AI for Assessing Peritoneal Metastases on Computed Tomography Imaging

Yao Tong
y.tong@student.tue.nl

***Abstract*—Peritoneal Metastases (PM) pose a diagnostic challenge in Computed Tomography (CT) imaging, presenting considerable interobserver variability even among expert radiologists. To standardize and enhance the determination for assessing PM, this study employs a two-fold approach. Firstly, a pretrained nnUNet is fine-tuned to achieve 13-region segmentation in the abdominal area. These 13 regions are defined according to Sugarbaker's Peritoneal Cancer Index (PCI). As the initial study addresses the abdomen 13 regions partition aligning with PCI criteria, preliminary experiments using three regions demonstrate 83.46% for inter-organ DSC, 77.25% for inter-organ NSD, and inter-organ HD95 of 6.07mm. Secondly, due to the absence of a publicly available dataset for PM nodules, we chose to transfer the detection task into a classification task. The PCI score dataset is carried out for classification purposes. Subsequently, the ResNet50 is fine-tuned with this dataset, to classify the CT image into the corresponding PCI score for three abdominal regions: R1, R2, and R3. Finally, we perform the experimental evaluations to validate the model's effectiveness on the PCI score dataset.**

***Index Terms*—Peritoneal Metastases, nnUNet, Segmentation, ResNet50, Classification**

I. Introduction

The peritoneum, which is the largest and most complex membrane in the human body, lines the space between the abdominal wall and pelvic cavities[1], and covers the intra-abdominal organs within the abdominal cavity. Metastases occur when cancer cells spread from their original tumor to other parts of the body. Peritoneal Metastases (PM) refer to metastases that spread to the peritoneum originating from the intra-abdominal organs. They are predominantly seen in patients with, among others, gastric, colorectal, and ovarian cancer. PM is often hard to diagnose on Computed Tomography (CT) imaging, thus, performing a diagnostic laparoscopy (DLS) becomes the most effective way to identify PM at the current stage. During DLS, the extent of PM is evaluated using a scoring system named Sugarbaker's Peritoneal Cancer Index (PCI) [2]. To calculate the PCI, the abdominal cavity is divided into 13 regions based on the anatomical structure, as illustrated in Figure 1. Based on the size of the tumor(s) in each region, a score of 0 to 3 will be assigned to that particular region. The sum of scores from all regions forms the final PCI, ranging from 0–39. The PCI score plays a crucial role in evaluating operability, locoregional treatment options, and prognosis [2].

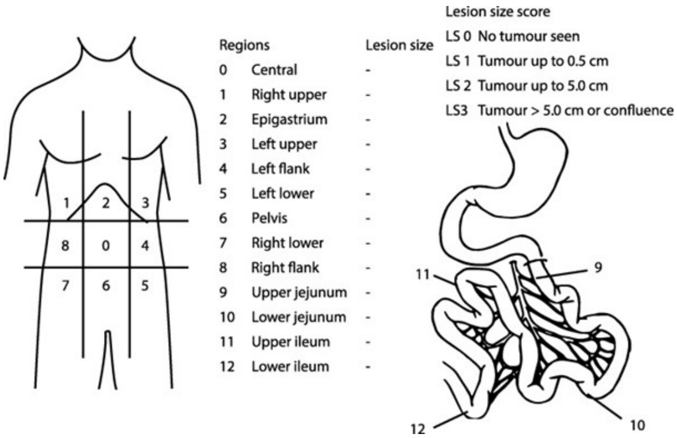

Fig. 1: Illustration of 13 regions partition for PCI scoring system. Obtained from [3]

Although DLS is the current gold standard for preoperative evaluation of PM, it cannot avoid the risks of its invasive procedure, including 1–2% rates for ileus, bleeding, bowel perforation, or infection [4, 5]. Reducing the need for DLS would be highly beneficial, thereby avoiding the aforementioned risks and costs associated with this procedure.

Assessing peritoneal metastasis (PM) on CT scans remains complex, with high interobserver variability even among expert radiologists. Interpreting imaging studies or pathological findings related to peritoneal metastases can be subjective, leading to different opinions between observers. For example, some radiologists may include the stomach and pancreas in Region 2, while others account for them in Region 3. The high interobserver variability also exists when determining the PCI score for each region. Some radiologists base the PCI score on the largest nodule in the region, while others sum the diameters of all nodules in that region for a more comprehensive assessment. Therefore, standardizing the 13 regions partition and assessment methods is crucial to ensure consistency and accuracy from a clinical perspective.

The clinical challenges listed above motivate us to consider using artificial intelligence techniques to facilitate the reporting for radiologists and to diminish the interobserver variability between them. To standardize and improve the determination of the radiological PCI in the assessment of PM, a two-step automated workflow would be desirable: (1) automatic segmentation of the abdominal regions on CT, and (2) automatic detection of PM in CT.

This might seem trivial, however, both still remain challenging tasks. Concerning segmentation tasks, existing research on abdominal region segmentation from CT imaging mainly focuses on single or multiple organs [6–8]. In contrast to organ segmentation, segmenting the 13 regions according to Sugarbaker's Peritoneal Cancer Index (PCI) specification differs. For example, the liver and its surrounding tissue are separated into Region 1 (R1) and Region 2 (R2), while Region 3 (R3) encompasses multiple organs such as the stomach, pancreas, spleen, and surrounding tissues. This presents unique challenges due to the high structural similarity between the abdominal organs, which complicates the precise determination of spatial boundaries and leads to difficulties in accurately segmenting adjacent regions. Furthermore, manual segmentation is a strenuous and time-consuming process, which restricts the availability of training data. This further poses the challenge of potentially influencing the performance of the final segmentation model.

Studies in [10–12] have demonstrated that transfer learning techniques effectively address the challenges posed by limited training data and accelerate the training process in medical imaging. Therefore, to tackle these challenges mentioned in the segmentation task, this study uses transfer learning techniques in addition to the pretrained nnUNet introduced in [14].The pretrained nnUNet is trained on large segmentation datasets, including abdominal organs. It has already acquired sufficient knowledge in segmenting abdominal organs from CT imaging. By employing fine-tuning techniques, this knowledge can be leveraged and adapted to partition the organs and their surrounding tissues into corresponding regions. This approach effectively reduces the difficulty in the determination of spatial boundaries for 13 regions. For convenience, the fine-tuned model is denoted as the Regions-3D-seg model throughout this paper.

The requirement for segmentation accuracy from the clinical perspective is not imposed. Therefore, the primary focus of this work will be on achieving the segmentation of the 13 regions to assist the PCI system in the PM diagnosis. The objective of this study is not oriented toward enhancing segmentation accuracy against the SOTA performance.

In PM nodule detection tasks, the algorithms for this specific task remain limited. In addition to that, PM are rare, especially in non-expert centers, seeing only a few cases per year. Therefore, obtaining a sufficient amount of annotated PM data is challenging.

In the absence of a publicly available dataset specifically tailored for annotated PM nodules, we have opted to transfer this detection task into a classification task. To achieve this, we collected the CT-scan data and its corresponding region-specific PCI scores from the Catharina Cancer Institute (CKI). The Region-3D-seg developed in the previous step is involved in generating the PCI score dataset for the classification task. Subsequently, we utilized this PCI score dataset to fine-tune the pre-trained ResNet50 model with the aim of classifying the PCI score for each region from the CT image.

In Summary, this work has two main objectives. The first goal is to develop and validate a deep-learning model that can automatically segment the abdomen into 13 regions. Secondly, develop a deep learning classifier that can assess the PCI scores of 13 regions from CT images.

II. Related work

A. Medical Image Segmentation

Ronneberger et al. [9] proposed UNet for medical image segmentation in 2015 and outperformed the state-of-the-art (SOTA) at that time by some margin. UNet is a typical encoder and decoder structure. The encoder part is named the contracting path, and the decoder part is called the expanding path. In image classification, it is common to utilize only the contracting part followed by some classifier. In this configuration, the output of the network is a single class label, and the localization information is not preserved. In many biomedical imaging applications, the desired output also includes localization information to create more actionable diagnostic insights. The expanding path combines up-sampling with high-resolution features from the contracting path. This concatenation results in more localized information and is one of the main reasons for the superior performance of U-net. For this reason, UNet architecture attracts the attention of researchers, inspiring them to propose new architectures based on UNet.

In 2018, Isensee et al. [13] proposed nnUNet, a robust adaptive framework incorporating both 2D-UNet and 3D-UNet. The traditional approaches often rely on the expertise of AI professionals, and slight deviations in hyper-parameter settings can significantly impact segmentation performance. Particularly in 3D medical image segmentation, challenges arise from dataset variations and resolution differences, making it difficult to establish an optimal processing pipeline. To address these issues, nnUNet proposed modifications on top of the UNet structure and came up with a method that automatically configures the training process pipeline to a given dataset. The method does not require manual tuning of the hyper-parameter settings, and has demonstrated superior performance against the SOTA methods across various datasets.

In 2023, Wasserthl et al. [14] introduced a robust segmentation model that can segment 104 anatomical structures from CT images. They used a 3D-nnUnet backbone to segment each anatomical structure with an isotropic resolution of $1.5\,\mathrm{mm}$. They combined five 3D-nnUNets, each specialized in a particular segmentation task, encompassing organs, vertebrae, cardiac structures, muscles, and ribs. In this work, we utilize one of the pre-trained 3D-UNet for the organ segmentation task and fine-tune it for 13 regions segmentation tasks.

B. PCI Score Classification

To tackle the challenges associated with deeper neural networks, such as accuracy saturation and training error raising for increasing network depth, He et al. [18] proposed a solution by incorporating residual blocks into the network architecture, leading to the development of Residual Networks (ResNet). Their experiment results have demonstrated that the ResNet is easier to optimize and exhibits improved accuracy with significantly increased depth compared to traditional architectures. Given the characteristics of PM nodules, which are often small and can occur in various locations within the abdomen region,

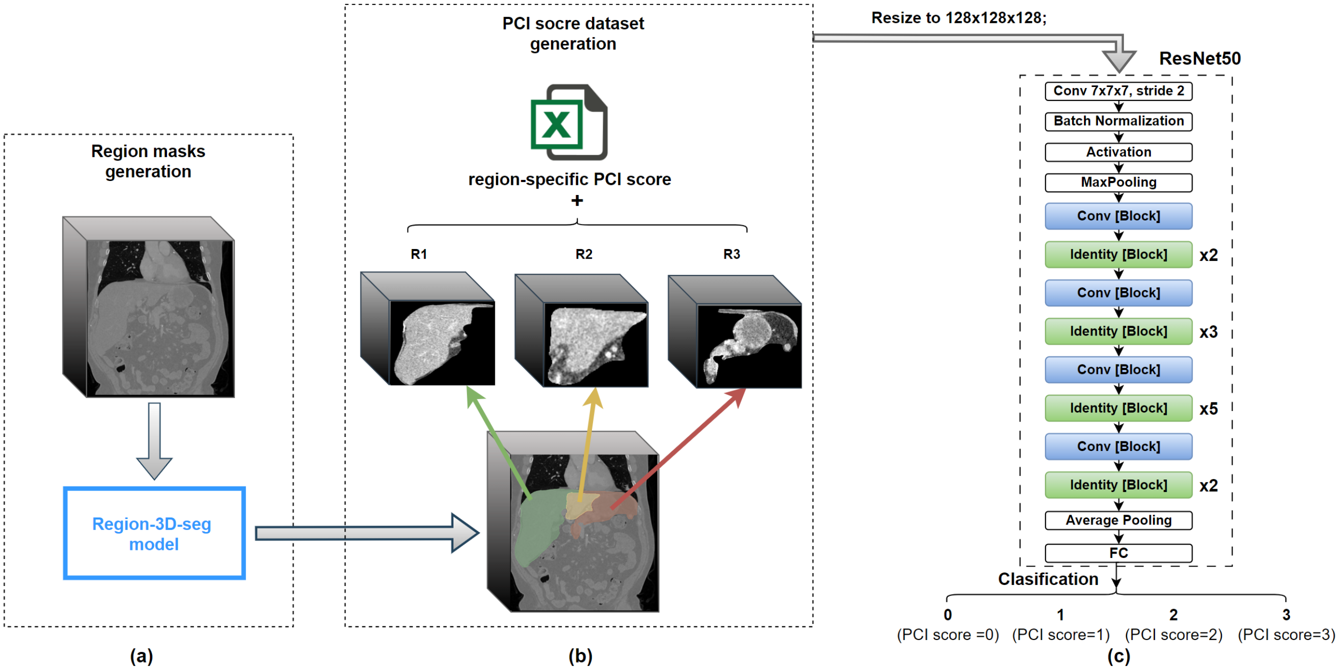

Fig. 2: Workflow of the proposed method. (a) Utilizing the Region-3D-seg model to generate region masks. (b) The CT scans have been cropped based on the masks of R1, R2, and R3 respectively. The cropped 3D image from each region is sent to the ResNet50 as independent input. (c) The overview of the ResNet50 architecture used in this work.

ResNet50 emerges as an optimal choice in terms of the feature extraction capability in medical imaging analysis [19].

In their work presented in [12], Chen et al. have collected many small-scale datasets to construct a comprehensive 3D medical image dataset. Using this dataset, they subsequently trained a ResNet50-based network, named Med3D. This pre-trained model can serve as a foundational model for transfer learning purposes, aiming to accelerate convergence and enhance performance in medical image analysis tasks. To demonstrate its efficiency, the Med3D transfer learning approach has proven effective in tasks such as pulmonary nodule classification, where the model can extract relevant features from medical images and make accurate predictions based on the transferred knowledge. Thus, leveraging the pre-trained weights from Med3D model could provide a good foundation for our model in PCI score classification tasks.

In 2024, Pai et al. [21] introduced a large-scale foundation model refer as fmcib model, which was trained on a comprehensive dataset of 11,467 radiographic lesions. These lesions were annotated from CT images and encompass a diverse range of types, including lung nodules, breast lesions, and perirectal lymph nodes from the deep lesion dataset [20]. The proposed model leverages ResNet50 as its backbone and has demonstrated remarkable effectiveness in classification tasks involving various types of nodules. Given these promising results, we hypothesize that the pre-trained weights from the fmcib model holds the potential for accurately classifying PCI scores in the context of our work.

III. Methods

A. Abdominal Region Segmentation

1) Data Preprocessing

The Hounsfield unit (HU) is used in CT raw data to express the intensity information, reflecting tissue properties [28]. The abdomen region needs to be divided into 13 regions, which include the peritoneum and different organs. The CT-normalization approach within [13] has shown satisfactory performance across diverse datasets related to organ segmentation tasks. Therefore, we will adhere to this convention and apply CT normalization to our dataset. The detailed CT normalization steps are shown as follows:

i. Collect the intensity value across the entire training set's foreground. The foreground is determined by the ground-truth masks.

ii. Compute the mean (μ) and standard deviation (σ) from the collected HU values.

iii. To avoid abnormally large or small isolated HU values occurring in CT images, we clipped out 0.05 and 0.95 percentile from the collected HU values, as shown in Figure 3.

iiii. Apply Z-score normalization as shown in eq (1):

$$Rescaled\ intensity = \frac{x_{clipped} - \mu}{\sigma} \quad (1)$$

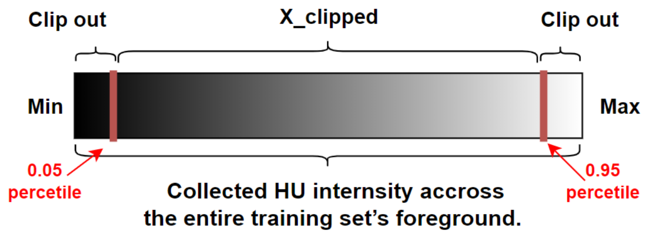
Clip out
X_clipped
Clip out
Min
Max
0.05
percetile
Collected HU internsity accross
the entire training set's foreground.
0.95
percetile

Fig. 3: Illustration of the clipping process for CT-normalization.

2) Model Fine-tuning

The 3D-nnUNet proposed in [13] is chosen as the baseline network since it performs well on the multi-organ segmentation in the abdominal region. The pre-trained weights from [14] are loaded to the baseline model for fine-tune purposes. To fine-tune the 3D-nnUNet, we adapted the architecture by replacing the linear classifier from the original model with a new classifier, which has 3 classes. Subsequently, we performed the linear probing on top of the backbone without freezing any layer.

3) Loss Function

The most commonly used loss functions for segmentation tasks are Dice loss (DL) and Cross-Entropy loss (CE). In this work, the DL and CE are combined as loss functions as shown in eq (2). The CE loss calculates the pixel-wise loss between the predicted results and the ground truth, while the DL aims to measure the similarity over prediction and ground truth segmentation.

$$\mathcal{L}_{total} = \mathcal{L}_{DL} + \mathcal{L}_{CE} \tag{2}$$

The DL function is defined in Eq(3). The A represents the ground truth segmentation in the image and B is the predicted results from the model. The $|A|$ and $|B|$ are separate sets over the corresponding pixels from the images.

$$\mathcal{L}_{DL} = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

The CE in Eq(4) is defined as follows, where y_c is the label of class c, 1 indicates class presents for the particular pixel, and 0 for class not present. The p_c is the predicted probability of a pixel belonging to class c.

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_c \log(p_c) \tag{4}$$

B. PCI Score Classification

1) Proposed Method

Figure 2 shows the workflow of the PCI score classification tasks, which are structured into three steps, denoted as (a) (b), and (c). In step (a), the trained Regions-3D-seg model, developed in this study for the segmentation task, is utilized to produce region masks from CT images. Subsequently, based on these masks, three regions are cropped out from the original image. Then followed by step (2), the masks are utilized to extract the contents of each region, with the background intensity set to 0 value. The PCI scores from each region serve as the labels, and along with the cropped-out regions, it forms the PCI score dataset. In step (c), each processed 3D region block is then resized to a spatial dimension of 128 x 128 x 128 voxels. These 3D regions are individually processed through a ResNet50-based classifier to classify them into 4 classes: 0, 1, 2, and 3, representing the specific PCI score for each respective region.

2) ResNet50 Architecture and Model Fine-tune

The architecture of ResNet50 employed in this work is shown in Figure2 (c), it contains 2 types of shortcut modules: convolutional block and identity block. In the identity block, there is no convolutional layer in the shortcut path, thus, the output dimension is the same as the input. For the convolutional block, the convolutional layer is added to the shortcut path, which results in the output dimension being larger than the input. These modules collectively establish the 16 processing blocks within the ResNet50. This design makes it possible to reduce the number of parameters while preserving model performance [18].

Pre-trained weights from Med3D [12] and fmcib [21] are individually loaded to the ResNet50 model for fine-tuning, with no frozen layers. The fully connected and final classification layers were then replaced with a new layer to classify the input 3D image into 4 classes.

3) Data Preprocessing

Following the suggestions from the radiologist at CKI, the HU values ranging from [-200, 300] were chosen and then normalized to range [0,1]. The spatial dimension of each cropped-out region varies from 172 x 184 x 95 voxels to 306 x 308 x 347 voxels. To make the input data of uniform size, we resize all of the data into the average shape across the entire dataset, resulting in the shape of 128 x 128 x 128 voxels.

IV. Experiments

A. Abdominal Region Segmentation

1) CKI Dataset

The manual annotation work is conducted by the Catharina Cancer Institute (CKI), thus, this dataset is named after this institution. The abdominal area is subdivided into 13 regions according to clinical specification [2], Figure 4 illustrates the 13 regions partition for the abdominal area. Two PhD candidates from CKI annotated 60 CT scans for 13 regions segmentation by using 3D Slicer software (Version 5.4.0). At the current stage, the CKI delivered 60 annotated CT scans for 3 regions (R1, R2, R3). 50 cases are randomly chosen as the training set and the rest of the 10 cases are for the validation set.

The delivered data is the 3D medical image and the spacing of each sample varies between $0.45\,\mathrm{mm}$ and $1\,\mathrm{mm}$. To

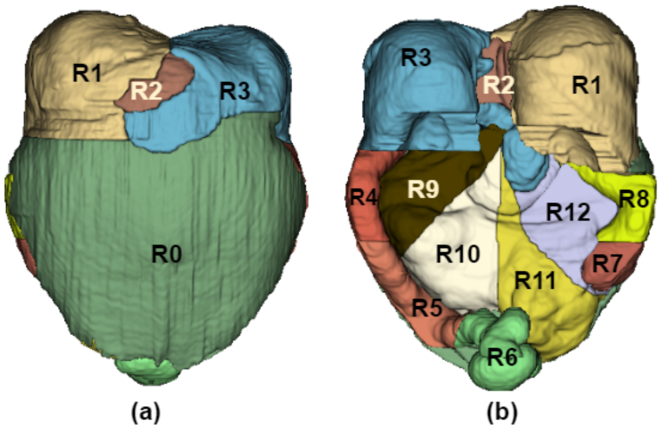

Fig. 4: Illustration of 13 regions partition for the abdominal area. The regions are arranged from Region 0 (R0) to Region 12 (R12). (a) The front view shows the region partition of R0, R1, R2, and R3. (b) The back view presents the partition of the region from R1 to R12 except R0.

enhance consistency and facilitate the comparative analysis of our experiments, we employ 2 resampling strategies:

- Strategy 1: Resampling data by using the average voxel spacing calculated across the entire dataset. The resultant value is set to $0.7 \times 0.7 \times\ 0.45\,\mathrm{mm}$.
- Strategy 2: Resampling data into $1.5\,\mathrm{mm}$ spacing to align with the pre-trained model, originally trained on data with $1.5\,\mathrm{mm}$ voxel spacing.

For convenience to discuss the experiment results in this study, we further labeled those 2 resampling strategies: R_{avg} for strategy 1 and $R_{1.5mm}$ for strategy 2.

2) *Dataset Augmentation*

nnUNET has provided an option to implement customer data augmentation functions by using batchgenerators [26], a framework for medical image augmentation developed by the authors of nnUNET as well. We have applied data augmentations as discussed in [13], with 6 augmentations selected out of the total available. This is because too extensive data augmentations may over complicate the training processes, which could lead to underfitting. The corresponding configurations for selected data augmentations are listed below:

- Random contrast adjustment (CA): Contrast range $r = [0.5, 2.0]$; probability $p = 0.15$.
- Random gaussian blur (GB): Blur sigma $\sigma = [0.5, 1.0]$; $p = 0.20$.
- Random Gaussian noise (GN): Noise variance $\sigma^2 = [0.0, 0.1]$; $p = 0.10$.
- Random rotation (R): Rotation angle $\alpha = [-30, 30]$ along x, y, z orientation; $p = 0.20$.
- Random scale (S): Scale range $s = [0.7, 1.7]$; $p = 0.20$.
- Random simulation low resolution (SLR): Zoom range $SLR\_s = [0.5, 1.0]$; Linear interpolate mode for downsampling ; Nearest Neighbor interpolate mode for upsampling; $p = 0.25$.

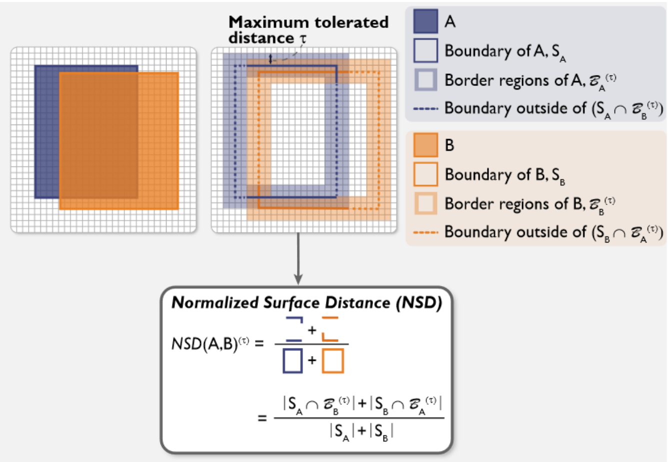
Maximum tolerated distance τ
A
Boundary of A, S_A
Border regions of A, $\mathcal{B}_A^{(\tau)}$
Boundary outside of $(S_A \cap \mathcal{B}_B^{(\tau)})$
B
Boundary of B, S_B
Border regions of B, $\mathcal{B}_B^{(\tau)}$
Boundary outside of $(S_B \cap \mathcal{B}_A^{(\tau)})$
Normalized Surface Distance (NSD)
$NSD(A,B)^{(\tau)} = \frac{|S_A \cap \mathcal{B}_B^{(\tau)}| + |S_B \cap \mathcal{B}_A^{(\tau)}|}{|S_A| + |S_B|}$

Fig. 5: Illustration of the normalized surface distance computation method. Obtained from [24]

To explore the optimal combination of augmentations for model training, we performed an ablation study. The combination of R, S, SLR, and CA produced superior results, as detailed in Section IV-A6.

3) *Evaluation metrics*

To evaluate the performance of Region-3D-seg model comprehensively, we employed 3 evaluation metrics: Dice Similarity Coefficient (DSC), Normalized surface distance (NSD) [23], and Hausdorff distance of 95 percentile (HD95) [22].

The DSC aims to measure the similarity between the predicted segmentation and the ground truth. The nnUNet employed the DSC as their evaluation standard. To ensure a fair comparison of experimental results, we follow this convention and conduct experiments with the DSC evaluation metrics.

Figure 5 shows the computation method for the NSD. It measures the overlap boundary area between prediction and target masks. The boundary area is defined by the threshold

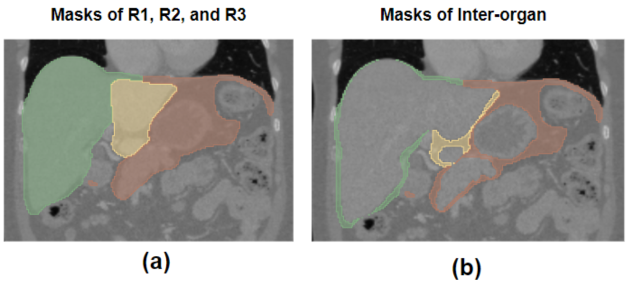

Fig. 6: (a) Masks for three regions are represented as follows: green for R1, yellow for R2, and red for R3. (b) Inter-organ masks are derived by subtracting the organ mask from the region masks. Specifically, the liver mask is subtracted from both R1 and R2, while the mask containing the stomach, pancreas, and spleen is removed from R3.

parameter τ in mm. The selection of τ is determined based on the domain-related requirements. In the study, a value of $\tau = 1\,\mathrm{mm}$ is adopted following the feedback from the clinicians at CKI.

Hausdorff Distance (HD) [22] calculates the maximum of all shortest distances for all points between the ground truth mask boundary and the prediction mask boundary. However, it is challenging to visually identify a single erroneously annotated pixel during the manual annotation process, which may significantly impact the performance of HD. HD95, a variant of HD, calculates the 95th percentile instead of the maximum. This approach can effectively disregard outliers, thus, HD95 is chosen as an evaluation metric in this work.

The metrics of DSC and NSD are aimed at measuring the overlapped area between prediction and ground truth. Therefore, the higher outcome represents the better-overlapped status, making higher results preferable. Conversely, HD95 measures the distance between the borders of predictions and ground truth. A lower result indicates closer proximity, making lower values desirable.

The model from [14] is applied to generate organ masks from CKI dataset. Next, we obtain the inter-organ mask by using the region masks and subtracting the organ masks. The resulting mask is presented in the Figure 6 (b). To assess the Region-3D-sge model's effectiveness in segmenting non-organ tissues, we applied the three metrics mentioned above to validate the inter-organ masks from both prediction and ground truth. Additionally, the mean DSC over three regions is computed for reference purposes.

4) Implementation

We employ the SGD optimizer with a momentum of 0.99 and weight decay of $3e-5$. Poly learning scheduler is utilized with an initial learning rate of $1e-2$. The He-initialization is employed when training from scratch. We train models with 500 epochs and evaluate after every epoch, the best results are presented in Table I.

5) Experiments Results

The experimental results are divided into quantitative and qualitative parts.

Quantitative results:

The quantitative experiment results can be found in Table I, where we train the baseline model with two different data resampling strategies, $R_{1.5mm}$ and R_{avg} respectively. We then compare these two baselines with different configuration options.

As shown in Table I, $BL\_R_{1.5mm}$ experiences a slight improvement in overall metrics with the addition of the data augmentation (DA) during the training process. However, employing DA alongside fine-tuning techniques does not yield further enhancement in model performance. This is because larger isotropic voxel spacing often contains less information for the model to learn from, resulting in challenges in accurately segmenting complex anatomical structures.

	mean DSC ↑ over 3 regions	**Inter-organ (mean)**		
		DSC↑	**NSD↑**	**HD95↓**
$BL\_R_{1.5mm}$	91.86%	82.33%	67.53%	6.44mm
+DA	92.24%	82.86%	67.82%	6.51mm
+DA+Fine-tune	92.23%	83.01%	68.04%	6.44mm
BL_R_{avg}	91.23%	80.58%	73.66%	7.46mm
+DA	91.77%	82.11%	76.00%	6.75mm
+DA+Fine-tune	**92.43%**	**83.46%**	**77.25%**	**6.07mm**

TABLE I: Experiment results from different train configurations. $BL\_R_{1.5mm}$ denotes the baseline trained from scratch with $R_{1.5mm}$ data resampling strategy. BL_R_{avg} for baseline model trained from scratch with R_{avg} data resampling strategy. The '+DA' means applying the CA, R, S, and SLR data augmentation during the training process. The data augmentation selection is determined through the ablation study. The '+Fine-tune' indicates fine-tuning the model by loading pre-trained weights from the totalsegmentator. The symbol ↑ indicates that higher experiment results are preferable, whereas ↓ indicates that lower values are desirable.

For BL_R_{avg}, the addition of the DA method improves the performance by +1.53%, and +2.43% for inter-organ DSC and inter-organ NSD respectively. Employing DA alongside fine-tuning techniques further enhances the performance of segmentation results, achieving 83.46% (+2.88%) inter-organ DSC and 77.25% (+3.59%) inter-organ NSD as compared to BL_R_{avg}. We believe this is attributed to the high-resolution data providing a sufficient amount of fine detail for the model to learn. Additionally, the 3D-UNet demonstrates its capability to handle this rich and complex anatomical structure information, thereby adapting it effectively to partition abdominal regions with precision. For this reason, the BL_R_{avg} model combined with DA and fine-tuning techniques denotes the Region-3D-seg model which is utilized to generate the PCI score dataset in classification tasks. For this reason, we denote the BL_R_{avg} model, combined with DA and fine-tuning techniques, as the Region-3D-seg model. This model is utilized to generate the PCI score dataset for classification tasks.

Qualitative results:

Figure 7 illustrates case ID 170 as a good segmentation result and case ID 138 as a comparatively poor outcome. Both results are obtained by using the BL_R_{avg} model together with DA and fine-tune techniques since it drives the best performance in Quantitative results. As shown in the visualization results from case ID 170, the prediction result is accurate for all three regions, even though each region has some irregular edges. We believe this is attributed to the powerful capacity of the 3D-UNet and the current conventional kernel is good enough to handle this complex segmentation shape.

However, in the prediction result from case ID 138, the model struggled to segment the inter-organ region precisely. As we can see from Error R1 and Error R3 in caseID 138, the mismatched areas within the blue dashed circles, it is noticeable that the model struggles with accurately segmenting the border of R1 and R3 particularly when they intersect

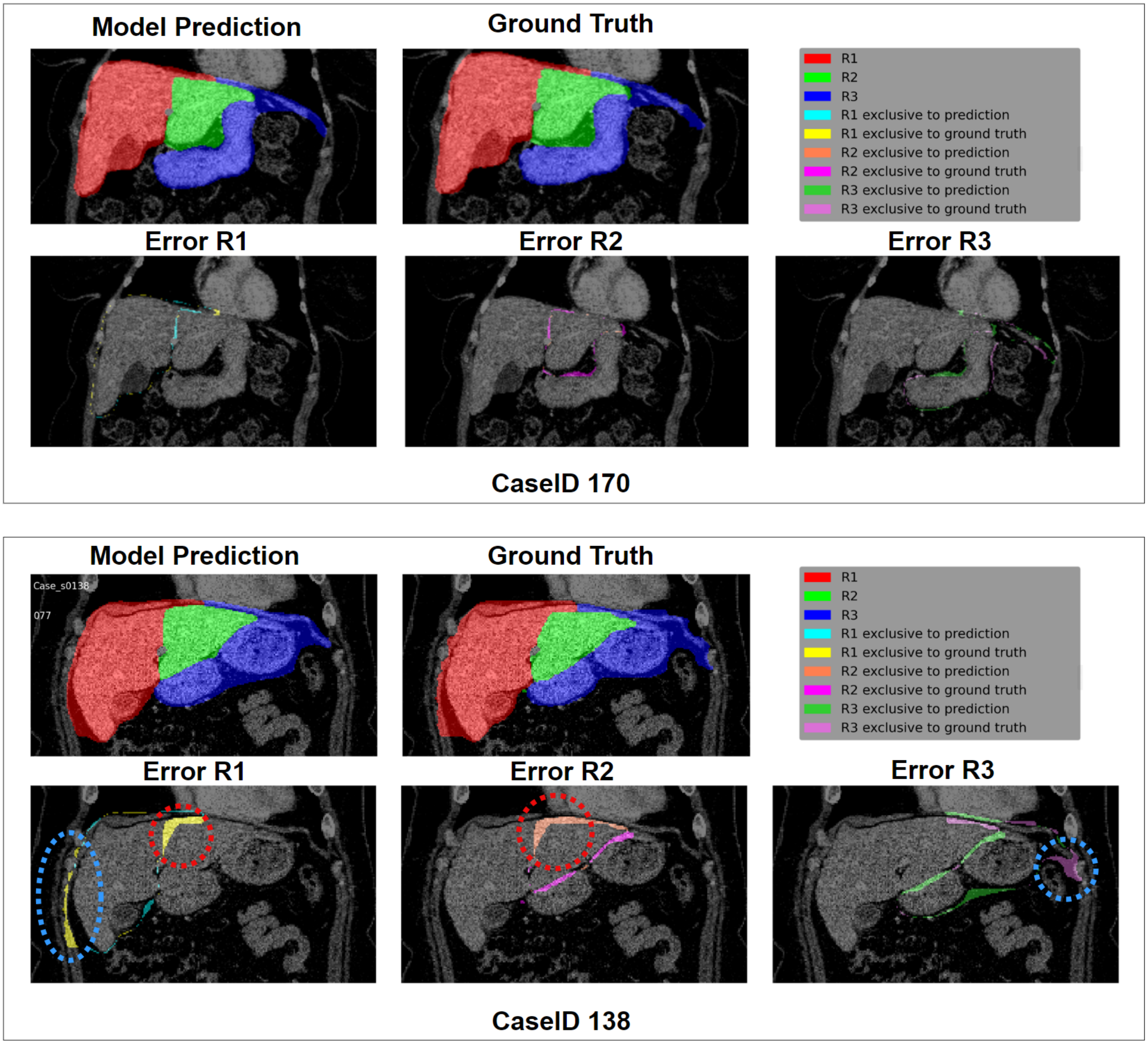
Model Prediction
Ground Truth
R1
R2
R3
R1 exclusive to prediction
R1 exclusive to ground truth
R2 exclusive to prediction
R2 exclusive to ground truth
R3 exclusive to prediction
R3 exclusive to ground truth
Error R1
Error R2
Error R3
CaseID 170
Model Prediction
Case_s0138
077
Ground Truth
R1
R2
R3
R1 exclusive to prediction
R1 exclusive to ground truth
R2 exclusive to prediction
R2 exclusive to ground truth
R3 exclusive to prediction
R3 exclusive to ground truth
Error R1
Error R2
Error R3
CaseID 138

Fig. 7: Visualization of good segmentation results from Case ID 170 (top) and bad segmentation results of Case ID 138 (bottom). Inter-organ dice for each region of case 170, R1=92.29%, R2=85.33%, R3=85.48%; For case 138, R1=85.05%, R2=73.96%, R3=84.11%.

with the abdominal wall. We believe this can be attributed to regions' borders contributing less to the loss magnitude. To address this issue, employing the inter-organ mask as a weighted map to the loss function can be a solution. This approach would raise the loss magnitude associated with the regions' edges, thereby facilitating more effective learning of border information by the model.

Moreover, in the Error R1 and Error R2 from case 138, mismatches between prediction and ground truth were observed in the yellow and coral areas highlighted within the red dashed circles. These discrepancies occurred because the annotators from CKI annotated these regions differently, causing confusion for the model in identifying these specific regions. To diminish these discrepancies, the elastic deformation augmentation [27] can be employed in the training process to enhance the robustness of the model. However, the augmentation method needs to be carefully experimented with to validate its effectiveness, since too extensive augmentation might downgrade the clinical suitability of the model. Therefore, in future work, we wish to conduct an ablation study regarding the elastic deformation method to explore its efficiency.

Augmentations	DSC↑ (%)	Inter-organ DSC ↑(%)	HD95↓ (mm)
NoDA	91.63	81.47	8.23
CA	91.60	**81.58**	8.37
GB	91.53	**81.49**	8.26
GN	91.45	81.37	8.52
R	**91.77**	**81.89**	**8.02**
S	**91.84**	**81.96**	**7.95**
SLR	91.58	**81.72**	**8.11**

TABLE II: Ablation study of data augmentation. 'NoDA' denotes no data augmentation applied during training. For the remaining experiments, only the specified data augmentation method was enabled during training. The symbol ↑ next to DSC and inter-organ DSC indicates that higher experiment results are preferable, whereas ↓ next to HD95 indicates that lower values are preferable. The values highlighted in bold indicate that adding the particular data augmentation during training yields better performance than the NoDA results.

R	S	SLR	CA	GB	DSC↑(%)	Inter-organ DSC ↑(%)	HD95↓ (mm)
✓	✓	✓	✗	✗	91.93	82.04	*8.07*
✓	✓	✓	✓	✗	***92.11***	***82.32***	**7.62**
✓	✓	✓	✓	✓	*91.81*	*81.82*	8.59

TABLE III: Ablation study of data augmentation combination.

6) Ablation Study

All experiments in the ablation study are performed on $BL\_R_{1.5mm}$ with 200 epochs. In table II, we evaluate the effectiveness of selected data augmentation methods in training $BL\_R_{1.5mm}$ model. The experiment includes results for 'NoDA' as a reference. The results highlighted in bold indicate that particular data augmentation drives better results compared to NoDA. The R, S, and SLR augmentations gained better performance in DSC, inter-organ DSC, and HD95. CA and GB drive better results in terms of inter-organ DSC. However, the GN downgrades the performance overall metrics. For this reason, we further conduct an ablation study for augmentation combination without involving the GN.

The ablation study of augmentations combination is shown in Tabel III, the combination of R, S, SLR, and CA yields the best outcome across three evaluation metrics.

B. PCI Score Classification

1) PCI score Dataset

To compose the PCI score dataset, we collected 239 CT scans from CKI, each containing PCI scores for 13 abdominal regions assessed by radiologists and surgeons across various institutions. There are 12 CT scans have been excluded due to lacking region-specific PCI scores. As a result, 227 CT scans are selected for the final PCI score dataset. For the current stage, the Region-3D-seg model can only segment 3 regions (R1, R2, R3), thus, we make use of those masks to extract 3 regions from each CT scan, which results in a total of 681 data samples, together with the corresponding region-specific PCI score as labels, forms the final PCI score dataset. After a thorough review of the dataset, we have noticed that the distribution of 4 classes is imbalanced. Therefore, the validation set was manually arranged to ensure all classes were distributed equally, and the remaining data were allocated to the training set. The statistics of the arranged training set and validation set are exhibited in table IV. Furthermore, the three regions extracted from the same CT scan were grouped together, in other words, the cropped R1, R2, and R3 from a given CT image were kept together in either the training or validation set.

2) Data Augmentation

Data augmentation is conducted using the MONAI framework [25], a Python library specialized for medical AI. Due to the limited training data availability, which restricts the number of training epochs. We selected only 3 data augmentations to not over-complicate the training process. These augmentations include random contrast adjustment with a gamma range of [0.9, 1.2], random rotation along the x, y, and z orientation with an angle range of [-15°, 15°], and random flip along x, y, z direction. As mentioned in the section IV-B1, the number of training samples of classes 1, 2, and 3 is way less than class 0. To address this issue, we employed a Weighted Random Sampler to reuse the training sample if the class is not 0 during the training process. Since the non-zero class will be reused during training, we assigned relatively high probabilities for each augmentation method. Specifically, a probability of 0.5 was set for random contrast adjustment, while a probability of 0.9 was applied to both random rotation and random flipping.

3) Evaluation metrics

The Precision, Recall, Accuracy, and Confusion Matrix are determined to evaluate the performance of the fine-tuned ResNet50 model. These metrics are computed based on four terminologies: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The Accuracy provides an overall assessment of model performance but cannot offer detailed information such as the cost associated with different errors. Thus, Precision and Recall are calculated for each class to provide detailed insights, along with the confusion matrix, enabling a comprehensive evaluation of model performance.

The equation of precision is shown in Eq (5), which

Class label	# of samples train set	% train set	# of samples validation set	% validation set
0	477	74.65%	11	26.19%
1	46	7.20%	9	21.43%
2	44	6.89%	13	30.95%
3	72	11.27%	9	21.43%
Total	639	100%	42	100%

TABLE IV: Statistics of PCI score dataset.

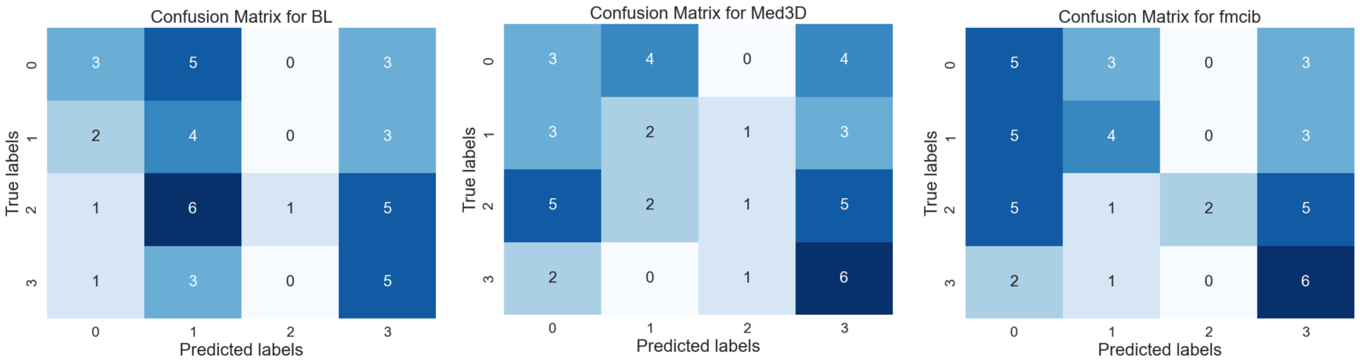

Fig. 8: Confusion matrix for BL, Med3D and fmcib respectively.

indicates the number of corrected predicted results out of the total prediction results.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall is aimed to measure the correctly identified class out of all positive results.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Accuracy measures the ratio of correctly classified cases out of the total number of samples in the dataset.

$$Accuracy = \frac{\text{\# of correct predictions}}{\text{Total \# of predicted samples}} \tag{7}$$

The Confusion Matrix presents the prediction results for all classes against their ground truth. It intuitively illustrates the model's classification ability for each class.

4) Implementation

We employ the AdamW optimizer [29] with weight decay of 1e-7. Poly learning scheduler is utilized with an initial learning rate of 8e-5. The He-initialization is used when training from scratch. To address the class-imbalanced distribution issue for the PCI score dataset, the weight random sampler is employed to sample from the PCI score dataset, such that the model sees approximately each class the same number of times. All models were trained for 50 epochs using batch size of 16, and evaluated after every epoch, the training results are presented in Table V.

5) Experiment Results

The experimental results, as presented in Table V, show that the performance of BL, Med3D, and fmcib falls considerably below expectations. Notably, Med3D performs even worse than BL, likely due to most of its training data focusing on the organ and tumor segmentation tasks. This presents challenges in identifying small nodules without specific annotations. Conversely, fmcib demonstrates better accuracy and recall, possibly because it is trained on datasets containing various lesion types, thereby acquiring better knowledge in identifying small lesions from medical images.

	Accuracy	Class	Precision	Recall
BL	0.31	0	0.43	0.27
		1	0.22	*0.44*
		2	1.00	0.08
		3	0.31	*0.56*
Med3D	0.29	0	0.23	0.27
		1	0.25	0.22
		2	0.33	0.08
		3	0.33	0.67
fmcib	0.4	0	0.36	0.45
		1	0.44	0.44
		2	1.00	0.15
		3	0.35	0.67

TABLE V: Experiment results. BL for base-line ResNet50 model trained from scratch. Med3D for loading the pre-trained weight from Med3D to BL. The fmcib means loading the pre-trained weights from the fmcib model.

However, upon reviewing the experimental results, it is obvious that the models struggle to learn from the data effectively. This can be attributed to the inconsistency of the PCI score dataset generated from Region-3D-seg model. As observed in the experimental results for the Region-3D-seg model, it faces challenges in segmenting the borders of the regions, where PM nodules may potentially exist. As a consequence, this leads to mismatches between the generated dataset and the PCI score labels. Moreover, the different PCI score assessment criteria among the radiologists from different institutions further complicate the classification tasks. Addressing this issue requires radiologists to reassess the PCI score for the generated dataset. In addition to that, the PM nodules are often tiny, making them challenging to identify. Applying filtering techniques such as wavelet transform [30], to preprocess the data to enhance the PM nodule visibility, thereby, facilitating the model to identify the PM nodules.

As shown in the confusion matrix depicted in Figure 8, all the models exhibit confusion in classifying classes 0, 1, and 3. For class 2, the model does not misclassify it with the other 3 classes, which means the class boundary between class 2 and the other 3 classes is clear. However, the TP for class 2 indicates that this class is relatively more challenging to classify than the other three classes. This is due to the fact that the training samples from class 2 only occupy 6.89% out of the total. This highlights the need for an improved data augmentation strategy to enhance the model's performance.

In future work, we would like to examine different data augmentation methods and filtering techniques to validate their effectiveness for training the model.

V. Conclusion

A. Abdominal Region Segmentation

In this study, we fine-tuned a pre-train nnUNET to achieve segmentation of 3 regions within the abdominal area, based on the region partition criteria for the Sugarbaker's Peritoneal Cancer Index. Through an ablation study, we identified an optimal combination of data augmentation. We comprehensively evaluated the model performance involving overlap metrics (DSC, NSD) and distance-based metric (HD95). Finally, our experiment results demonstrate that the BL_R_{avg} model, combined with selected data augmentation and fine-tuning techniques, yielded promising results: 83.46% for inter-organ DSC, 77.25% for inter-organ NSD, and an inter-organ HD95 of $6.07\,\mathrm{mm}$.

B. PCI Score Classification

In this work, faced with the absence of a publicly available dataset specifically annotated for PM nodules, we opted to transform the detection task into a classification task. To achieve this, we collected raw CT data from CKI and utilized the Region-3D-seg model, originally developed for Abdominal Region Segmentation, to generate region masks and extract region blocks. The PCI scores from each region served as labels, along with the cropped-out regions, constitute the PCI score dataset for the classification task. Subsequently, we fine-tuned the ResNet50 model using this dataset, with pre-trained weights from Med3D and fmcib. The experiment results show that the proposed method is a naive solution for PCI score classification tasks. This work only provides a benchmark dataset and method for this question. In future work, we wish to employ more advanced data augmentation methods and filtering techniques to improve the classification performance for this problem.

References

[1] Cortés-Guiral D, Hübner M, Alyami M, Bhatt A, Ceelen W, Glehen O, Lordick F, Ramsay R, Sgarbura O, Van Der Speeten K, Turaga KK, Chand M., "Primary and metastatic peritoneal surface malignancies," Nat Rev Dis Primers. 2021 Dec 16;7(1):91. doi: 10.1038/s41572-021-00326-6. PMID: 34916522.

[2] Jacquet, P., Sugarbaker, P.H. (1996). Clinical research methodologies in diagnosis and staging of patients with peritoneal carcinomatosis. Cancer Treatment and Research, vol 82. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-1247-5_23

[3] Leimkühler M, de Haas RJ, Pol VEH, Hemmer PHJ, Been LB, van Ginkel RJ, Kruijff S, de Bock GH, van Leeuwen BL. Adding diagnostic laparoscopy to computed tomography for the evaluation of peritoneal metastases in patients with colorectal cancer: A retrospective cohort study. Surg Oncol. 2020 Jun;33:135-140. doi: 10.1016/j.suronc.2020.02.010. Epub 2020 Feb 15. PMID: 32561078.

[4] Lemmens VE, Bosscha K, van der Schelling G, Brenninkmeijer S, Coebergh JW, de Hingh IH. Improving outcome for patients with pancreatic cancer through centralization. Br J Surg. 2011 Oct;98(10):1455-62. doi: 10.1002/bjs.7581. Epub 2011 Jun 29. PMID: 21717423.

[5] Chua TC, Moran BJ, Sugarbaker PH, Levine EA, Glehen O, Gilly FN, Baratti D, Deraco M, Elias D, Sardi A, Liauw W, Yan TD, Barrios P, Gómez Portilla A, de Hingh IH, Ceelen WP, Pelz JO, Piso P, González-Moreno S, Van Der Speeten K, Morris DL. Early- and long-term outcome data of patients with pseudomyxoma peritonei from appendiceal origin treated by a strategy of cytoreductive surgery and hyperthermic intraperitoneal chemotherapy. J Clin Oncol. 2012 Jul 10;30(20):2449-56. doi: 10.1200/JCO.2011.39.7166. Epub 2012 May 21. PMID: 22614976.

[6] P. G. Rózsa, M. Kovács and S. Nagy, "On Applying Expert Knowledge for Spline-Based Segmentation of Liver on Computed Tomography Images," 2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), Nis, Serbia, 2023, pp. 21-24, doi: 10.1109/ICEST58410.2023.10187395.

[7] F. Ferdinandus, E. I. Setiawan, E. M. Yuniarno and M. H. Purnomo, "Lung Segmentation using MultiResUNet CNN based on Computed Tomography Image," 2022 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 2022, pp. 1-6, doi: 10.1109/ISITIA56226.2022.9855353.

[8] J. Hu and K. Wang, "Abdominal Multi-Organ Segmentation Based on nnUNet," 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2023, pp. 2026-2030, doi: 10.1109/ICSP58490.2023.10248829.

[9] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arXiv:1505.04597. https://doi.org/10.48550/arXiv.1505.04597.

[10] Raghu, Maithra et al. "Transfusion: Understanding Transfer Learning for Medical Imaging." Neural Information Processing Systems (2019).

[11] Kim, H.E., Cosa-Linan, A., Santhanam, N. et al. Transfer learning for medical image classification: a literature review. BMC Med Imaging 22, 69 (2022). https://doi.org/10.1186/s12880-022-00793-7.

[12] Chen, S., Ma, K., Zheng, Y. (2019). Med3D: Transfer Learning for 3D Medical Image Analysis. ArXiv, abs/1904.00625.

[13] Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. nnU-Net: a self-configuring method for deep learning-based biomed-

ical image segmentation. Nat Methods 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

[14] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence. https://doi.org/10.1148/ryai.230024

[15] J. Hu and K. Wang, ”Abdominal Multi-Organ Segmentation Based on nnUNet,” 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2023, pp. 2026-2030, doi: 10.1109/ICSP58490.2023.10248829.

[16] Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

[17] Isensee, F. et al. (2023) 'Extending NNU-net is all you need', Informatik aktuell, pp. 12–17. doi:10.1007/978-3-658-41657-7_7.

[18] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[19] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, Pinkie Anggia, Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer, Procedia Computer Science, Volume 179, 2021, Pages 423-431, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.025.

[20] Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging (Bellingham). 2018 Jul;5(3):036501. doi: 10.1117/1.JMI.5.3.036501. Epub 2018 Jul 20. PMID: 30035154; PMCID: PMC6052252.

[21] Pai, S., Bontempi, D., Hadzic, I. et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell 6, 354–367 (2024). https://doi.org/10.1038/s42256-024-00807-9

[22] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. 1993. Comparing images using the Hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence 15, 9 (1993), 850–863.

[23] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, Patel Y, Meyer C, Askham H, Romera-Paredes B, Kelly C, Karthikesalingam A, Chu C, Carnell D, Boon C, D'Souza D, Moinuddin SA, Garie B, McQuinlan Y, Ireland S, Hampton K, Fuller K, Montgomery H, Rees G, Suleyman M, Back T, Hughes CO, Ledsam JR, Ronneberger O. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res. 2021 Jul 12;23(7):e26151. doi: 10.2196/26151. PMID: 34255661; PMCID: PMC8314151.

[24] Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Radsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., Farahani, K., Glocker, B., Heckmann-Notzel, D., Isensee, F., Jannin, P., Kahn, C.E., Kleesiek, J., Kurç, T.M., Kozubek, M., Landman, B.A., Litjens, G.J., Maier-Hein, K.H., Menze, B.H., Muller, H., Petersen, J., Reyes, M., Rieke, N., Stieltjes, B., Summers, R.M., Tsaftaris, S.A., Ginneken, B.V., Kopp-Schneider, A., Jager, P.F., & Maier-Hein, L. (2021). Common Limitations of Image Processing Metrics: A Picture Story. ArXiv, abs/2104.05642.

[25] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murray, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Zalbagi Darestani, M., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B. S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P. F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L. A., Roth, H. R., Xu, D., Bericat, D., Floca, R., Zhou, S. K., Shuaib, H., Farahani, K., Maier-Hein, K. H., Aylward, S., Dogra, P., Ourselin, S., Feng, A. (2022). MONAI: An open-source framework for deep learning in healthcare. https://doi.org/https://doi.org/10.48550/arXiv.2211.02701

[26] Isensee Fabian, Jäger Paul, Wasserthal Jakob, Zimmerer David, Petersen Jens, Kohl Simon, Schock Justus, Klein Andre, Roß Tobias, Wirkert Sebastian, Neher Peter, Dinkelacker Stefan, Köhler Gregor, Maier-Hein Klaus (2020). batchgenerators - a python framework for data augmentation. doi:10.5281/zenodo.3632567.

[27] Bar-David D, Bar-David L, Shapira Y, Leibu R, Dori D, Gebara A, Schneor R, Fischer A, Soudry S. Elastic Deformation of Optical Coherence Tomography Images of Diabetic Macular Edema for Deep-Learning Models Training: How Far to Go? IEEE J Transl Eng Health Med. 2023 Jul 24;11:487-494. doi: 10.1109/JTEHM.2023.3294904. PMID: 37817823; PMCID: PMC10561735.

[28] Razi T, Niknami M, Alavi Ghazani F. Relationship between Hounsfield Unit in CT Scan and Gray Scale in CBCT. J Dent Res Dent Clin Dent Prospects. 2014 Spring;8(2):107-10. doi: 10.5681/joddd.2014.019. Epub 2014 Jun 11. PMID: 25093055; PMCID: PMC4120902.

[29] Loshchilov, Ilya and Frank Hutter. “Decoupled Weight Decay Regularization.” International Conference on Learning Representations (2017).

[30] Guo X, Liu X, Wang H, Liang Z, Wu W, He Q, Li K, Wang W. Enhanced CT images by the wavelet transform improving diagnostic accuracy of chest nodules. J Digit Imaging. 2011 Feb;24(1):44-9. doi: 10.1007/s10278-009-9248-y. Epub 2009 Nov 24. PMID: 19937084; PMCID: PMC3046798.