Computerized Medical Imaging and Graphics 117 (2024) 102434

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Main challenges on the curation of large scale datasets for pancreas segmentation using deep learning in multi-phase CT scans: Focus on cardinality, manual refinement, and annotation quality

Matteo Cavicchioli [a,b,*], Andrea Moglia [a], Ludovica Pierelli [b], Giacomo Pugliese [b,1], Pietro Cerveri [a,c,1]

[a] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy
[b] Fondazione MIAS (AIMS Academy), Piazza dell'Ospedale Maggiore 3, Milano, 20162, Italy
[c] Department of Industrial and Information Engineering, University of Pavia, Via Adolfo Ferrata 5, Pavia, 27100, Italy

ARTICLE INFO

Dataset link: https://github.com/hal9000-lab/Pancreas_2.5D

Keywords:
Pancreas segmentation
Deep learning pancreas segmentation
Artificial intelligence surgery
Artificial intelligence surgical planning
Pancreas dataset
Medical imaging dataset acquisition
Medical imaging dataset curation

ABSTRACT

Accurate segmentation of the pancreas in computed tomography (CT) holds paramount importance in diagnostics, surgical planning, and interventions. Recent studies have proposed supervised deep-learning models for segmentation, but their efficacy relies on the quality and quantity of the training data. Most of such works employed small-scale public datasets, without proving the efficacy of generalization to external datasets. This study explored the optimization of pancreas segmentation accuracy by pinpointing the ideal dataset size, understanding resource implications, examining manual refinement impact, and assessing the influence of anatomical subregions. We present the AIMS-1300 dataset encompassing 1,300 CT scans. Its manual annotation by medical experts required 938 h. A 2.5D UNet was implemented to assess the impact of training sample size on segmentation accuracy by partitioning the original AIMS-1300 dataset into 11 smaller subsets of progressively increasing numerosity. The findings revealed that training sets exceeding 440 CTs did not lead to better segmentation performance. In contrast, nnU-Net and UNet with Attention Gate reached a plateau for 585 CTs. Tests on generalization on the publicly available AMOS-CT dataset confirmed this outcome. As the size of the partition of the AIMS-1300 training set increases, the number of error slices decreases, reaching a minimum with 730 and 440 CTs, for AIMS-1300 and AMOS-CT datasets, respectively. Segmentation metrics on the AIMS-1300 and AMOS-CT datasets improved more on the head than the body and tail of the pancreas as the dataset size increased. By carefully considering the task and the characteristics of the available data, researchers can develop deep learning models without sacrificing performance even with limited data. This could accelerate developing and deploying artificial intelligence tools for pancreas surgery and other surgical data science applications.

1. Introduction

Despite recent diagnostic and treatment advancements, the five-year survival rate of pancreatic ductal adenocarcinoma has remained low (13%) due to its late diagnosis and substantial lack of effective treatment (Siegel et al., 2024). The pancreaticoduodenectomy, also known as the Whipple procedure, is a surgical operation for pancreas tumor removal, demanding highly skilled surgeons to perform successfully several critical tasks. For instance, the anastomosis between the remaining pancreas, after tumor removal, and the jejunum requires extensive expertise, taking advantage of advanced image-based surgical planning. In this context, the three-dimensional visualization of the pancreas is crucial to enhance the clinical understanding of the surgical site, empowering pre-operative planning, surgical simulation, and intra-operative navigation (Morineau et al., 2009; Nguyen and Melstrom, 2020). For this reason, image processing of multi-phase (i.e. arterial and venous phases) computed tomography (CT) scans, including segmentation and reconstruction, is mandatory to generate a digital representation of organs and lesions (Santambrogio et al., 2022; Zhang et al., 2023a). However, pancreas segmentation is challenging

* Corresponding author at: Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy.

E-mail addresses: matteo.cavicchioli@polimi.it, matteo.cavicchioli@aimsacademy.org (M. Cavicchioli).

[1] Equally contributing authors.

https://doi.org/10.1016/j.compmedimag.2024.102434
Received 6 February 2024; Received in revised form 20 June 2024; Accepted 7 September 2024
Available online 13 September 2024

for several reasons. First, the organ is irregular in shape and easily deformed. Its shape, size, aspect ratio, position, and orientation inside the abdomen vary substantially among patients (Dai et al., 2023; Man et al., 2019). Second, the pancreas occupies less than 0.5% of the entire CT volume (Man et al., 2019). Third, the lack of contrast around the boundaries of the abdominal organs makes the segmentation demanding, requiring extensive expert annotation and cross-validation (Moglia et al., 2024). As a result, manual pancreas segmentation at acceptable clinical quality is not only time-consuming and costly, but also prone to uncertainties. Hence, automatic procedures have been recently investigated exploiting recent advances in deep learning (DL) with a special focus on encoder–decoder convolutional neural networks (CNNs), such as the UNet model (Huang et al., 2021; Liu et al., 2022; Li et al., 2022; Tong et al., 2023). Supervised DL is a data-driven technique to train CNNs based on labeled image datasets. To make training trustworthy, the dataset must capture the full diversity and complexity of the problem domain as much as possible. This makes mandatory a large number of samples, featuring high-quality annotations. However, gathering and accurately labeling hundreds and even thousands of abdominal multi-phase CT scans raises economic and skilled effort issues. Nonetheless, data scarcity due to the reduced size of datasets and low data quality are acknowledged to prevent reliable training of CNNs, degrading the segmentation accuracy and slowing down the adoption of such models in clinics (Qayyum et al., 2020; Lim et al., 2022). Specific studies addressed data scarcity, exploring innovative strategies like data augmentation, transfer learning, semi-supervised learning, and even self-supervised learning. The enlargement of the training dataset by augmentation through techniques like rotation, zooming, or flipping, was shown to only partially enhance model robustness, leaving generalizability still an issue (Bansal et al., 2022; Garcea et al., 2022; Tajbakhsh et al., 2020). Likewise, transfer learning, leveraging heterogeneous knowledge pre-stored in the network, was proved to be beneficial as soon as pre-trained models are from domains closely related to the target domain (Tajbakhsh et al., 2020; Valverde et al., 2021). In addition, the effectiveness of transfer learning was shown to be affected by the quality and relevance of the data source (Kora et al., 2022). Recent efforts, focusing on semi-supervised and self-supervised learning (Kora et al., 2022; Senkyire and Liu, 2021; Zhang et al., 2023b; Rani et al., 2023) showcased interesting results in exploiting pseudo labels, although at the cost of increased model complexity and unsustainable computational infrastructures. Despite advances in data augmentation and other methodologies, acquiring and annotating large, high-quality datasets for pancreas segmentation remains challenging. This raises a critical question whose answer still needs to be met: *How much data are needed to train an effective DL model?*

Answering this question by understanding the correlation between dataset size, cost, and accuracy is of paramount importance to optimize resource allocation while ensuring the highest standards of patient care. Leveraging the data-gathering processes operated by the AIMS Academy Foundation in Milan, Italy (https://www.aimsacademy.org/en/), we analyzed the process behind the collection procedure. We provided insights into the challenges associated with large dataset acquisition, annotation, and manual refinement, running an extensive experimental study to investigate the influence of such features on DL-driven pancreas segmentation and its subregions. Valuable insights into the optimal dataset size, its impact on model performance, and practical guidelines for future research were carried out. To ensure a comprehensive examination, we addressed the following research questions:

- **RQ1**: How does the process of acquiring a large medical image dataset impact the time and resources required for its creation?
- **RQ2**: Is there an optimal dataset size for training deep learning models to achieve accurate pancreas segmentation, considering the trade-off between performance and resource limitations?
- **RQ3**: To what extent does the size and quality of a training dataset influence the time and the entity of manual refinement for automatic pancreas segmentation results?
- **RQ4**: How do anatomical subregions variability affect the required dataset size to attain the targeted segmentation accuracy?

The remainder of this paper is structured as follows. In Section 2, we provide a literature review on both public and private datasets for pancreas segmentation. Section 3 presents the overall study outline and the experiment configurations. Then, we address the challenges related to the research questions for the creation of the AIMS-1300 dataset (data acquisition, optimal size for segmentation, manual refinement, and impact of anatomical subregions) sequentially in Sections 4, 5, 6, and 7. We provide context, takeaway messages, theoretical motivations, and results for each challenge. Section 8 discusses the queries initially posed in Section 1. Finally, in Section 9, we discuss the results and present our conclusions. The code is available at https://github.com/hal9000-lab/Pancreas-2.5D.

2. Related work

2.1. Public datasets for pancreas segmentation

Six datasets for pancreas segmentation are publicly available online (Table 1). The National Institutes of Health (NIH) and the Medical Segmentation Decathlon (MSD) datasets include only annotations of the pancreas. The other four incorporate the segmentation of multiple abdominal organs, namely 8 (Multi-organ Abdominal CT Reference Standard Segmentations), 15 (AMOS-CT), 16 (WORD), and 4 (AbdomenCT-1k). Only AMOS-CT and AbdomenCT-1k are multi-vendor and multicenter, with data from two and 12 centers, respectively (Ji et al., 2022; Ma et al., 2021). In the NIH dataset, the pancreas was manually labeled by a medical student and then verified by an experienced radiologist (Roth et al., 2015; Ma et al., 2021). The images of the MSD dataset were provided by the Memorial Sloan Kettering Cancer Center (New York, NY, United States). The issues related to the manual annotation of the pancreatic parenchyma and lesions (cyst or tumor) were described in Simpson et al. (2019). The Multi-organ Abdominal CT Reference Standard Segmentations dataset was manually labeled by a research fellow under the supervision of a board-certified radiologist (Gibson et al., 2018). In the AMOS-CT dataset, 50 out of 500 CTs were initially annotated by humans. Then, one three-dimensional (3D) UNet was trained using these 50 CTs to pre-label the remaining ones (coarse stage). Five junior radiologists refined the segmentation results. To further reduce errors, three senior radiologists with more than 10 years of experience checked and validated the results (fine stage). The process was iterated several times to reach a final consensus on the well-labeled annotations (Ji et al., 2022). For the WORD dataset, a senior oncologist with seven years of experience used ITK-SNAP to manually segment all organs. Subsequently, an expert oncologist with over 20 years of experience checked the annotations for the final consensus (Luo et al., 2021). For the AbdomenCT-1k dataset, 15 junior annotators (one to five years of experience) used ITK-SNAP to manually segment the organs under the supervision of two board-certified radiologists. Then, one senior radiologist with more than 10 years of experience checked the annotations. After annotation, UNet models were trained to find the possible errors, which were double-checked by the senior radiologist (Ma et al., 2021). The dataset grouped the MSD Pancreas (420 cases), the NIH (82 cases), and 20 CT scans of patients with pancreas cancer from Nanjing University (Ma et al., 2021).

The NIH dataset is considered the most common one for testing pancreas segmentation algorithms (Moglia et al., 2024). UNet coupled with spatial and channel attention reported the highest Dice Score Coefficient (DSC) (see paragraph 3.4, and Eq. (6) for detailed definition) equal to 0.91, followed by a two-stage architecture, exploiting UNet

Table 1
Publicly available datasets of abdominal CT scans endowed with segmentation of the pancreas morphology and lesions. Missing data indicate varying numbers.

	Name	Country	Size	Phase	Application	Resolution	Number of slices	Slice thickness [mm]	Annotation
Roth et al. (2015)	National Institute of Health (NIH)	United States	82	Arterial	Parenchyma	512 × 512	181–466	[1.5:2.5]	Manual
Simpson et al. (2019)	Medical Segmentation Decathlon (MSD)	United States	420	Venous	Parenchyma Tumors	–	–	2.5	Manual
Gibson et al. (2018)	Multi-organ Abdominal CT Reference Standard Segmentation	United States	90 (43 from NIH 47 from Beyond the Cranial Vault)	Arterial	Parenchyma Tumors (4 abdominal organs)	–	–	–	Manual
Ji et al. (2022)	AMOS-CT	China	500	Venous Arterial	Tumors (15 abdominal organs)	–	–	2.5	Semiautomatic
Luo et al. (2021)	WORD	China	150	Venous Arterial	Radiotherapy (16 abdominal organs)	512 × 512	159–330	[2.5:3.0]	Manual
Ma et al. (2021)	AbdomenCT-1k	China	1112 (MSD, NIH, 50 from Nanjing University)	Venous Arterial	Parenchyma Tumors (4 abdominal organs)	–	–	–	Manual

for localization and vision transformer for segmentation, achieving a DSC of 0.89 (Shan and Yan, 2021; Dai et al., 2023). The latter model reported the highest DSC on the MSD dataset (0.91) (Dai et al., 2023). AbdomenCT-1k and AMOS-CT datasets were used in the same study, proposing a two-stage 3D UNet for localization and a UNet with multi-branches feature attention as encoder, and feature attention aggregation as decoder (Li et al., 2023a). This architecture achieved a DSC of 0.86 and 0.78 on Abdomen1k-CT and AMOS-CT, respectively (Li et al., 2023a). Tian et al. (2023) described a two-stage architecture combining nnU-Net (coarse stage) and a variational model, embedding the directional and magnitude information of the boundary intensity gradient. It was shown to improve boundary delineation (fine stage) reporting a DSC of 0.89 on AbdomnenCT-1k, 0.87 on NIH, and 0.87 on MSD, respectively. The trained model on AbdomenCT-1k was tested for generalization on 50 CT scans acquired at Jiangsu Province Hospital (China), achieving a DSC of 0.90 (Tian et al., 2023). The WORD dataset was used to set the benchmark with 10 different DL architectures. 3D nnU-Net achieved the highest DSC (0.85) (Luo et al., 2021). A 3D DenseVNet reported a DSC of 0.78 on the Multi-organ Abdominal CT Reference Standard Segmentation (Gibson et al., 2018). A comprehensive analysis of DL for pancreas segmentation using public datasets was documented by a recent systematic review (Moglia et al., 2024).

2.2. Private datasets for pancreas segmentation

A dataset of 1917 CT scans without pancreatic pathology was collected internally at Mayo Clinic (Rochester, MN, United States) and segmented by two expert radiologists (Panda et al., 2021). A two-stage UNet model was used for localization and segmentation. The training dataset was stratified on different subsets (200; 500; 800; 1000; 1200; 1500; and 1628 cases), achieving a DSC ranging from about 0.74 to 0.91. The same model was tested for generalization on The Cancer Imaging Archive (TCIA) and NIH datasets, obtaining a Dice score of 0.96 and 0.89, respectively (Panda et al., 2021) A dataset of 1150 CTs was built at John Hopkins University (Baltimore, MD, United States) from 575 healthy individuals, evaluated as potential living-related renal organ donors, and undergoing the dual phase (venous and arterial) imaging protocol. A total of 21 structures (organs and other anatomic landmarks) were annotated (Park et al., 2020). A two-stage attention network achieved a DSC of 0.87 (Park et al., 2020). At the Gil Medical Center, Gachon University College of Medicine in Incheon (South Korea), a dataset of 1006 CTs was collected and annotated (no tumor lesions). It was used to train four 3D UNet models before generalization on the NIH dataset (Lim et al., 2022). The internal and NIH datasets reported DSC of 0.84 and 0.73, respectively (Lim et al., 2022). At the Peking Union Medical College Hospital (Beijing, China), a dataset of 224 CTs was annotated by two junior radiologists and checked by an expert. It was used to train a 3D encoder and 2D decoder before generalizing on an external dataset of 66 CTs labeled by a young and checked by an expert radiologist at HeHan cancer center in Zhengzhou, China (Qu et al., 2022). The model reached a DSC of 0.90 on the training dataset and 0.86 on the generalization dataset (Qu et al., 2022). A private dataset of 104 CT scans (Li et al., 2022) with cancer cases was collected at Renji Hospital in Shanghai (China). It was used in conjunction with NIH and MSD datasets in two studies (Li et al., 2022, 2023b). In the first one, two datasets were trained in turn, while the third one was tested for generalization. A two-stage architecture based on 3D UNet was employed, reaching a DSC of 0.84 on the private dataset (Li et al., 2022). In the second study, a two-stage 3D UNet architecture was adopted to differentiate between organs (first stage) and to enhance the characterization of high-uncertainty regions (second stage). During the test for generalization on the Renji Hospital dataset, the model obtained a DSC of 0.73 and 0.83 when trained, respectively, on NIH and MSD datasets (Li et al., 2023b). A detailed analysis of deep learning techniques for pancreas segmentation using private datasets was provided by a recent systematic review (Moglia et al., 2024).

3. Methodology of the work

In this section, we detail the methodology employed in this study. We first outline the study design and then comprehensively explain the selected DL methodology. Subsequently, we present the training strategy adopted and a detailed description of the metrics used for evaluation.

3.1. Study outline

The study addresses the issues highlighted in Table 2, aligning with the research questions presented in Section 1. We thoroughly investigate the data collection process by leveraging information from

Table 2
Investigated challenges represented by four main research questions.

	Research objectives
Dataset acquisition (RQ1)	Medical images collection Manual annotation procedure Expert check and consensus for approval
Optimal dataset size (RQ2)	Accuracy-based optimal size Metrics performance trend Dataset size plateau effect Generalization performance
Manual adjustments (RQ3)	Manual corrections according to dataset size Slices of errors based metric
Anatomical subregions (RQ4)	Complexity-Accuracy qualitative correlation Pancreatic regions analysis

the data acquisition, annotation, and validation procedures performed by the Fondazione MIAS (AIMS Academy). Furthermore, to address the research questions, we conducted extensive experiments training our proposed DL model on the AIMS-1300 CT angiography (CTA) dataset.

3.2. Selected DL methodology

For our experimental analysis, we selected a CNN utilizing the 2.5D UNet architecture. This architecture bridges the gap between traditional 2D and 3D architectures, offering an optimal blend of contextual information and spatial resolution. Such a compromise enabled the model to leverage both the computational efficiency of 2D models and the depth of information of 3D models. Furthermore, the reduced computational demand of the 2.5D UNet made it especially fitting for our study. Our approach was built upon the foundation laid by Fantazzini et al. (2020), who introduced the 2.5D UNet architecture for aortic arch segmentation. We adapted this framework for the task of pancreas segmentation (Fig. 1). First, a 2D UNet trained on downsampled axial slices of CTAs generated a preliminary, coarse segmentation of the pancreas. This segmentation was used to locate the overall bounding box, embedding all pancreases in the dataset, in the centroid of the coarse pancreas. Then, a pancreas-encompassing subvolume from the original full-resolution CTAs was extracted. This operation minimized the confounding influence of other abdominal structures during subsequent refinement stages, excluding irrelevant backgrounds. The full-resolution subvolumes were then sliced and fed to three 2D UNets with the same configuration. Each network operated on a different anatomical axis (axial, coronal, and sagittal), producing three different pancreas segmentations, one for each axis. The three network outputs were sequences of 2D probability maps, each based on the corresponding axial view. 3D volumes were obtained by concatenating subsequent slices together. The final pancreas segmentation was obtained by an integrated multi-view approach based on majority voting. Before training, the CTAs were preprocessed to improve model accuracy. The volumes were windowed in the range of [−150, +250] Hounsfield Unit (HU) and then resampled to a common spacing of [0.79 mm, 0.79 mm, 3.0 mm] using interpolation. Since the networks' output activation was a sigmoid function, the Binary Cross-Entropy loss (CE) was used as:

$$CE = -\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \tag{1}$$

quantifying the discrepancy between the value predicted by the network $\hat{y}_i$ and the target value (ground truth pancreas segmentation) y_i, where N represents the total number of pixels in a 2D CTA slice.

3.3. Training strategy

The selected DL model was trained on eleven subsets, including 25; 50; 100; 150; 295; 440; 585; 730; 875; 1020; and 1170 samples of the AIMS-1300 dataset. Each sample corresponded to a single CTA scan acquired in the arterial phase with dimensions of M×512 × 512, where M denoted the number of axial slices. The eleven trained models were subsequently evaluated on a common hold-out test set, including 130 samples of AIMS-1300 (Fig. 2). DL tasks were developed using TensorFlow (Abadi et al., 2015) with Keras (Chollet et al., 2015) on an NVIDIA GeForce GTX 1080 GPU with 8 GB of graphics memory (VRAM). The optimal set of hyperparameters was determined through grid search. Each model underwent training for 100 epochs, using a learning rate of 0.0001 and a batch size of 6 slices. The total training time was 14 days, whereas grid search needed 20 days.

3.4. Evaluation metrics

In this paper, we carefully selected a set of six metrics to evaluate the network performances from different perspectives: DSC, Precision (PR), Recall (RE), Relative Volume Difference (RVD), Hausdorff Distance (HD), and Average Symmetric Surface Distance (ASSD). The selected metrics, each focusing on distinct aspects, enabled a comprehensive evaluation of the models. The given metrics are formally presented considering the medical volume as a set of points $X = [x_1, x_2, \dots, x_n]$ with $|X| = w \times h \times d = n$, where w, h, d are the width, height and depth of the grid in which the volume is defined. Each voxel of volume X has a corresponding label in both the ground truth segmentation S_g and the automatic segmentation predicted by the model S_p. We defined as $S_g(x) \in \{0, 1\}$ the label assigned to voxel x by the segmentation S_g, and $S_p(x) \in \{0, 1\}$ the label assigned to voxel x by the segmentation S_p. Since the formulations of the size-based and overlap-based metrics can be derived from the four cardinalities of the confusion matrix, namely true positive (TP), false positive (FP), true Negative (TN), false Negative (FN), we defined them as follows:

$$TP = \left|\{x \in X : S_g(x) = 1 \text{ and } S_p(x) = 1\}\right| \tag{2}$$

$$FP = \left|\{x \in X : S_g(x) = 0 \text{ and } S_p(x) = 1\}\right| \tag{3}$$

$$FN = \left|\{x \in X : S_g(x) = 1 \text{ and } S_p(x) = 0\}\right| \tag{4}$$

$$TN = \left|\{x \in X : S_g(x) = 0 \text{ and } S_p(x) = 0\}\right| \tag{5}$$

where $|\cdot|$ denotes the count of the set.

Dice Score Coefficient (DSC) is the most popular performance metric in medical image segmentation. It measures the overlap between the ground truth segmentation and the predicted segmentation (Kamnitsas et al., 2017; Ronneberger et al., 2015; Taha and Hanbury, 2015; Zijdenbos et al., 1994; Li et al., 2019). A score of 0 indicates no overlap, while a score of 1 indicates perfect overlap. DSC is insensitive to the size and shape of the target region, making it a relatively easy-to-interpret metric. However, DSC does not differentiate between under-segmentation and over-segmentation errors nor considers the spatial distribution of classified voxels. It is defined by:

$$DSC = \frac{2|S_p \cap S_g|}{|S_p| + |S_g|} = \frac{2TP}{2TP + FN + FP} \tag{6}$$

Precision (PR) and **Recall (RE)** are complementary metrics that can address DSC's limitations in distinguishing between under and over-segmentation errors. PR measures the proportion of identified positive voxels that are TP, while RE measures the proportion of TP voxels that are correctly identified. Both metrics range from 0 to 1, focusing on incorrectly predicted voxels (Taha and Hanbury, 2015; Hicks et al., 2022). PR is defined by:

$$PR = \frac{|S_p \cap S_g|}{|S_p|} = \frac{TP}{TP + FP} \tag{7}$$

while RE is defined by:

$$RE = \frac{|S_p \cap S_g|}{|S_g|} = \frac{TP}{TP + FN} \tag{8}$$

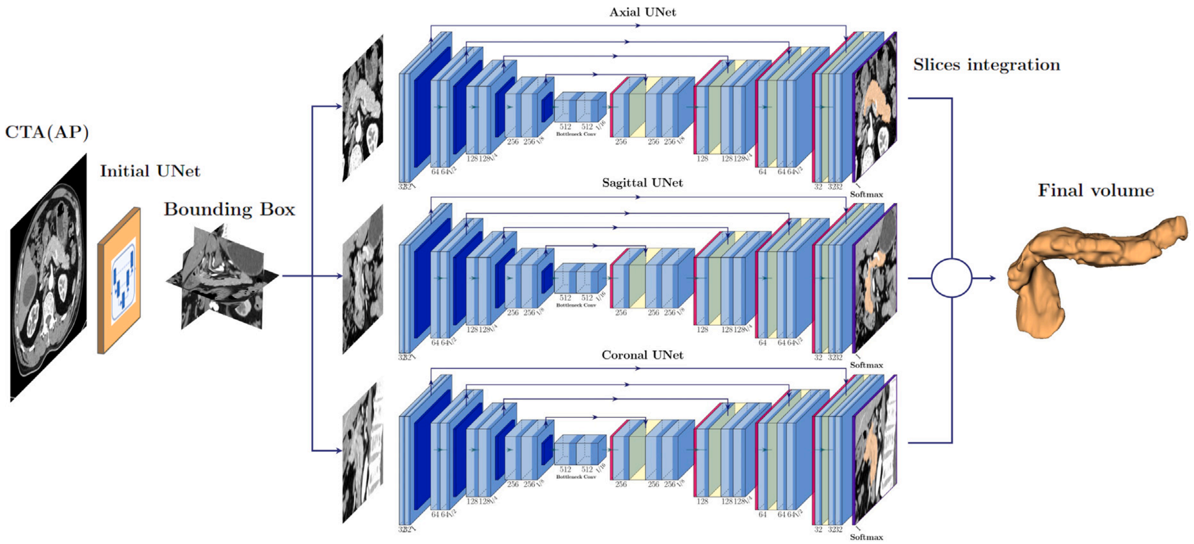


Fig. 1. The proposed 2.5D UNet architecture to examine the correlation between segmentation quality and dataset size.

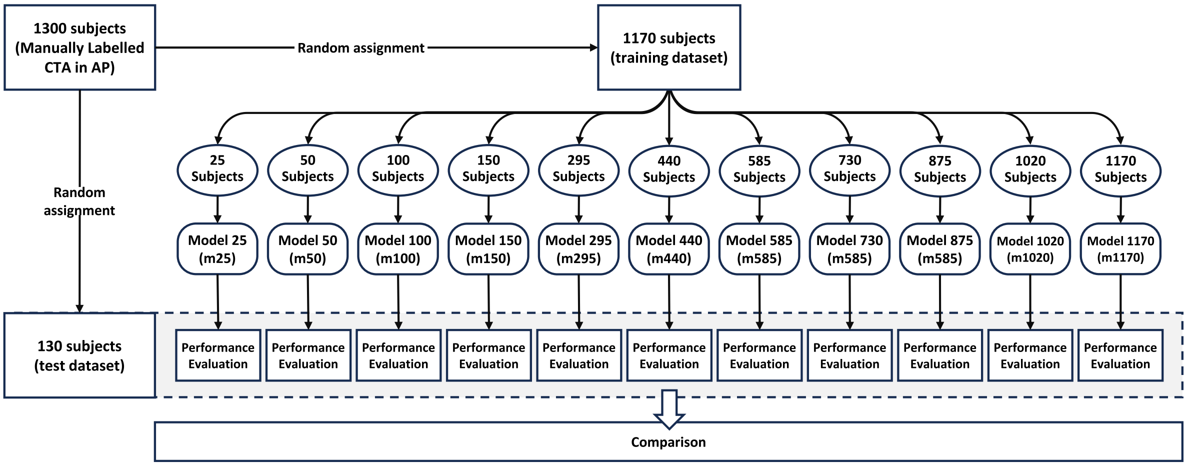


Fig. 2. Training workflow overview. As depicted, the AIMS-1300 dataset was partitioned into eleven subsets of an increasing number of samples during training, obtaining eleven segmentation models that were then tested on an independent set of 130 samples.

Relative Volume Difference (RVD) is a size-based metric that measures the percentage difference in volume between the machine-segmented region and the ground truth, normalized by the reference volume. It is an asymmetric metric that can indicate over-segmentation (RVD $>$ 0) or under-segmentation (RVD $<$ 0). The absolute value of RVD indicates the magnitude of the voxel set difference, but it does not consider the spatial distribution of voxels (Yeghiazaryan and Voiculescu, 2018). RVD is defined by:

$$RVD = \frac{|S_p| - |S_g|}{|S_g|} = \frac{(TP + FP) - (TP + FN)}{TP + FN} \tag{9}$$

Hausdorf Distance (HD) is notably known as a boundary metric and is designed to evaluate the shape similarity between two sets of points within a defined metric space. HD is unaffected by point correlations, focusing only on pairwise voxel distances. However, it has a pronounced sensitivity to outliers. It is defined by:

$$HD(S_g, S_p) = \max(h(S_g, S_p), h(S_p, S_g)) \tag{10}$$

where $h(S_g, S_p)$ is called the directed Hausdorff distance and is given by:

$$h(S_g, S_p) = \max_{x_g \in S_g} \min_{x_p \in S_p} \|x_g - x_p\| \tag{11}$$

where $\|x_g - x_p\|$ is some norm, e.g. Euclidean distance.

Average Symmetric Surface Distance (ASSD) is defined as the average of all the distances from points on the boundary of the ground truth segmentation to the boundary of the predicted segmentation, and vice-versa (Yeghiazaryan and Voiculescu, 2018). The ASSD can be expressed as:

$$ASSD(S_g, S_p) = \frac{1}{|S_g| + |S_p|} \left(\sum_{x_{sg} \in S(S_g)} d(x_{sg}, S(S_p)) + \sum_{x_s p \in S(S_p)} d(x_s p, S(S_g)) \right) \tag{12}$$

where $d(x_{sg}, S(S_p))$ is defined as:

$$d(x_{sg}, S(S_p)) = \min_{s_{sp} \in S(S_p)} (\|s_{sg} - s_{sp}\|) \tag{13}$$

with $S(S_g)$ and $S(S_p)$ representing surfaces of S_g and S_p respectively.

4. Image data acquisition (RQ1)

In DL for biomedical image segmentation, the dataset quality is essential for reliable performance and accurate results (Cobo et al., 2023; Tajbakhsh et al., 2020). A quality dataset includes clear images and validated annotations that represent the subject's population

Table 3
Properties of AIMS-1300 dataset.

Name	Country	Size	Images	Phase	Application	Resolution	Spacing [mm]	Slice thickness [mm]	Annotation	Pathological
Fondazione MIAS (AIMS Academy)	Italy	1300	CTA	Arterial	Parenchyma	512 × 512	[0.57:0.97] × [0.57:0.97]	[0.8:4.0]	Manual	✓

well and feature appropriate cardinality and class frequency. However, image collection, annotation, and validation of quality datasets are expensive and time-consuming (Paullada et al., 2021). Several advanced data augmentation methodologies were introduced to address this challenge (Tajbakhsh et al., 2020). However, the most careful strategy to ensure the reliability of DL models in the medical domain remains the manually acquired dataset (Arora et al., 2023; Roh et al., 2019; Tajbakhsh et al., 2020). Furthermore, the annotation process continues to be a significant obstacle in automated medical image segmentation, often demanding more resources and time than the actual development of the algorithm (Tajbakhsh et al., 2020). This raises an often-overlooked conceptual question:

> *What is the allocation of economic, human, and time resources required to acquire a clinically useful dataset?*

In this section, we systematically explored the dataset acquisition process of 1300 CTAs performed by AIMS Academy, delving into the problems of quality, costs, and times driven by theoretical motivations. The main takeaways in this section are summarized below.

1. Gathering 1300 CTA scans from various facilities highlighted the logistical complexities of assembling a diverse and comprehensive dataset.
2. Precisely calculating the annotation time underlined the need for efficient time management and resource allocation in future works.
3. Quality checks by medical experts to reach a consensus not only ensure accuracy but also highlight the need for expert involvement, which could be a resource-intensive process.

4.1. Theoretical motivation

Several key motivations drove our exploration of the dataset construction process. Firstly, resource optimization was crucial. Analyzing the entire process, from image collection to annotation and approval from experts, provided insights into quantifying human, time, and financial resources. Secondly, a precise computation of the above-mentioned resources would benefit the included stakeholders (e.g., hospitals, medical centers, and private companies) investing in creating the dataset. Lastly, a thorough assessment of the entire dataset creation process enabled us to identify potential issues for possible future improvements.

4.2. AIMS-1300 dataset

The AIMS-1300 dataset was constructed by acquiring abdominal CTAs from April 2019 to December 2019. The initial phase of the process was particularly time-consuming and involved the acquisition of images. This effort was related to image availability, anonymization, and provider agreements. The duration of this phase was challenging to quantify, mainly due to the varying bureaucratic procedures of different providers, often extending to several months. Following the necessary approvals, AIMS Academy collected 1300 anonymized DICOM (Digital Imaging and Communications in Medicine) images in compliance with the GDPR (General Data Protection Regulation) privacy and security law. These acquired CTAs followed a specific acquisition protocol, starting with a non-contrast phase and progressing to injecting an iodinated contrast agent. The CT angiography protocol progressed through distinct phases at specific intervals following the injection. The arterial phase (AP) was acquired 20–30 s after the injection. The portal venous phase (PVP) follows the AP around 60–70 s after the injection. Finally, the late phase, which captures delayed contrast retention, was obtained approximately 3–4 min post-injection. This sequential process provided a holistic view of the structure and vasculature of the pancreas. Furthermore, the CTAs in the dataset featured an axial resolution of 512 × 512 pixels. The dimensions of these images were characterized by spacing parameters $h \times w \times d$, with h and w spanning from 0.57 mm to 0.97 mm and d varying between 0.80 mm and 4.0 mm. AIMS-1300 strictly adhered to the DICOM standard to ensure uniformity and consistency across the collected images. Table 3 presents a complete overview of the dataset features. After data collection, AIMS-1300 started a robust annotation process to label the pancreas on the abdominal CTA in the AP according to internal assessments. To minimize human error and guarantee the quality of annotations, a dedicated team of three annotators, each with at least one year of experience, underwent specialized training. This training was conducted by experienced physicians, each with over seven years of experience. The purpose was to equip the annotators with the skills to execute segmentations adhering to a designed annotation protocol specifically tailored for pancreas parenchyma. DICOM images were manually labeled using 3D Slicer (www.slicer.org), with the annotators delineating the pancreas area on each slice, marking and labeling the region of interest. Given the intricacy and variance of pancreas morphology in the dataset, AIMS-1300 provided annotators with explicit criteria and guidelines, emphasizing the importance of capturing the entire organ, including the head, body, and tail. To achieve consensus on the dataset, expert physicians conducted the review process in two distinct stages: initially, concurrent with the annotation phase, and subsequently, upon the dataset's completion. They first reported discrepancies between annotations in consensus meetings, where they jointly reviewed erroneous annotations and chose the most accurate representation. Finally, they performed random spot checks on 60% of the annotated images, evaluating the accuracy of the labels and ensuring that they conformed to predetermined guidelines. The original annotators reviewed and corrected all anomalies detected during the two consensus phases. Three distinct cases of CTA and ground truth (GT) reconstruction are depicted in Fig. 3.

4.3. Results

The data acquisition process carried out by AIMS Academy began with the collection of 1300 CTA scans from multiple medical facilities. Analysis of this process allowed us to answer the open question **RQ1**. All CTAs were acquired using a standardized imaging protocol on healthy and pathological subjects. The population included 1165 healthy subjects: 100 subjects with adenocarcinoma, 31 with cystic tumors, and 4 with neuroendocrine tumors. The total time T_a required for annotation (slice by slice) was divided into two distinct parts: a fixed duration of 80 h allocated for annotators' training T_t and a variable time T_l to label the CTAs. The latter varied based on factors such as the number of images, the complexity of the structures involved, and the resolution of the images. The labeling process of the parenchyma of the pancreas required 40 min for each scan, on average. Notably, the pancreas annotation time on a 1.5 mm resolution CTA was twice that of a 3 mm CTA. The entire annotation process, considering the average time, took about 938 h, obtained as follows:

$$T_a = T_t + T_l = 80\,\mathrm{h} + \left(\frac{0.66\,\mathrm{h}}{CTA} \times 1300\,CTA\right) \approx 938\,\mathrm{h}. \tag{14}$$

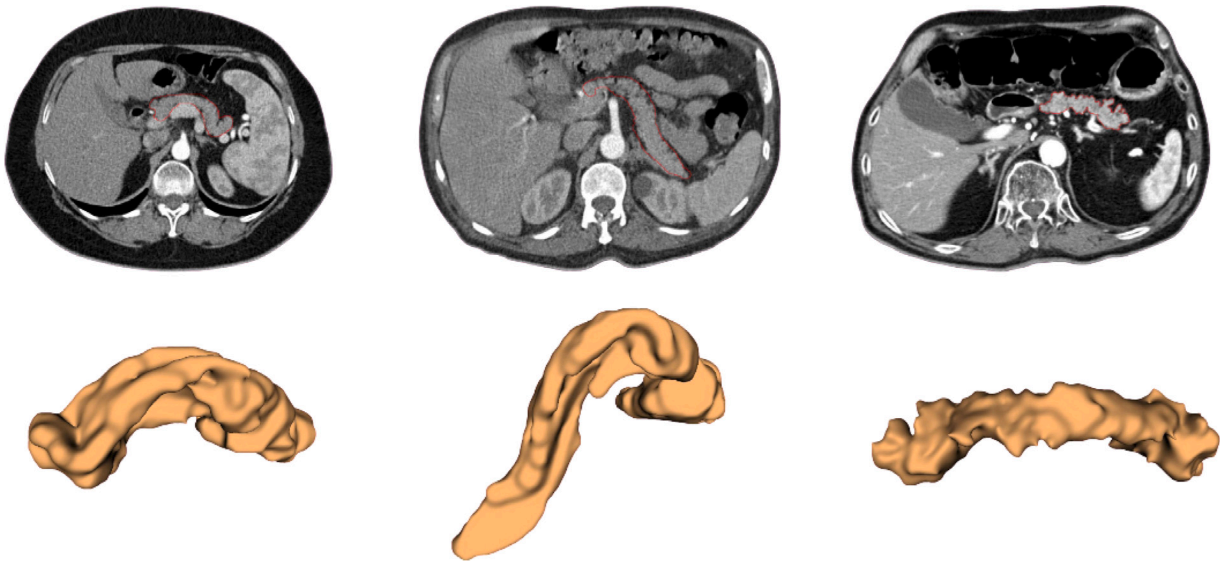


Fig. 3. Three cases of AIMS-1300 dataset: 2D axial view of CTAs in AP (top), and 3D perspective view of pancreas ground truth (bottom).

Therefore, the time required to increment the dataset size of 100 CTs was equivalent to about 70 h, considering only the annotation process. The last phase involved the validation of the dataset quality, conducted by three experienced physicians. Quantifying the precise duration was challenging due to validator availability and the amount of detected errors. As expected, addressing significant errors demanded additional time from the validators.

5. Optimal dataset size (RQ2)

As revealed by a recent review, the prevailing architecture for pancreatic parenchyma segmentation is based on the UNet model (Ronneberger et al., 2015; Moglia et al., 2024). This architecture typically features symmetrical encoding and decoding paths linked by skip connections. Introducing skip connections to combine features at different levels achieved remarkable performance gains (Drozdzal et al., 2016). Despite novel alternative models, exemplified by generative-adversarial networks and transformers, U-shaped architecture remains the most common architectural basis. Due to its intrinsic complexity, UNet requires a training set with a substantial amount of annotated data to achieve generalizability and robustness (Çiçek et al., 2016). In particular, the reproducibility, generalizability, and accuracy of a neural network in delivering reliable results are greatly influenced by the representativeness of its training dataset. Therefore, it is crucial that the dataset well mirrors the target population (Renard et al., 2020; Alzubaidi et al., 2021; Çiçek et al., 2016; Esteva et al., 2019). In domains where precision is paramount, like medical image analysis, the adequacy of both the quality and quantity of data becomes even more critical for the model to reach its optimal performance level (Paullada et al., 2021). Given the considerable time and resources required to acquire high-quality datasets, it becomes essential to have a tool that can determine the optimal dataset size in advance. This tool would not only preserve resources but also set an upper bound for data augmentation strategies. Such a tool AIMS-1300 to maximize network learning by identifying the most efficient dataset size, ensuring effective learning without the unnecessary use of extensive resources. This approach is particularly valuable in fields where data collection is both costly and labor-intensive.

Drawing on our experience at AIMS Academy, this section encompasses a series of extensive experiments following the methodology described in Section 3.2 to determine the optimal dataset cardinality for pancreatic parenchyma segmentation. For this purpose we compared the proposed 2.5 UNet with two extensively investigated networks, namely UNet with Attention Gate (UNet-AG) (Vaswani et al., 2017; Oktay et al., 2018) and nnU-Net (Isensee et al., 2021). UNet-AG and nnU-Net were both trained using three different loss functions, namely Binary Cross-Entropy loss (CE), Dice loss (DL) (Li et al., 2019), and Focal loss (FL) (Lin et al., 2017). We adopted the 2D versions of UNet-AG and nnU-Net networks instead of their 3D counterparts to preserve computational efficiency. For each network, we derived a trend curve relating performance metrics to dataset size. This trend curve is not only exploitable for pancreatic parenchyma segmentation but also generalizable to other organs. After testing the three DL models on the AIMS-1300 dataset, we assessed their generalization capability by testing them on the AMOS-CT dataset.

The main takeaways messages of this section are summarized below.

1. The relationship between performance metrics and the size of the training dataset exhibited a logarithmic trend.
2. Beyond a certain dataset size, increasing the dataset size resulted in negligible performance improvements.
3. An optimal dataset size can be determined based on the desired level of accuracy.
4. Insights into optimal dataset cardinality can guide the calibration of data augmentation strategies.

5.1. Theoretical motivation

Exploring the correlation between segmentation metrics and dataset size is fundamental in the quest of the required performances. Identifying the ideal dataset size is important to avoid unnecessary data collection costs and improve model performance on unseen data. Standardizing dataset sizes would lead to the definition of benchmarks to foster an objective comparison among different models and the adoption of data augmentation strategies. Optimal dataset size is also important to train effective DL models, ensuring they adapt to data variations and perform consistently in clinical settings. Ultimately, having this knowledge could underpin clinical validation and the development of more effective segmentation tools.

5.2. Results on AIMS-1300 dataset

Motivated by the theoretical considerations previously discussed, we explored how the variation in training dataset size impacted performance metrics, addressing **RQ2**. We ran the analysis according to the

Table 4
Metrics results for the proposed 2.5D UNet model (Fig. 1), trained with different dataset sizes and tested on 130 samples of AIMS-1300 dataset. The best values are in bold.

(a)

DSC	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	0.70 (0.20)	0.79 (0.16)	0.83 (0.12)	0.83 (0.12)	0.85 (0.10)	0.86 (0.10)	0.86 (0.10)	0.87 (0.09)	0.87 (0.09)	0.87 (0.08)	**0.88 (0.08)**
Mean ± STD	0.65 ± 0.17	0.74 ± 0.13	0.79 ± 0.11	0.79 ± 0.12	0.82 ± 0.11	0.83 ± 0.11	0.83 ± 0.10	0.84 ± 0.10	0.84 ± 0.10	0.84 ± 0.10	**0.84** ± **0.10**
Min, Max	0.00, 0.87	0.24, 0.90	0.23, 0.92	0.22, 0.92	0.25, 0.93	0.24, 0.93	0.24, 0.94	0.23, 0.94	0.24, 0.94	0.24, 0.94	0.24, 0.94

(b)

PR	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	0.83 (0.23)	0.82 (0.19)	0.84 (0.18)	0.84 (0.14)	0.87 (0.12)	0.86 (0.11)	0.88 (0.11)	0.87 (0.10)	0.88 (0.10)	0.88 (0.09)	**0.89 (0.08)**
Mean ± STD	0.75 ± 0.22	0.76 ± 0.19	0.79 ± 0.15	0.79 ± 0.16	0.83 ± 0.13	0.83 ± 0.13	0.84 ± 0.11	0.84 ± 0.12	0.84 ± 0.12	0.85 ± 0.11	**0.86** ± **0.10**
Min, Max	0.00, 0.98	0.16, 0.97	0.14, 0.97	0.13, 0.97	0.15, 0.98	0.14, 0.97	0.14, 0.97	0.14, 0.97	0.14, 0.97	0.14, 0.97	0.15, 0.97

(c)

RE	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	0.64 (0.19)	0.78 (0.14)	0.84 (0.12)	0.85 (0.12)	0.86 (0.11)	0.87 (0.10)	0.87 (0.10)	0.88 (0.10)	0.88 (0.10)	0.87 (0.10)	**0.88 (0.10)**
Mean ± STD	0.62 ± 0.14	0.76 ± 0.11	0.82 ± 0.11	0.83 ± 0.11	0.83 ± 0.11	0.84 ± 0.11	0.84 ± 0.11	0.85 ± 0.11	0.85 ± 0.11	0.85 ± 0.11	**0.85** ± **0.11**
Min, Max	0.00, 0.89	0.44, 0.96	0.36, 0.96	0.33, 0.96	0.35, 0.96	0.45, 0.97	0.44, 0.97	0.37, 0.97	0.44, 0.97	0.44, 0.97	0.38, 0.97

(d)

RVD	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	26.36 (55.30)	3.41 (31.15)	−2.22 (26.33)	−2.47 (23.02)	0.57 (21.76)	−2.25 (17.37)	0.83 (16.37)	−1.13 (16.53)	−1.44 (17.24)	**−0.26 (17.02)**	1.00 (15.04)
Mean ± STD	26.46 ± 47.96	2.54 ± 31.62	−0.99 ± 24.90	−2.34 ± 26.42	2.18 ± 25.50	−0.06 ± 23.41	2.78 ± 22.76	**0.62** ± **23.82**	0.66 ± 22.34	2.93 ± 22.53	3.64 ± 23.45
Min, Max	−94.52, 182.66	−76.10, 96.43	−78.57, 104.99	−78.24, 91.37	−79.43, 109.70	−80.91, 111.95	−79.40, 117.15	−78.66, 111.38	−79.99, 109.96	−80.09, 109.58	−77.78, 115.63

(e)

HD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	73.20 (29.65)	33.06 (26.07)	26.63 (21.76)	22.64 (19.47)	20.65 (15.81)	15.27 (11.11)	15.54 (11.26)	14.91 (10.96)	14.17 (12.19)	14.86 (10.96)	**12.73 (11.23)**
Mean ± STD	73.84 ± 24.11	39.72 ± 21.80	31.20 ± 18.43	27.58 ± 17.53	24.87 ± 16.61	19.60 ± 13.46	19.05 ± 13.04	18.90 ± 13.12	18.97 ± 13.60	18.90 ± 12.73	**18.36** ± **13.84**
Min, Max	25.79, 181.15	12.14, 116.57	8.61, 100.90	8.04, 100.05	8.01, 101.15	6.97, 86.85	6.97, 85.35	6.50, 86.11	6.18, 88.36	6.76, 84.17	6.76, 83.30

(f)

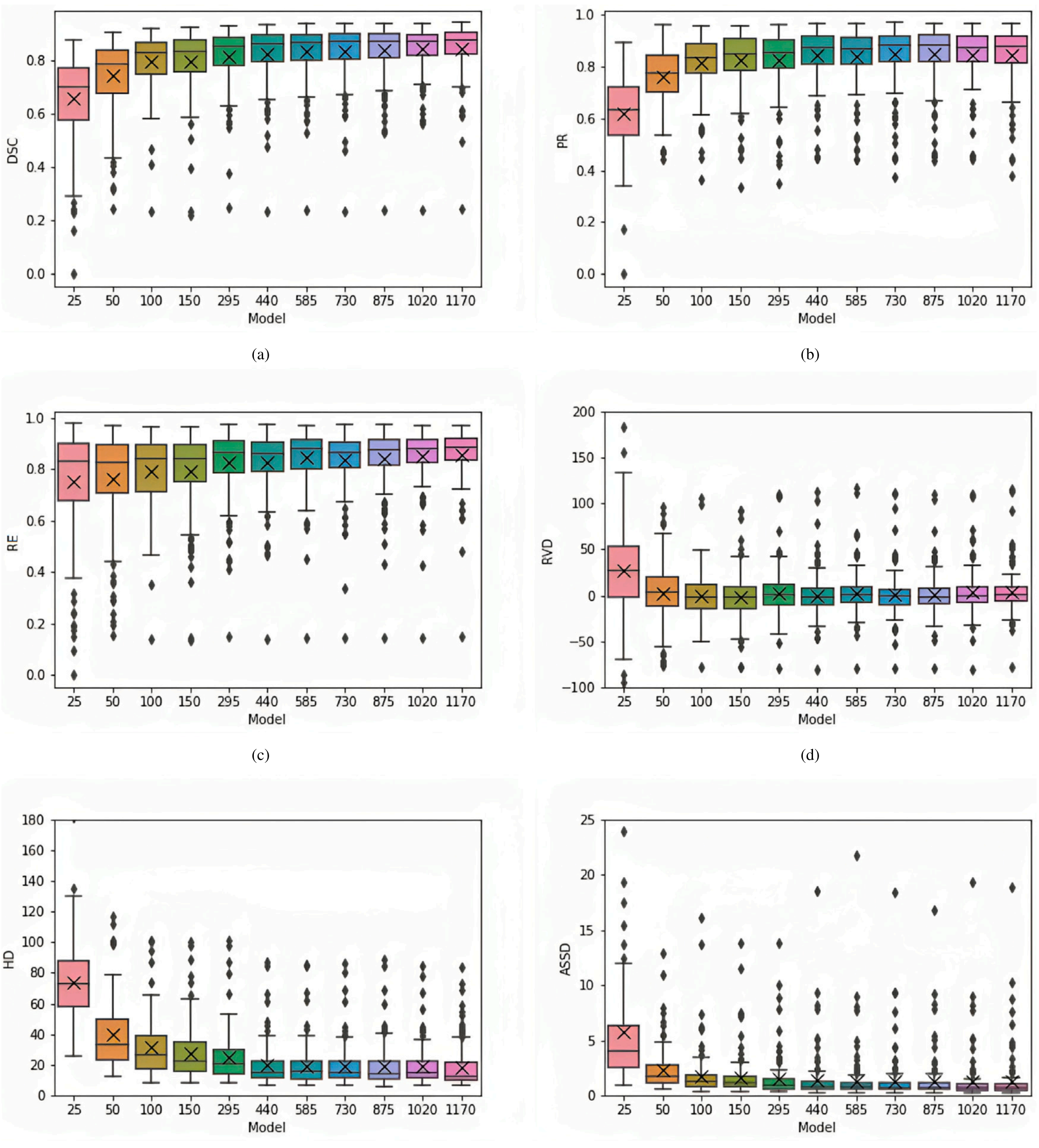
ASSD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
Median (IQR)	4.07 (3.79)	1.68 (1.56)	1.22 (1.06)	1.12 (0.96)	0.92 (0.84)	0.79 (0.67)	0.75 (0.59)	0.71 (0.60)	0.70 (0.60)	0.70 (0.52)	**0.66 (0.51)**
Mean ± STD	5.74 ± 9.24	2.26 ± 1.86	1.73 ± 2.00	1.61 ± 1.85	1.44 ± 1.82	1.35 ± 2.06	1.29 ± 2.24	1.27 ± 2.09	1.23 ± 1.95	**1.21** ± **2.09**	1.25 ± 2.16
Min, Max	0.95, 101.32	0.56, 12.84	0.37, 16.12	0.33, 13.78	0.32, 13.83	0.27, 18.50	0.27, 21.78	0.25, 18.41	0.25, 16.82	0.26, 19.32	0.25, 18.88

workflow depicted in Fig. 2, presenting results obtained from the AIMS-1300 test set with 130 subjects. Fig. 4 shows the distributions of the six reference metrics for the different models. Table 4 presents numerical results using median and interquartile range to account for the non-normal distribution of values. These metrics were further supported by including mean, standard deviation, minimum, and maximum for a comprehensive analysis. Regarding DSC, the median value showed a significant rise in Table 4a, ranging from 0.70(0.20) of m25 to 0.88(0.08) of m1170. However, as can be observed in Fig. 4(a), beyond m440, the rate of increase in the DSC began to diminish. The similarity metrics, PR and RE, also showed an increasing trend in Figs. 4(b) and 4(c), respectively. Specifically, PR went from 0.83(0.23) of m25 to 0.89(0.08) of m1170 in Table 4b, while RE went from 0.64(0.19) of m255 to 0.88(0.10) of m1170 (Table 4c). Also, these two metrics showed a decrease in the rate of improvements, starting after m295 and m440 for PR and RE, respectively. The RVD varied across different dataset sizes, with the median RVD values fluctuating from positive to negative and then back toward neutral, as shown in Fig. 4(d). The smallest dataset sizes (m25, m50) showed a high median RVD in Table 4d, suggesting a considerable discrepancy between the predicted and ground truth volumes. As the dataset size increased (m100 to m440), there was a notable improvement, with median RVD values becoming negative, indicating that the predicted volume was, on average, slightly larger than the ground truth volume. The RVD became closer to 0 with m1020, reaching −0.26(17.24). For ASSD depicted in Fig. 4(f), there was a clear decrease in the median value, showing improved similarity from 4.07 mm to 0.66 mm in Table 4f. This trend was consistent across other statistical measures, such as the interquartile range, mean, and standard deviation indicating an overall reduction in surface distance as the sample size increases. There was also a general decrease in the median value for HD, from 73.20 mm at m25 to 12.73 mm at m1170, suggesting a reduced spread in the maximum distances between the surfaces. The interquartile range, mean, and standard deviation followed a similar decrease pattern. Interestingly, the HD values in Table 4e started in a significantly higher range and showed a sharper decrease than the ASSD. This trend could be related to the HD's extreme sensitivity to outliers. This sensitivity became less pronounced as the sample size expanded, as shown in Fig. 4(e). Both ASSD and HD showed a decrease in the rate of improvement after m440. An analysis of the trends illustrated in Fig. 4 unveiled a consistent pattern where the expansion of the dataset size was associated with discernible enhancements in the metrics. Furthermore, all trends showed a plateau where the rate of improvement reduced. These observations from the graphical data showed that an optimal dataset size exists where further data inclusion does not necessarily translate into significant improvements in metrics. Identifying this optimal cardinality (called upper bound) is essential for the efficient training of models, ensuring optimal resource allocation without reducing the quality of segmentations. To statistically estimate the upper bound, we performed a non-parametric Kruskal–Wallis test with post-hoc comparison ($p<0.05$). We chose the Kruskal–Wallis test for its effectiveness in handling non-normally distributed data. As a result, the ASSD metrics proved to be the most restrictive one, with a statistically significant transition between 295 and 440 subjects ($p = 0.005$), while we did not detect any significant change ($p>0.25$) for the next greater dataset sizes. Thus, we noted that at least 400 samples in the training set should be required to ensure high-quality segmentation of the pancreas.

The results on DSC, HD, and ASSD for the 2.5D UNet, UNet-AG, and nnU-Net are reported in Table 5. An Appendix was created with the complete list of the results. As expected, the quality metrics improved for all three networks while the training set size increased. Overall, UNet-AG and nnU-Net networks trained using the cross-entropy loss provided the best results. Unlike our proposed 2.5D UNet, a significant difference was found in the transition between 440 and 585 cases in the training set ($p = 0.005$) in both the other two models. This difference underscored the enhanced sensitivity of the entire 2D models to variations in training data volume, a factor less pronounced in our 2.5D approach, which effectively integrates spatial context within a computationally efficient framework. A comparison of the metrics scores showed that our proposed 2.5D UNet outperformed UNet-AG and nnU-Net in terms of DSC, HD, and ASSD. The gap was larger for smaller partitions of the AIMS-1300 dataset used for training. The performance gap shrank for datasets with a cardinality of 440 cases or larger.

5.3. Generalization on AMOS-CT dataset

This section presents the results obtained from our models, which were trained with different AIMS-1300 dataset sizes and tested on the AMOS-CT public dataset. It is important to note that since our networks



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. Box plots for quality scores computed over the AIMS-1300 test set (130 samples) according to AIMS-1300 dataset cardinality.

were specifically trained using images taken during the arterial phase, only the CTA scans in the AP were considered from the 300 publicly available cases (200 training, 100 validation). Overall, 74 cases were identified and included in our analysis. Our study, as shown in Table 6, compared the performance of the proposed 2.5D UNet with UNet-AG and nnU-Net on the AMOS-CT dataset. The complete list of metrics is available in an Appendix. Notably, the performance trends observed in the AMOS-CT dataset closely mirrored those seen in our AIMS-1300 test dataset. This finding underscores the crucial role of dataset size in model performance. We observed a slowing down of the improvement in all AMOS-CT metrics beyond the model m440, reflecting the trends of the AIMS-1300 metrics and suggesting a potential plateau in generalization gains with increasing dataset size. The analysis of both UNet-AG and nnU-Net confirmed the best results when trained using the cross-entropy loss. Unlike our 2.5D UNet, a significant difference was found in the transition between 440 and 585 cases in the training set (p = 0.005) in both models.

6. Manual refinement of the results (RQ3)

In the clinical field, the DL models should support rather than replace the expertise of radiologists in a semi-automatic approach (Liew, 2018; Rudie et al., 2021). For this reason, models should be truly effective and should deliver accurate results, minimizing the need for manual correction. Precisely, the time and entity of manual refinements required by the masks predicted from these networks need a quantitative investigation (Sander et al., 2020; Plaza et al., 2012). In this section, we assumed that radiologists performed manual corrections on 2D images. To identify a metric capable of quantifying the extent of manual correction, we identified the slices within the predicted volumes exhibiting over- or under-segmentation errors. The frequency

Table 5
Comparison of the adopted 2.5D model, UNet-AG, and nnU-Net on DSC, HD, and ASSD metrics, evaluated on the AIMS-1300 test set of 130 subjects. Metrics are expressed as median with interquartile range (IQR).

DSC	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	0.54 (0.25)	0.68 (0.19)	0.74 (0.15)	0.70 (0.20)	0.81 (0.13)	0.83 (0.10)	0.85 (0.09)	0.85 (0.11)	0.86 (0.09)	0.85 (0.10)	0.86 (0.09)
UNet-AG-DL	0.53 (0.19)	0.65 (0.20)	0.69 (0.18)	0.73 (0.16)	0.71 (0.20)	0.74 (0.14)	0.73 (0.13)	0.72 (0.13)	0.81 (0.10)	0.83 (0.10)	0.86 (0.10)
UNet-AG-FL	0.49 (0.20)	0.65 (0.22)	0.70 (0.19)	0.77 (0.16)	0.82 (0.13)	0.81 (0.13)	0.81 (0.10)	0.84 (0.11)	0.85 (0.10)	0.86 (0.09)	0.86 (0.10)
nnU-Net-CE	0.46 (0.18)	0.62 (0.19)	0.76 (0.14)	0.79 (0.16)	0.82 (0.13)	0.83 (0.11)	0.85 (0.10)	0.85 (0.11)	0.85 (0.10)	0.86 (0.10)	0.86 (0.09)
nnU-Net-DL	0.48 (0.19)	0.43 (0.13)	0.62 (0.18)	0.64 (0.17)	0.75 (0.16)	0.61 (0.22)	0.74 (0.16)	0.76 (0.13)	0.76 (0.15)	0.81 (0.12)	0.81 (0.13)
nnU-Net-FL	0.40 (0.21)	0.63 (0.22)	0.69 (0.20)	0.76 (0.16)	0.79 (0.14)	0.82 (0.12)	0.84 (0.11)	0.83 (0.11)	0.85 (0.11)	0.85 (0.10)	0.86 (0.09)
Our model	0.70 (0.20)	0.79 (0.16)	0.83 (0.12)	0.83 (0.12)	0.85 (0.10)	0.86 (0.10)	0.86 (0.10)	0.87 (0.09)	0.87 (0.09)	0.87 (0.08)	0.88 (0.08)
HD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	107.82 (27.84)	94.45 (24.74)	101.39 (23.80)	68.56 (33.02)	70.21 (47.44)	65.93 (37.51)	38.46 (40.68)	43.86 (55.43)	32.65 (35.04)	30.17 (40.96)	22.03 (35.38)
UNet-AG-DL	97.40 (24.84)	75.47 (23.43)	74.16 (38.27)	61.92 (31.06)	64.28 (29.87)	76.06 (37.57)	68.53 (36.73)	79.61 (24.04)	42.35 (42.11)	47.00 (42.18)	30.04 (40.76)
UNet-AG-FL	107.92 (19.36)	95.27 (39.72)	99.23 (24.19)	84.76 (32.56)	82.75 (37.11)	59.99 (43.53)	85.27 (30.85)	55.51 (50.84)	39.40 (51.39)	37.28 (46.77)	52.12 (57.16)
nnU-Net-CE	112.50 (17.93)	103.41 (17.56)	91.07 (24.26)	82.57 (29.29)	72.72 (33.37)	76.51 (34.80)	71.05 (35.30)	65.40 (40.10)	62.60 (43.52)	56.76 (53.92)	71.37 (54.68)
nnU-Net-DL	101.57 (21.22)	103.43 (19.52)	93.55 (25.47)	94.41 (24.19)	61.73 (26.86)	90.54 (25.96)	67.71 (23.62)	67.14 (33.58)	66.55 (27.81)	54.49 (37.52)	46.86 (39.04)
nnU-Net-FL	114.06 (14.44)	97.25 (20.27)	97.43 (24.89)	82.64 (31.78)	83.81 (26.30)	80.14 (30.35)	71.82 (37.63)	64.74 (35.10)	70.27 (39.74)	62.10 (45.85)	64.09 (52.62)
Our model	73.20 (29.65)	33.06 (26.07)	26.63 (21.76)	22.64 (19.47)	20.65 (15.81)	15.27 (11.11)	15.54 (11.26)	14.91 (10.96)	14.17 (12.19)	14.86 (10.96)	12.73 (11.23)
ASSD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	11.48 (8.64)	8.35 (7.37)	8.09 (8.11)	3.92 (3.38)	1.96 (2.84)	1.57 (1.95)	1.18 (1.35)	1.31 (1.90)	0.96 (0.96)	1.05 (1.09)	0.96 (0.85)
UNet-AG-DL	13.81 (11.17)	6.60 (5.32)	4.86 (5.02)	3.67 (3.89)	4.40 (3.54)	4.25 (3.89)	3.80 (3.46)	5.86 (6.74)	1.75 (2.05)	1.60 (1.56)	1.26 (1.12)
UNet-AG-FL	17.38 (10.25)	6.94 (9.06)	6.52 (6.27)	3.22 (2.96)	2.21 (2.68)	1.83 (1.97)	3.79 (4.33)	1.40 (1.62)	1.10 (1.05)	0.91 (0.95)	1.23 (1.27)
nnU-Net-CE	25.56 (8.87)	14.47 (8.03)	5.84 (7.03)	3.99 (3.95)	2.03 (2.97)	2.07 (2.16)	1.64 (1.74)	1.48 (1.49)	1.57 (1.67)	1.37 (1.44)	1.25 (1.42)
nnU-Net-DL	16.16 (8.65)	20.67 (8.19)	9.37 (6.47)	8.37 (5.35)	3.32 (2.77)	8.89 (7.88)	4.00 (3.11)	3.65 (3.33)	3.66 (3.40)	2.18 (2.14)	2.14 (1.79)
nnU-Net-FL	30.78 (7.19)	11.38 (6.06)	11.14 (7.65)	3.00 (3.09)	3.39 (3.71)	2.32 (2.11)	1.76 (1.74)	1.71 (1.79)	1.60 (1.90)	1.23 (1.66)	1.13 (1.15)
Our model	4.07 (3.79)	1.68 (1.56)	1.22 (1.06)	1.12 (0.96)	0.92 (0.84)	0.79 (0.67)	0.75 (0.59)	0.71 (0.60)	0.70 (0.60)	0.70 (0.52)	0.66 (0.51)

Table 6
Comparison of the adopted 2.5D model, UNet-AG, and nnU-Net on DSC, HD, and ASSD metrics, evaluated on the AMOS-CT test set of 74 subjects. Metrics are expressed as median with interquartile range (IQR).

DSC	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	0.57 (0.22)	0.68 (0.21)	0.73 (0.22)	0.70 (0.22)	0.78 (0.12)	0.81 (0.11)	0.82 (0.13)	0.83 (0.13)	0.83 (0.12)	0.84 (0.12)	0.85 (0.11)
UNet-AG-DL	0.45 (0.24)	0.65 (0.25)	0.65 (0.24)	0.69 (0.23)	0.63 (0.27)	0.69 (0.25)	0.71 (0.17)	0.70 (0.22)	0.81 (0.13)	0.81 (0.13)	0.85 (0.13)
UNet-AG-FL	0.52 (0.23)	0.65 (0.22)	0.71 (0.14)	0.75 (0.20)	0.80 (0.12)	0.78 (0.17)	0.82 (0.10)	0.84 (0.14)	0.85 (0.13)	0.84 (0.13)	0.84 (0.11)
nnU-Net-CE	0.38 (0.25)	0.54 (0.31)	0.68 (0.23)	0.74 (0.23)	0.81 (0.14)	0.80 (0.14)	0.82 (0.14)	0.83 (0.12)	0.82 (0.14)	0.85 (0.13)	0.84 (0.12)
nnU-Net-DL	0.41 (0.41)	0.38 (0.31)	0.61 (0.26)	0.55 (0.35)	0.63 (0.30)	0.64 (0.23)	0.70 (0.22)	0.73 (0.19)	0.68 (0.24)	0.76 (0.16)	0.78 (0.19)
nnU-Net-FL	0.21 (0.19)	0.43 (0.42)	0.61 (0.24)	0.74 (0.15)	0.78 (0.15)	0.80 (0.13)	0.82 (0.12)	0.81 (0.15)	0.83 (0.12)	0.80 (0.16)	0.83 (0.14)
Our model	0.68 (0.36)	0.78 (0.28)	0.82 (0.13)	0.84 (0.12)	0.85 (0.09)	0.86 (0.08)	0.86 (0.09)	0.86 (0.08)	0.87 (0.09)	0.87 (0.08)	0.87 (0.09)
HD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	79.39 (49.43)	80.35 (46.53)	75.61 (48.36)	63.51 (27.99)	45.55 (49.25)	53.70 (53.14)	19.50 (29.53)	35.98 (45.24)	19.59 (17.97)	18.00 (32.47)	19.01 (25.85)
UNet-AG-DL	76.60 (34.13)	60.04 (31.96)	55.30 (28.52)	55.83 (25.59)	61.04 (29.67)	40.78 (39.31)	53.73 (42.07)	69.54 (27.23)	31.11 (32.06)	34.61 (46.71)	28.81 (38.50)
UNet-AG-FL	84.79 (48.79)	70.87 (33.17)	73.08 (38.77)	67.86 (41.80)	69.57 (36.17)	45.65 (41.91)	43.99 (53.53)	28.32 (38.91)	21.48 (41.71)	14.23 (21.52)	18.34 (37.26)
nnU-Net-CE	86.37 (52.75)	82.10 (42.21)	67.38 (33.22)	65.66 (35.87)	39.57 (53.35)	44.49 (39.11)	31.59 (35.63)	32.80 (44.91)	29.72 (39.42)	26.29 (41.46)	41.31 (57.03)
nnU-Net-DL	74.35 (39.67)	81.79 (34.50)	69.43 (33.57)	73.40 (48.48)	57.27 (25.97)	60.39 (33.32)	54.08 (23.73)	60.67 (30.93)	54.19 (31.46)	39.41 (37.97)	35.86 (31.96)
nnU-Net-FL	94.71 (56.35)	70.69 (32.81)	85.88 (31.35)	69.79 (42.59)	58.71 (41.96)	46.32 (44.60)	45.88 (51.04)	35.02 (40.95)	29.29 (28.92)	24.67 (31.11)	35.82 (49.82)
Our model	54.72 (34.46)	23.66 (19.88)	17.72 (12.08)	14.97 (12.85)	12.86 (9.06)	11.93 (6.45)	11.74 (7.74)	11.67 (7.95)	11.25 (6.55)	11.91 (8.85)	10.64 (6.44)
ASSD [mm]	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
UNet-AG-CE	7.00 (7.35)	6.38 (5.56)	3.59 (4.43)	4.47 (4.36)	1.70 (1.71)	1.59 (2.05)	1.30 (0.90)	1.26 (1.26)	1.25 (0.77)	1.12 (1.28)	1.02 (1.11)
UNet-AG-DL	8.52 (8.68)	4.62 (5.09)	4.12 (3.69)	3.26 (2.91)	6.47 (5.24)	3.52 (3.67)	3.31 (3.86)	4.95 (6.53)	1.70 (1.39)	1.72 (1.34)	1.30 (1.00)
UNet-AG-FL	13.18 (6.97)	5.90 (6.32)	3.91 (3.17)	2.46 (2.13)	1.83 (1.93)	1.74 (1.62)	1.61 (1.59)	1.40 (1.13)	1.26 (0.93)	1.15 (0.99)	1.03 (1.01)
nnU-Net-CE	20.23 (8.54)	13.57 (10.11)	3.45 (4.76)	3.25 (3.34)	1.49 (1.24)	1.69 (1.84)	1.34 (1.18)	1.34 (1.05)	1.34 (1.05)	1.32 (0.94)	1.22 (0.89)
nnU-Net-DL	13.26 (10.14)	21.12 (12.25)	5.39 (4.78)	7.24 (6.68)	4.54 (3.96)	5.39 (5.03)	3.67 (3.39)	3.32 (3.72)	3.74 (3.55)	2.20 (1.92)	1.90 (1.53)
nnU-Net-FL	30.79 (11.71)	11.02 (8.79)	8.51 (6.31)	2.59 (2.26)	2.55 (2.38)	1.85 (1.77)	1.54 (1.45)	1.42 (1.26)	1.32 (1.03)	1.47 (1.23)	1.30 (1.11)
Our model	3.47 (3.35)	1.56 (1.59)	1.14 (0.53)	1.08 (0.76)	0.97 (0.56)	0.86 (0.57)	0.88 (0.49)	0.81 (0.48)	0.84 (0.50)	0.81 (0.48)	0.82 (0.52)

Table 7
Slices with errors on the 130 subjects of the test set of AIMS-1300 dataset for the different partitions of the training dataset. Results for the proposed 2.5D UNet. Data are grouped by model segmentation (MS), ground truth (GT), and total number of slices with errors. The best values are in bold.

Slices	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
MS-only	1566	716	496	442	396	288	281	**232**	261	263	274
GT-only	89	79	110	112	102	115	**82**	110	95	93	97
Total error slices	1655	795	606	554	498	403	363	**342**	356	356	371

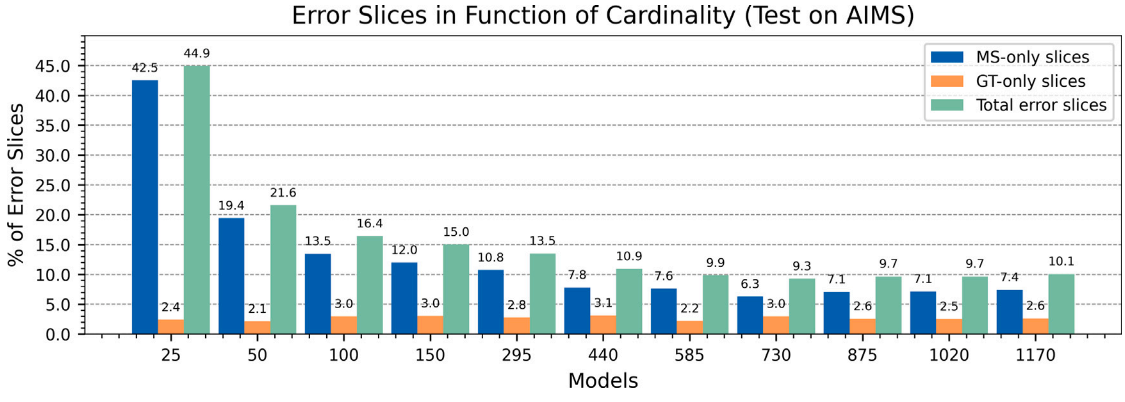


Fig. 5. Trend of error slices on the 130 subjects of the test set of AIMS-1300, expressed as percentages of the total pancreatic slices (3683 slices).

Table 8
Slices with errors on the 74 AP subjects of the test set of AMOS-CT dataset. Results for the proposed 2.5D UNet. Data are grouped by model segmentation (MS), ground truth (GT), and total number of slices with errors. The best values are in bold.

Slices	m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
MS-only	500	183	168	163	101	96	119	**86**	98	136	118
GT-only	86	89	58	68	55	55	**52**	71	67	45	59
Total error slices	586	272	226	231	156	**151**	171	157	165	181	177

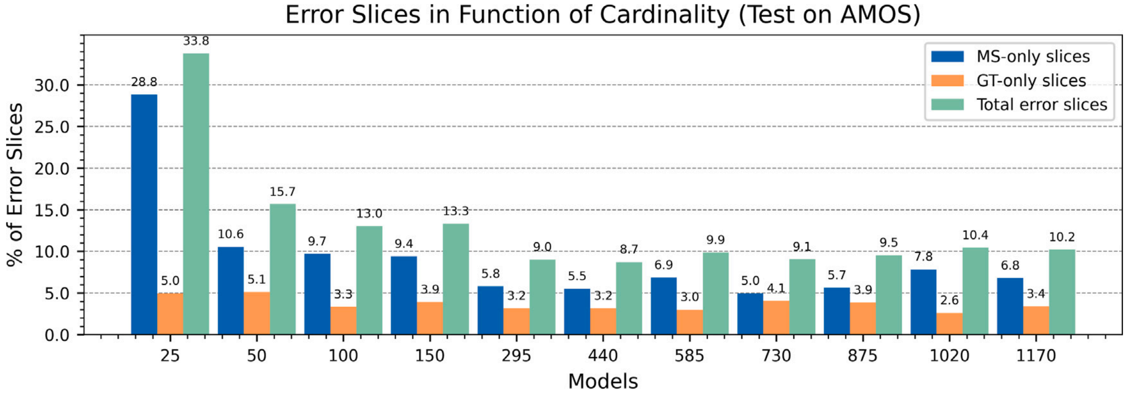


Fig. 6. Trend of error slices on the 74 subjects of the test set of AMOS-CT, expressed as percentages of the total pancreatic slices (1734 slices).

of these slices provided a quantitative measure of the extent of modifications. Multiplying this measure by the estimated time required to correct a single slice leads to the time requested to correct an entire volume or batch. By following this approach, we could quantitatively assess the relationship between the dataset size and the amount of manual correction required after model inference. The takeaways messages in this section are as follows:

1. Larger datasets improve the accuracy and reduce the time and extent of manual corrections in DL models for radiology.
2. Quantitative analysis of slices of error frequency helps determine the optimal dataset size for robust and efficient radiological and clinical applications.

6.1. Theoretical motivation

The motivations behind this analysis were twofold: to assess model reliability and to quantify how errors impact clinical workflows. While the PR, RE, and RVD metrics effectively measure over-segmentation and under-segmentation errors, they fall in quantifying the time and extent required by manual correction, as this correction is typically applied on 2D slices. This analysis is valuable for targeting artificial intelligence models to make them clinically effective, enhance model reliability, and optimize clinical costs. Moreover, identifying the ideal dataset size for DL models is pivotal in balancing automated efficiency and minimal manual intervention.

6.2. Results

Inspired by the theoretical motivation above, we conducted investigations to understand how the size of the dataset influences the need for manual corrections in image segmentation, answering the **RQ3**. More specifically, after overlapping the predicted mask and GT of each slice of the test dataset, the number of slices showing only the ground truth (GT-only slices) or only the predicted model segmentations (MS-only slices) were counted. GT-only slices highlight under-segmentation errors, while MS-only slices point out over-segmentation errors. The sum of GT-only slices and MS-only slices yielded the total number of slices with errors. Fig. 5 displays the percentage of error slices in the 130-subjects test set of AIMS-1300, evaluated upon the total number of pancreas slices included in the GT images (3683 pancreas GT slices out of 20,109 total CTA slices). The actual number of these slices is systematically reported in Table 7, providing a precise count and a more quantitative view of the errors observed in the dataset. It can be noted that as the dataset size increased, the number of error slices decreased significantly. The most significant reduction occurred between m25 and m50, where the total error slices were halved from 1655 to 795. From m50 to m440, the number of error slices decreased from 795 to 403. After that, the rate of errors slowed down, as can be appreciated in Fig. 5. The model with the lowest total error slices was m730, producing 342 total error slices representing 9.3% of the total pancreas slices. Notably, over 70% of the error slices were attributed to slices with only automatic segmentation. Interestingly, the increase in dataset size was perceived more significantly in the MS-only slices, indicating a greater reduction in over-segmentation compared to under-segmentation as the dataset size increases. To generalize our findings, we performed the same evaluation on the 74-subjects test set of AMOS-CT. Fig. 6 displays the percentage of error slices evaluated upon the total number of pancreas slices included in the GT images (1734 pancreas GT slices out of 11,678 total CTA slices), while Table 8 provides a precise count of the slices. It is interesting to note that also in this case, the MS-only slices benefited more from an increase in dataset size. For AMOS-CT, the cardinality that resulted in the fewest errors was m440, with 151 total error slices representing 8.7% of the total pancreas slices.

7. Anatomical subregions (RQ4)

To expand our approach to optimal cardinality, we analyzed the performances of the proposed 2.5D UNet on another segmentation task, i.e. the segmentation of the pancreas subregions (**RQ4**). The pancreas subregions are head, body, and tail, with the head showing greater anatomical variability than the body and tail, and being the subregion where the majority (60%–70%) of pancreas cancers occur (Sureka et al., 2021; Vareedayah et al., 2018). To address **RQ4**, the test set of AIMS-1300 with 130 CTs was divided into three equal longitudinal subregions along its main axis: head, body, and tail, as depicted in Fig. 7. The 2.5D UNet network was then tested on these regions, allowing for a comprehensive examination of their performance. The main takeaways messages from this section are summarized below:

1. The greater the complexity and anatomical variability of the subregion, the more significant the impact of a large dataset on the accuracy of automatic predictions.

Table 9
Distribution of the selected metrics on the test set of 130 subjects of AIMS-1300 dataset for the proposed 2.5D UNet, trained with different dataset sizes. Test volumes were divided into three pancreas subregions (head (P1), body (P2), and tail (P3)). Data expressed as median with interquartile range (IQR), with better values in bold.

Median (IQR)		m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
DSC	P1	0.64 (0.22)	0.76 (0.20)	0.81 (0.19)	0.80 (0.20)	0.83 (0.14)	0.84 (0.15)	0.85 (0.14)	0.85 (0.14)	0.85 (0.16)	0.85 (0.13)	**0.85 (0.13)**
	P2	0.72 (0.16)	0.80 (0.13)	0.85 (0.10)	0.85 (0.09)	0.87 (0.08)	0.87 (0.06)	0.87 (0.07)	0.88 (0.06)	0.88 (0.06)	0.88 (0.05)	**0.88 (0.07)**
	P3	0.78 (0.18)	0.83 (0.16)	0.86 (0.12)	0.87 (0.12)	0.87 (0.10)	0.88 (0.09)	0.89 (0.08)	0.89 (0.08)	0.89 (0.08)	0.89 (0.07)	**0.89 (0.07)**
PR	P1	0.64 (0.28)	0.76 (0.20)	0.81 (0.15)	0.80 (0.13)	0.83 (0.13)	0.84 (0.13)	0.85 (0.13)	0.85 (0.12)	0.85 (0.12)	0.85 (0.14)	**0.85 (0.12)**
	P2	0.72 (0.18)	0.80 (0.14)	0.85 (0.11)	0.85 (0.10)	0.87 (0.10)	0.87 (0.10)	0.87 (0.10)	0.88 (0.09)	0.88 (0.09)	0.88 (0.09)	**0.88 (0.10)**
	P3	0.78 (0.15)	0.83 (0.15)	0.86 (0.13)	0.87 (0.12)	0.87 (0.11)	0.88 (0.09)	0.89 (0.08)	0.89 (0.09)	0.89 (0.09)	0.89 (0.09)	**0.89 (0.09)**
RE	P1	0.81 (0.33)	0.78 (0.31)	0.80 (0.23)	0.80 (0.28)	0.83 (0.21)	0.82 (0.18)	0.85 (0.14)	0.82 (0.17)	0.83 (0.17)	0.84 (0.15)	**0.85 (0.15)**
	P2	0.86 (0.21)	0.84 (0.19)	0.86 (0.16)	0.85 (0.14)	0.88 (0.10)	0.89 (0.10)	0.90 (0.10)	0.89 (0.09)	0.89 (0.10)	0.89 (0.08)	**0.90 (0.08)**
	P3	0.85 (0.23)	0.86 (0.19)	0.88 (0.13)	0.88 (0.12)	0.88 (0.10)	0.88 (0.10)	0.89 (0.11)	0.88 (0.10)	0.89 (0.10)	0.89 (0.08)	**0.89 (0.08)**
RVD	P1	27.19 (88.08)	−2.38 (53.96)	−7.87 (36.51)	−8.98 (37.22)	−6.02 (29.88)	−5.17 (27.69)	−3.27 (22.36)	−6.15 (22.23)	−6.48 (22.33)	−4.73 (22.90)	**−2.21 (23.02)**
	P2	26.12 (54.97)	4.48 (33.08)	−1.54 (22.42)	−3.02 (21.67)	1.51 (19.04)	**0.40 (16.57)**	1.18 (17.61)	−0.67 (15.17)	−0.75 (16.52)	1.27 (16.20)	1.82 (15.80)
	P3	15.46 (38.24)	0.09 (27.39)	−2.16 (22.72)	−1.49 (19.96)	−0.69 (17.84)	−2.81 (14.89)	**−0.08 (16.06)**	−1.11 (15.14)	−1.86 (12.56)	−0.09 (14.50)	−0.55 (15.69)
HD [mm]	P1	72.22 (49.27)	19.57 (23.77)	14.23 (13.98)	13.37 (13.04)	10.83 (8.83)	9.96 (6.59)	10.12 (7.56)	9.97 (7.37)	**9.38 (6.18)**	10.15 (6.65)	9.52 (5.77)
	P2	47.87 (32.81)	20.59 (18.59)	16.89 (16.09)	16.19 (12.61)	13.43 (11.69)	11.61 (8.20)	11.36 (9.25)	10.88 (8.44)	11.01 (9.01)	10.87 (7.74)	**10.63 (6.43)**
	P3	35.38 (29.29)	19.16 (15.20)	17.38 (14.34)	14.37 (11.45)	12.30 (10.25)	11.77 (7.81)	10.69 (6.95)	10.82 (7.78)	10.34 (7.09)	11.05 (6.62)	**10.07 (7.31)**
ASSD [mm]	P1	4.71 (6.29)	1.75 (1.79)	1.28 (1.40)	1.28 (1.49)	1.02 (0.84)	0.93 (0.76)	0.90 (0.76)	0.88 (0.75)	0.87 (0.66)	0.86 (0.62)	**0.81 (0.65)**
	P2	3.17 (3.39)	1.50 (1.47)	1.06 (0.95)	1.04 (0.84)	0.89 (0.66)	0.79 (0.54)	0.76 (0.51)	0.71 (0.47)	0.72 (0.52)	0.72 (0.45)	**0.67 (0.43)**
	P3	2.17 (1.47)	1.14 (1.24)	0.84 (0.78)	0.76 (0.68)	0.66 (0.63)	0.61 (0.55)	0.60 (0.48)	0.57 (0.43)	0.57 (0.43)	**0.56 (0.43)**	0.57 (0.42)

Table 10
Distribution of the selected metrics on the test set of 74 AP subjects of AMOS-CT dataset for the proposed 2.5D UNet, trained with different dataset sizes. Test volumes were divided into three pancreas subregions (head (P1), body (P2), and tail (P3)). Data expressed as median with interquartile range (IQR), with better values in bold.

Median (IQR)		m25	m50	m100	m150	m295	m440	m585	m730	m875	m1020	m1170
DSC	P1	0.60 (0.47)	0.73 (0.33)	0.80 (0.26)	0.80 (0.27)	0.80 (0.30)	0.83 (0.24)	0.85 (0.19)	0.84 (0.23)	0.84 (0.22)	**0.85 (0.16)**	0.85 (0.17)
	P2	0.70 (0.24)	0.78 (0.21)	0.83 (0.15)	0.84 (0.14)	0.85 (0.14)	0.85 (0.14)	0.87 (0.13)	0.86 (0.14)	0.86 (0.13)	**0.86 (0.11)**	0.86 (0.13)
	P3	0.68 (0.39)	0.76 (0.30)	0.79 (0.22)	0.81 (0.20)	0.81 (0.22)	0.83 (0.22)	0.84 (0.17)	0.84 (0.17)	0.83 (0.18)	**0.84 (0.15)**	0.84 (0.19)
PR	P1	0.72 (0.36)	0.84 (0.21)	0.86 (0.16)	0.87 (0.15)	0.90 (0.14)	0.89 (0.14)	0.88 (0.13)	0.90 (0.12)	0.89 (0.13)	0.84 (0.15)	**0.88 (0.13)**
	P2	0.71 (0.22)	0.82 (0.17)	0.83 (0.13)	0.83 (0.14)	0.86 (0.12)	0.86 (0.11)	0.86 (0.12)	0.88 (0.10)	0.87 (0.11)	0.83 (0.12)	**0.85 (0.11)**
	P3	0.72 (0.28)	0.82 (0.21)	0.81 (0.18)	0.82 (0.18)	0.87 (0.14)	0.88 (0.14)	0.88 (0.13)	0.86 (0.14)	0.85 (0.14)	0.84 (0.16)	**0.85 (0.15)**
RE	P1	0.59 (0.64)	0.70 (0.48)	0.78 (0.40)	0.81 (0.43)	0.74 (0.46)	0.79 (0.38)	0.90 (0.31)	0.84 (0.38)	0.85 (0.36)	**0.92 (0.24)**	0.88 (0.25)
	P2	0.76 (0.40)	0.79 (0.33)	0.88 (0.24)	0.91 (0.20)	0.88 (0.24)	0.89 (0.24)	0.92 (0.18)	0.90 (0.22)	0.91 (0.21)	**0.94 (0.15)**	0.92 (0.19)
	P3	0.74 (0.55)	0.78 (0.46)	0.85 (0.34)	0.86 (0.30)	0.82 (0.37)	0.85 (0.33)	0.87 (0.28)	0.89 (0.23)	0.90 (0.25)	**0.91 (0.20)**	0.90 (0.24)
RVD	P1	−17.86 (74.14)	−15.09 (58.47)	−9.97 (45.39)	−8.16 (50.68)	−17.70 (52.71)	−9.34 (42.49)	2.67 (38.69)	−5.38 (45.98)	−4.50 (40.46)	8.50 (31.90)	**0.24 (30.83)**
	P2	3.48 (62.06)	−3.17 (45.69)	3.66 (30.51)	6.91 (26.35)	2.39 (32.74)	**2.35 (31.30)**	5.47 (25.88)	3.21 (28.54)	3.25 (29.43)	10.28 (23.69)	6.90 (26.85)
	P3	1.75 (80.98)	−1.54 (57.94)	2.57 (46.95)	5.62 (43.29)	−4.97 (43.02)	−2.66 (38.24)	**−0.90 (34.20)**	3.48 (29.42)	4.56 (33.84)	8.36 (28.37)	5.94 (32.23)
HD [mm]	P1	22.17 (39.56)	11.08 (9.53)	8.70 (6.00)	8.07 (6.22)	8.15 (7.20)	7.29 (6.72)	6.76 (5.60)	6.76 (6.29)	7.08 (5.76)	7.12 (6.38)	**6.38 (6.14)**
	P2	40.01 (50.95)	16.78 (21.66)	12.88 (13.13)	12.10 (14.53)	11.25 (12.23)	11.02 (11.61)	**9.56 (11.11)**	9.90 (10.24)	9.59 (10.91)	10.56 (11.01)	10.07 (10.81)
	P3	29.43 (33.12)	16.91 (20.00)	14.20 (13.55)	12.37 (11.26)	11.19 (11.21)	10.41 (10.52)	**9.56 (8.95)**	9.63 (8.86)	10.13 (10.28)	9.91 (8.72)	9.91 (9.47)
ASSD [mm]	P1	4.53 (7.64)	2.64 (2.74)	1.96 (1.84)	1.86 (1.95)	1.89 (2.03)	1.63 (1.73)	1.40 (1.46)	1.50 (1.73)	1.49 (1.43)	1.44 (1.31)	**1.40 (1.38)**
	P2	5.50 (7.30)	2.87 (3.60)	2.08 (2.01)	1.98 (2.14)	1.89 (1.75)	1.79 (1.87)	**1.60 (1.48)**	1.63 (1.62)	1.66 (1.55)	1.74 (1.55)	1.73 (1.46)
	P3	4.35 (5.05)	2.82 (3.04)	2.36 (1.93)	2.16 (1.73)	2.05 (1.86)	1.81 (1.57)	**1.71 (1.46)**	1.74 (1.49)	1.77 (1.56)	1.76 (1.38)	1.74 (1.51)

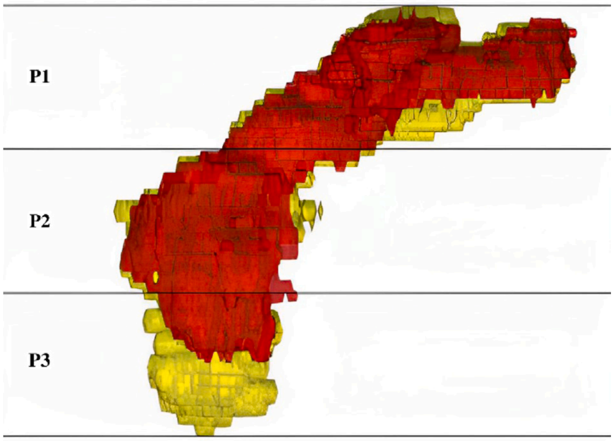


Fig. 7. Example of GT (red) and MS (yellow) volumes divided into three pancreas subregions: head (P1), body (P2), and tail (P3).

2. This analysis allowed for a qualitative transposition of the previously obtained results to structures with different degrees of complexity.
3. Models with lower cardinality were more prone to errors in the head subregion. Increasing cardinality yields more significant benefits for the head subregion, thus improving performance.

7.1. Theoretical motivation

The motivations behind our analysis of anatomical variability shared a common goal of optimizing resource allocation in medical imaging. By investigating the relationship between anatomical variability and dataset size in the pancreas, we aimed to establish a methodological framework for conducting similar assessments of other anatomical structures. This approach was expected to provide insights into efficient data collection and annotation strategies to improve the ability of algorithms to handle structural variability and to simplify model evaluation processes in other domains. Furthermore, understanding the quantitative trends linking the anatomical variability of pancreas subregions to the dataset size could provide qualitative guidance for transferring results to different structures. This provided a valuable reference point for future research on segmenting structures with varying anatomical complexity, including those beyond the pancreas.

7.2. Results

The metrics values in Table 9 obtained evaluating the eleven models on 130 subjects of the test set of AIMS-1300 improved as the training dataset size increased. These results are reported using the median and interquartile range. Notably, P1 consistently underperformed compared to P2 and P3, as evidenced in Table 9. In models m730, m875, m1020, and m1170, the median DSC for P1 is 0.85(0.13), slightly below the median DSC values for P2 and P3, which were 0.88(0.07) and 0.89(0.07), respectively. PR showed the same results as DSC, while the maximum median RE for P1 was 0.85(0.13), for P2 was 0.90(0.08), and for P3 was 0.89(0.08), both for m1170. RVD showed better performances in P1 with −2.21(23.02) for m1170, in P2 with 0.40(16.57) for m440, and in P3 with −0.08(16.06) for m585. HD showed better performances in P1 with 9.38(6.18) mm for m875, in P2 and P3 with 0.63(6.43) mm and 10.07(7.31) mm for m1170. Instead, ASSD showed better performances in P3 with 0.56(0.43) mm for m1020, and in P1 and P2 with 0.81(0.65) mm and 0.67(0.43) mm for m1170. Interestingly, P1 showed a larger variation than P2 and P3 in the values of the metrics in the different models. Specifically, DSC, PR, HD, and ASSD of P1 greatly improved by the increase in dataset size. Both P2 and P3 benefit from increased dataset size, albeit to a lesser extent. To assess the reproducibility and generalizability of our findings, we evaluated the eleven trained models on the 74 AP subjects from the AMOS-CT test set. The results, as detailed in Table 10, showed similar patterns to those observed with the AIMS test set. Notably, the head subregions presented marked improvements with increased training dataset size.

8. Discussion and future directions

8.1. Dataset acquisition

Significant findings emerged from thoroughly analyzing AIMS Academy's data collection and annotation procedures. These insights are particularly relevant when contextualized in preoperative planning and medical image processing. Central to this discussion is manual data collection and annotation, avoiding weakly supervised methods. Paullada et al. (2021) emphasized that manually annotated datasets, albeit more expensive in terms of time, money, and effort, are pivotal in clinical contexts. Moreover, manual segmentation is still regarded as the most reliable method, although at the expense of scalability (Tajbakhsh et al., 2020; Roh et al., 2019; Arora et al., 2023). The initial phase of DICOM image acquisition underlines the critical importance of anonymization and supplier agreements. This phase encompasses two distinct stages: firstly, identifying suitable providers, a process often extended due to privacy policies and data management regulations; secondly, the bureaucratic time involved, which includes negotiating agreements and ensuring compliance with data protection standards. These stages, especially when combined, substantially extend the overall duration required to initiate research projects, as discussed by Roh et al. (2019). Our analysis shows an extensive 80-hour training to mitigate intra-observer and inter-observer variability, a well-documented limitation in medical image annotation (Renard et al., 2020). This training period was essential to ensure the reliability of the manual segmentation process. As stated by Roh et al. (2019), the human effort required during the annotation process was challenging to quantify. Our findings show that the time needed for pancreas segmentation varied significantly with the resolution of the CTA images, given that annotation was performed slice-by-slice on the 3D volumes. Our analysis provided insight that annotating a dataset of 1300 pancreatic volumes required 938 h, including the 80 h dedicated to training annotators on the segmentation protocol. Lastly, incorporating verification by domain experts, such as clinicians and radiologists, although time-intensive, significantly enhanced the reproducibility and reliability of the dataset. This step, often overlooked in data collection methodologies, is paramount for the clinical applicability of the dataset (Paullada et al., 2021).

8.2. Dataset size and accuracy

Given the approach taken in this study, which explored the relationship between training dataset size and performance metrics, the results obtained offered a significant contribution to the field of DL in medical image analysis (Fang et al., 2021). Our work is among the first ones aiming to answer the unsolved question regarding the optimal dataset size. A previous study on 1917 CTs on the pancreas compared different training subsets (from 200 to 1628 CTs) on the same testing set of 289 CTs, reporting a DSC ranging from 0.74 to 0.91 (Panda et al., 2021). These findings were crucial for supervised DL methodologies and may provide a target dataset size for data augmentation. We thoroughly assessed a comprehensive set of metrics across models trained with different dataset sizes, yielding several notable observations. As the dataset size increased, DSC improved significantly, from 0.70 to 0.89, PR improved from 0.83 to 0.89, and RE improved from 0.64 to 0.88. The improvements in these three metrics mean an increased overlap between the model's predictions and the ground truth, reducing over- and under-segmentation errors (Taha and Hanbury, 2015). The reduction in RVD from 26.36 to 1.00 indicates that increased dataset size resulted in more precise volumetric predictions. The decrease in HD from 73.20 mm to 12.73 mm and the ASSD from 4.07 mm to 0.66 mm indicated that increasing the dataset size also enhanced the accuracy of the predicted edges. Overall, an improvement trend was observed across all metrics, meaning larger datasets yield more accurate results. Further investigation into the metric distributions of the test dataset revealed that increased dataset size resulted in a progressive narrowing of these distributions. This observation highlighted that enlarging the dataset cardinality improved the model's generalizability, thereby reducing errors on unseen data, as Freiesleben and Grote (2023) noted. Further observations showed that the trend of the overall metrics relative to dataset cardinality showed only marginal increases in accuracy beyond a specific size. As pointed out by Roh et al. (2019), it could be challenging to determine whether the data and labels are sufficient. This left the critical issue of assessing whether data had been collected in sufficient volume unsolved. Our study provided an objective answer to this question in the context of pancreatic segmentation, revealing that increasing the dataset size undoubtedly improved the network's performance, thus increasing its generalizability. Nevertheless, it was observed that the extent of improvement significantly diminished beyond an identified optimal dataset cardinality. Using the Kruskal–Wallis statistical test with post-hoc comparison ($p < 0.05$) enabled us to determine the optimal cardinality for pancreatic segmentation. According to the most restrictive metric (ASSD), this optimal range fell to around 400 subjects. Above this dataset size, almost all metrics started to reduce their improvement rate and were statistically identical, within a 5% significance level. Remarkably, our findings on dataset cardinality showed that the proposed 2.5D UNet achieved comparable results to those of nnU-Net and UNet-AG with only half the sample size.

8.3. Manual refinement of the results

The examination of error slices served not only as an extra metric for assessing the model's accuracy but, more notably, as a quantitative measure for estimating the time and extent of manual refinement needed to correct the automatic segmentation of the pancreas. According to Sander et al. (2020), identifying slices with segmentation errors is key to achieving accurate segmentation in clinical settings. This process aids in the manual correction of automatic segmentations in daily practice. Our analysis aligned with this perspective, revealing a trend of reducing error slices as dataset sizes increased, which was consistent with the observed trends in the reported metrics. This observation suggested that more extensive datasets enhanced model accuracy, decreasing the frequency and complexity of manual interventions required. Furthermore, our work introduced the concept of using error slice evaluation as a guiding metric to minimize manual correction. As a result, it allowed us to quantify the time for manual correction of an entire volume or batch, equivalent to the time required for the radiologist correction of one single 2D slice, multiplied by the number of slices with errors. Our approach, firstly evaluated on the test portion of the AIMS-CT dataset, was corroborated by assessing its generalization to an external dataset, AMOS-CT in our case. Overall, this methodology in medical image segmentation research offered a practical measure for assessing model efficiency, especially in balancing automatic accuracy with the need for manual refinement. Our findings agreed with those reported by Sander et al. (2020), where a higher number of detected slices containing segmentation errors corresponded to an increased workload for manual correction.

8.4. Anatomical subregions

In the literature, published works on optimal cardinality are sporadic, e.g., the one on pancreas segmentation on 1917 CTs (Panda et al., 2021). However, to date, no studies have been published on the segmentation of pancreas subregions. In our work, we aimed to extend qualitatively our approach to other segmentation tasks, exemplified by pancreas subregions (RQ4). Additionally, by recognizing the lack of an effective method for evaluating segmentation results (Renard et al., 2020; Villarini et al., 2021), we proposed a novel approach to the assessment of DL models for pancreas segmentation based on its subregions. Our findings highlighted the influence of anatomical variability on the performance of DL models for the segmentation of pancreas subregions. The results across all models' cardinalities showed a lower performance of the pancreas head than the body and tail in both the tested datasets, namely AIMS-1300 and AMOS-CT. This finding could be associated with the complexity and variability of the shape of the head subregion, in agreement with Sureka et al. (2021). This was further supported by the observation that the head subregion exhibited more significant performance improvements with increasing dataset size, making it the most challenging subregion for the segmentation task. Consequently, it may be advisable to increase the dataset cardinality for those tasks targeting this subregion.

8.5. Limitations and future directions

While this study provided valuable insights into the segmentation of the pancreas using DL, it encountered certain limitations that suggested directions for future research. The focus on the pancreas might limit the generalizability of the findings to other anatomical structures, underlining the need for further studies across diverse abdominal organs. While innovative, the approach of using error slice evaluation as a metric for manual refinement may only partially capture the complexities of manual corrections required for more intricate or subtle segmentation errors. Beyond dataset size, future research should explore the impact of image quality, resolution, and patient variability on segmentation accuracy. These factors are likely to play a crucial role in real-world performance. Addressing these limitations could broaden the understanding of DL model performance in medical image segmentation and enhance their practical applicability in different clinical settings. Future studies might also incorporate advanced DL techniques like transfer learning to refine the segmentation process further, especially in resource-constrained environments. Lastly, while our study made significant contributions to the field of medical image analysis, we acknowledge a limitation in our research methodology. The scope of our evaluation did not extend to 3D network models due to their computational cost. This aspect remains an area for future exploration, as 3D networks offer distinct capabilities and challenges.

9. Conclusions

This study advanced the field of DL for medical image analysis of the pancreas by addressing many open questions. Published evidence highlighted a notable scarcity of datasets and a tendency to prioritize technical over clinical validation, often deviating from the fundamental objective of medical imaging, i.e., improving patient care. Through our findings, we provided useful insights for effective DL tasks, offering a balance between resource optimization and improved outcomes. Our analysis yielded practical tools for the design of DL segmentation of the pancreas. Our analysis of dataset acquisition offered valuable perspectives on the resources required for manual collection. We further investigated the influence of dataset size on the performance of DL models, identifying 440 subjects as the dataset cardinality where the rate of accuracy improvement began to plateau. These results could be further exploited to refine potential DL strategies targeting the optimal cardinality instead of huge augmentations. Insights into manual refinements to meet expert consensus were provided by monitoring error slices. Furthermore, our examination of pancreatic subregions not only deepened our understanding of pancreatic complexity but also introduced qualitative metrics that could be applied to other organs. Lastly, our study bridged the gap between technical innovation and clinical utility, empowering computer scientists, clinicians, and stakeholders to make informed decisions with the ultimate goal of enhancing patient care.

CRediT authorship contribution statement

Matteo Cavicchioli: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Andrea Moglia:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Ludovica Pierelli:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Giacomo Pugliese:** Data curation, Funding acquisition, Writing – review & editing, Conceptualization. **Pietro Cerveri:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The dataset is available upon request. The code is available on GitHub at the following url: https://github.com/hal9000-lab/Pancreas_2.5D.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors did not use any tool or software based on generative AI and AI-assisted technologies.

Acknowledgments

This work was funded by Fondazione MIAS (AIMS Academy) and the FAIR project in the PNRR-PE program of the Italian Ministry of University and Research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2024.102434.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL https://www.tensorflow.org/, Software available from tensorflow.org.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8, 1–74.

Arora, A., Alderman, J.E., Palmer, J., Ganapathi, S., Laws, E., McCradden, M.D., Oakden-Rayner, L., Pfohl, S.R., Ghassemi, M., McKay, F., et al., 2023. The value of standards for health datasets in artificial intelligence-based applications. Nat. Med. 1–10.

Bansal, M.A., Sharma, D.R., Kathuria, D.M., 2022. A systematic review on data scarcity problem in deep learning: solution and applications. ACM Comput. Surv. 54 (10s), 1–29.

Chollet, F., et al., 2015. Keras. https://keras.io.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp. 424–432.

Cobo, M., Menéndez Fernández-Miranda, P., Bastarrika, G., Lloret Iglesias, L., 2023. Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows. Sci. Data 10 (1), 732.

Dai, S., Zhu, Y., Jiang, X., Yu, F., Lin, J., Yang, D., 2023. TD-Net: Trans-Deformer network for automatic pancreas segmentation. Neurocomputing 517, 279–293.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 179–187.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24–29.

Fang, Y., Wang, J., Ou, X., Ying, H., Hu, C., Zhang, Z., Hu, W., 2021. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys. Med. Biol. 66 (18), 185012.

Fantazzini, A., Esposito, M., Finotello, A., Auricchio, F., Pane, B., Basso, C., Spinella, G., Conti, M., 2020. 3D automatic segmentation of aortic computed tomography angiography combining multi-view 2D convolutional neural networks. Cardiovasc. Eng. Technol. 11, 576–586.

Freiesleben, T., Grote, T., 2023. Beyond generalization: a theory of robustness in machine learning. Synthese 202 (4), 109.

Garcea, F., Serra, A., Lamberti, F., Morra, L., 2022. Data augmentation for medical imaging: A systematic literature review. Comput. Biol. Med. 106391.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.

Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., Parasa, S., 2022. On evaluation metrics for medical applications of artificial intelligence. Sci. Rep. 12 (1), 5979.

Huang, M., Huang, C., Yuan, J., Kong, D., 2021. A semiautomated deep learning approach for pancreas segmentation. J. Healthc. Eng. 2021.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Commun. 18 (2), 203–211. http://dx.doi.org/10.1038/s41592-020-01008-z.

Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhannng, L., Ma, W., Wan, X., et al., 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Adv. Neural Inf. Process. Syst. 35, 36722–36732.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P., Acharya, U.R., 2022. Transfer learning techniques for medical image analysis: A review. Biocybern. Biomed. Eng. 42 (1), 79–107.

Li, J., Chen, T., Qian, X., 2022. Generalizable pancreas segmentation modeling in CT imaging via meta-learning and latent-space feature flow generation. IEEE J. Biomed. Health Inf. 27 (1), 374–385.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J., 2019. Dice loss for data-imbalanced NLP tasks. http://dx.doi.org/10.48550/ARXIV.1911.02855.

Li, L., Zhao, H., Wang, H., Li, W., Zheng, S., 2023a. Automatic abdominal segmentation using novel 3D self-adjustable organ aware deep network in CT images. Biomed. Signal Process. Control 84, 104691.

Li, J., Zhu, H., Chen, T., Qian, X., 2023b. Generalizable pancreas segmentation via a dual self-supervised learning framework. IEEE J. Biomed. Health Inf..

Liew, C., 2018. The future of radiology augmented with artificial intelligence: a strategy for success. Eur. J. Radiol. 102, 152–156.

Lim, S.-H., Kim, Y.J., Park, Y.-H., Kim, D., Kim, K.G., Lee, D.-H., 2022. Automated pancreas segmentation and volumetry using deep neural network on computed tomography. Sci. Rep. 12 (1), 4075.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. http://dx.doi.org/10.48550/ARXIV.1708.02002.

Liu, S., Liang, S., Huang, X., Yuan, X., Zhong, T., Zhang, Y., 2022. Graph-enhanced U-Net for semi-supervised segmentation of pancreas from abdomen CT scan. Phys. Med. Biol. 67 (15), 155017.

Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2021. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. arXiv preprint arXiv:2111.02403.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6695–6714.

Man, Y., Huang, Y., Feng, J., Li, X., Wu, F., 2019. Deep Q learning driven CT pancreas segmentation with geometry-aware U-Net. IEEE Trans. Med. Imaging 38 (8), 1971–1980.

Moglia, A., Cavicchioli, M., Mainardi, L., Cerveri, P., 2024. Deep Learning for Pancreas Segmentation: a Systematic Review. arXiv preprint arXiv:2407.16313.

Morineau, T., Morandi, X., Le Moëllic, N., Diabira, S., Riffaud, L., Haegelen, C., Hénaux, P.-L., Jannin, P., 2009. Decision making during preoperative surgical planning. Hum. Factors 51, 67–77.

Nguyen, A.H., Melstrom, L.G., 2020. Use of imaging as staging and surgical planning for pancreatic surgery. Hepatobiliary Surg. Nutr. 9 (5), 603.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Panda, A., Korfiatis, P., Suman, G., Garg, S.K., Polley, E.C., Singh, D.P., Chari, S.T., Goenka, A.H., 2021. Two-stage deep learning model for fully automated pancreas segmentation on computed tomography: Comparison with intra-reader and inter-reader reliability at full and reduced radiation dose on an external dataset. Med. Phys. 48 (5), 2468–2481.

Park, S., Chu, L., Fishman, E., Yuille, A., Vogelstein, B., Kinzler, K., Horton, K., Hruban, R., Zinreich, E., Fouladi, D.F., et al., 2020. Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation. Diagn. Interv. Imaging 101 (1), 35–44.

Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A., 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns 2 (11).

Plaza, S.M., Scheffer, L.K., Saunders, M., et al., 2012. Minimizing manual image segmentation turn-around time for neuronal reconstruction by embracing uncertainty. PLoS One 7 (9), 1–14.

Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A., 2020. Secure and robust machine learning for healthcare: A survey. IEEE Rev. Biomed. Eng. 14, 156–180.

Qu, T., Wang, X., Fang, C., Mao, L., Li, J., Li, P., Qu, J., Li, X., Xue, H., Yu, Y., et al., 2022. M3Net: A multi-scale multi-view framework for multi-phase pancreas segmentation based on cross-phase non-local attention. Med. Image Anal. 75, 102232.

Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K., 2023. Self-supervised learning: A succinct review. Arch. Comput. Methods Eng. 30 (4), 2761–2775.

Renard, F., Guedria, S., Palma, N.D., Vuillerme, N., 2020. Variability and reproducibility in deep learning for medical image segmentation. Sci. Rep. 10 (1), 13724.

Roh, Y., Heo, G., Whang, S.E., 2019. A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Trans. Knowl. Data Eng. 33 (4), 1328–1347.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E., Summers, R.M., 2015. DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation. arXiv:1506.06448.

Rudie, J.D., Duda, J., Duong, M.T., Chen, P.-H., Xie, L., Kurtz, R., Ware, J.B., Choi, J., Mattay, R.R., Botzolakis, E.J., et al., 2021. Brain mri deep learning and bayesian inference system augments radiology resident performance. J. Digit. Imaging 34 (4), 1049–1058.

Sander, J., de Vos, B.D., Išgum, I., 2020. Automatic segmentation with detection of local segmentation failures in cardiac MRI. Sci. Rep. 10 (1), 21769.

Santambrogio, R., Vertemati, M., Picardi, E., Zappa, M., et al., 2022. Planning the treatment: preoperative 3D reconstruction. Laparosc. Surg. 6, 1–8.

Senkyire, I.B., Liu, Z., 2021. Supervised and semi-supervised methods for abdominal organ segmentation: A review. Int. J. Autom. Comput. 18 (6), 887–914.

Shan, T., Yan, J., 2021. SCA-Net: A spatial and channel attention network for medical image segmentation. IEEE Access 9, 160926–160937.

Siegel, R.L., Giaquinto, A.N., Jemal, A., 2024. Cancer statistics, 2024. CA: Cancer J. Clin. 74 (1), 12–49. http://dx.doi.org/10.3322/caac.21820.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063.

Sureka, B., Jha, S., Yadav, A., Varshney, V., Soni, S., Vishnoi, J.R., Yadav, T., Garg, P.K., Khera, P.S., Misra, S., 2021. MDCT evaluation of pancreatic contour variations in head, neck, body and tail: surgical and radiological significance. Surg. Radiol. Anat. 43, 1405–1412.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15 (1), 1–28.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693.

Tian, L., Zou, L., Yang, X., 2023. A two-stage data-model driven pancreas segmentation strategy embedding directional information of the boundary intensity gradient and deep adaptive pointwise parameters. Phys. Med. Biol..

Tong, N., Xu, Y., Zhang, J., Gou, S., Li, M., 2023. Robust and efficient abdominal CT segmentation using shape constrained multi-scale attention network. Phys. Medica 110, 102595.

Valverde, J.M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R., Tohka, J., 2021. Transfer learning in magnetic resonance brain imaging: A systematic review. J. Imaging 7 (4), 66.

Vareedayah, A.A., Alkaade, S., Taylor, J.R., 2018. Pancreatic adenocarcinoma. Mo. Med. 115 (3), 230.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Villarini, B., Asaturyan, H., Kurugol, S., Afacan, O., Bell, J.D., Thomas, E.L., 2021. 3D Deep learning for anatomical structure segmentation in multiple imaging modalities. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems. CBMS, IEEE, pp. 166–171.

Yeghiazaryan, V., Voiculescu, I., 2018. Family of boundary overlap metrics for the evaluation of medical image segmentation. J. Med. Imaging 5 (1), 015006.

Zhang, Y., Yang, Y., Chen, S., Ji, J., Ge, H., Huang, H., 2023a. Clinical application of 3D reconstruction in pancreatic surgery: a narrative review. J. Pancreatol. 6 (01), 18–22.

Zhang, C., Zheng, H., Gu, Y., 2023b. Dive into the details of self-supervised learning for medical image analysis. Med. Image Anal. 89, 102879. http://dx.doi.org/10.1016/j.media.2023.102879, URL https://www.sciencedirect.com/science/article/pii/S1361841523001391.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging 13 (4), 716–724.