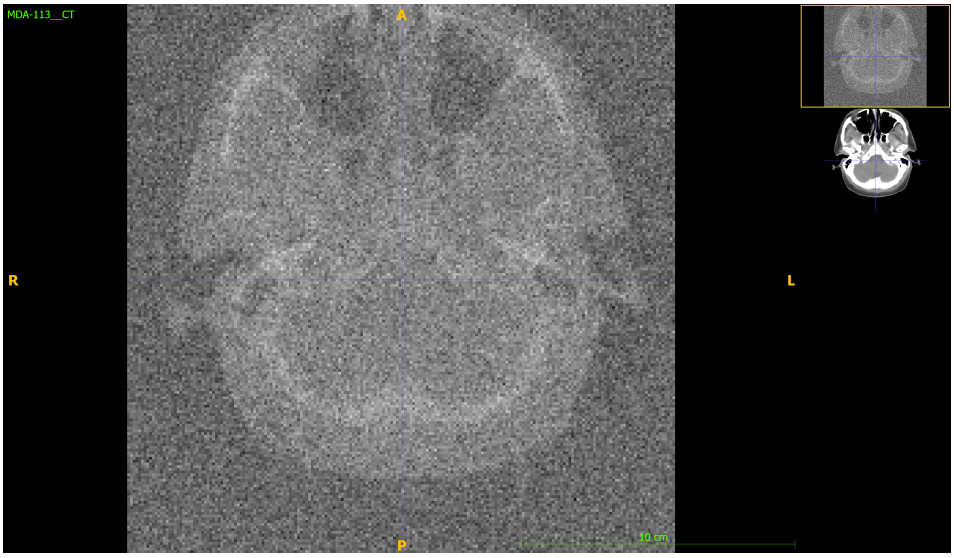


Figure B.2. Noise perturbation on CT

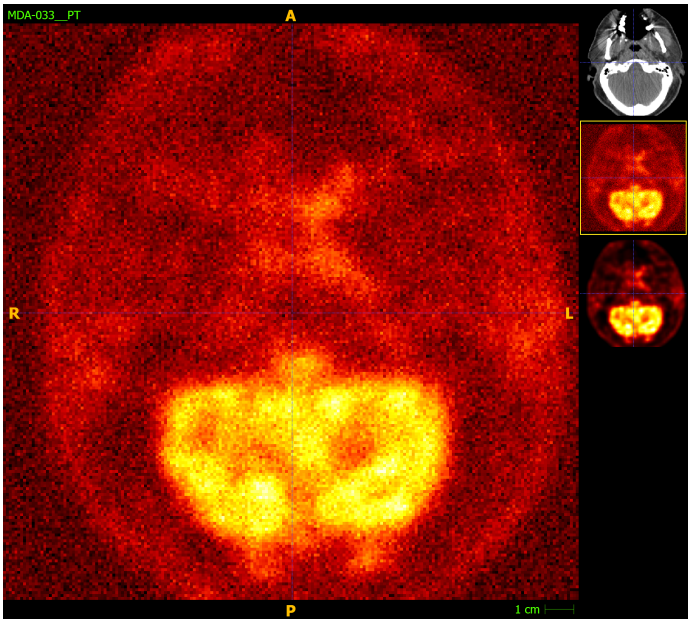


Figure B.3. Noise perturbation on PET