https://doi.org/10.1016/j.cmpb.2024.108494

A Comprehensive Benchmarking of A U-Net based Model for Midbrain Auto-segmentation on Transcranial Sonography

Hong-yu KANG[1, #], Wei ZHANG[2, #, *], Shuai LI[1], Xinyi WANG[1], Yu SUN[2], Xin SUN[2], Fang-xian LI[2], Chao HOU[2], Sai-kit LAM[1, 3], Yong-ping ZHENG[1, 3, *]

[1]Department of Biomedical Engineering, The Hong Kong Polytechnic University, HKSAR, China.

[2]Department of Ultrasound, Beijing Tiantan Hospital, Capital Medical University, Beijing, China.

[3]Research Institute of Smart Ageing, The Hong Kong Polytechnic University, HKSAR, China.

[#]The two authors contributed equally to this work.

***Correspondence:**

Yongping ZHENG, PhD

Department of Biomedical Engineering

The Hong Kong Polytechnic University

Hung Hom, Kowloon

Hong Kong, China

Email: yongping.zheng@polyu.edu.hk

Wei ZHANG, MD

Department of Ultrasound

Beijing Tiantan Hospital

Capital Medical University

NO.119, South 4th Ring West Road

Fengtai District

Beijing 100070, China

Email: ultrazhangwei@126.com

Highlights

- The nnU-Net model yielded the best midbrain segmentation agreement, achieving the top-ranked averaged scores compared U-Net and U-Net+++ on original TCS Images.
- The nnU-Net demonstrated the best-performing performance in aspects of segmentation agreement, model stability, and time efficiency.
- The segmentation power of the nnU-Net network remained robust against ultrasound imaging noise.
- This large cohort study is the first of its kind to benchmark the best-performing state-of-the-art deep neural network for TCS-based midbrain auto-segmentation in a wide spectrum of application perspectives.

Abstract

Background and Objective: Transcranial sonography-based grading of Parkinson's Disease has gained increasing attention in recent years, and it is currently used for assistive differential diagnosis in some specialized centers. To this end, accurate midbrain segmentation is considered an important initial step. However, current practice is manual, time-consuming, and bias-prone due to the subjective nature. Relevant studies in the literature are scarce and lacks comprehensive model evaluations from application perspectives. Herein, we aimed to benchmark the best-performing U-Net model for objective, stable and robust midbrain auto-segmentation using transcranial sonography images.

Methods: A total of 584 patients who were suspected of Parkinson's Disease were retrospectively enrolled from Beijing Tiantan Hospital. The dataset was divided into training (n=416), validation (n=104), and testing (n=64) sets. Three state-of-the-art deep-learning networks (U-Net, U-Net+++, and nnU-Net) were utilized to develop segmentation models, under 5-fold cross-validation and three randomization seeds for safeguarding model validity and stability. Model evaluation was conducted in testing set in three key aspects: (i) segmentation agreement using DICE coefficients (DICE), Intersection over Union (IoU), and Hausdorff Distance (HD); (ii) model stability using standard deviations of segmentation agreement metrics; (iii) prediction time efficiency, and (iv) model robustness against various degrees of ultrasound imaging noise produced by the salt-and-pepper noise and Gaussian noise.

Results: The nnU-Net achieved the best segmentation agreement (averaged DICE: 0.910, IoU: 0.836, HD: 2.793-mm) and time efficiency (1.456-s). Under mild noise corruption, the nnU-Net outperformed others with averaged scores of DICE (0.904), IoU (0.827), HD (2.941 mm) in the salt-and-pepper noise (signal-to-noise ratio, SNR = 0.95), and DICE (0.906), IoU (0.830), HD (2.967 mm) in the Gaussian noise (sigma value, $\sigma = 0.1$); by contrast, intriguingly, performance of the U-Net and U-Net+++ models were remarkably degraded. Under increasing levels of simulated noise corruption (SNR decreased from 0.95 to 0.75; σ increased from 0.1 to 0.5), the nnU-Net network exhibited marginal decline in segmentation agreement meanwhile yielding decent performance as if there were absence of noise corruption.

Conclusions: The nnU-Net model was the best-performing midbrain segmentation model in terms of segmentation agreement, stability, time efficiency and robustness, providing the community with an objective, effective and automated alternative. Moving forward, a multi-center multi-vendor study is warranted when it comes to clinical implementation.

Keywords: Parkinson's Disease, Midbrain, Transcranial Sonography, Auto-segmentation, Deep Learning.

1 Introduction

Parkinson's Disease (PD) is a chronic, progressive and devastating neurodegenerative disorder, accounting for over 9.4 million of sufferers across the globe in 2020 [1] and it has been considered as the fastest-growing neurological illness [2, 3]. Early diagnosis and frequent surveillance are the key antidote to this crippling disease. Among various imaging approaches, transcranial sonography (TCS) possesses unique particularities of being non-invasive, radiation-free, rapid, low cost, highly accessible, high patient compliance, easy-to-operate, as well as proven to offer the correlations of its hyper echoic pattern in brain substantia nigra with incidence and severity of PD [4-8]. In light of this, several international guidelines on TCS-based PD assessment and grading have been well-documented to aid neurologists in offering enhanced healthcare delivery to PD sufferers in aspects of early diagnosis and treatment outcome assessment [9-11]. Its popularity is on a rapid rise around the world [12, 13].

In clinical practice, an accurate segmentation of the midbrain structure is considered a crucial initial task. [14, 15] in TCS-based PD grading and severity assessment [16, 17]. However, traditional practice of midbrain segmentation is entirely a manual process, which suffers from two major drawbacks. First, the time-consuming nature of the practice has not only posed significant clinical burdens, especially in busy clinics, but also called for experienced clinicians for accurate midbrain segmentation which may not be easily overcome in resource-demanding regions, such as developing and under-developed countries. Second, the issue of subjectiveness is intrinsically embedded in the nature of the manual segmentation process, which may in turn incur bias in subsequent PD grading and assessment. In the contemporary era of AI in medicine, there is a pressing demand for an automated, efficient, and objective technique for midbrain segmentation in order to lay a solid foundation toward revolutionizing the clinical practice of TCS-based PD assessment in the long run.

With this regard, the U-Net deep-neural network and its variants, including the U-Net+++ and the nnU-Net, have been gaining increasing popularity in the community for medical image segmentation. In 2015, the U-Net neural network was first introduced; its design presents a U-shaped architecture and it utilizes an encoder-decoder structure with skip connections to capture contextual information for achieving precise localization [18]. In 2020, an extension of the U-Net network called U-Net +++ was introduced. It is coupled with nested pathways and deeply supervised full-scale skip connections, enabling multi-scale contextual information [19]. In 2021, the nnU-Net model, an adaptive, unified extensible training framework, was introduced for segmenting medical images. Instead of altering the network architecture, the authors proposed a comprehensive training workflow structure. [20]. The segmentation capabilities of these three neural networks on various imaging modalities have been well-documented, including computed tomography [21-23], magnetic resonance imaging (MRI) [24-26] and ultrasound [27-30]. In spite of their promising segmentation power, the application of AI in midbrain segmentation is scare in the literature.

TCS-based midbrain segmentation can be broadly classified into two categories according to the current body of literature: traditional and AI-based approaches. For traditional method, Sakalauskas et al. revealed a modified shape-based active contour segmentation algorithm for midbrain using a dataset of 40 TCS images in 2013, which achieved a DICE of 73.1±7.5% [17]. Later in 2016, Sakalauskas et al. integrated an experience-based statistical shape model with an intensity-amplitude invariant edge detector, for extracting the fuzzy boundaries of the midbrain on TCS images using a dataset of 130 TCS images, which achieved a mean DICE and Hausdorff Distance (HD) between 87.6-89.6% and 3.60-6.09 mm, respectively [31]. For AI-based strategies, the only relevant research work found in the literature was conducted by Milletari et al. (2017), they enrolled 34 subjects and developed a Hough-CNN network for midbrain segmentation, which achieved a DICE between 0.77-0.85,

depending on different configurations of the convolutional layers [32]. These abovementioned studies have reflected the importance and laid a foundation of TCS-based automated midbrain segmentation over the past decade.

Nevertheless, the contemporary era of AI has placed a growing emphasis on the sample size and a boarder scope of model evaluations from application perspectives, including and beyond the segmentation accuracy. The only relevant research in the literature by Milletari et al. may not be capable of providing sufficient values of their developed models from clinical implementation perspectives, on top of the inadequacy of sample size [32]. To address these challenges, this study aimed to conduct a large-cohort study and benchmark a U-Net based network from U-Net, U-Net+++, and nnU-Net for midbrain auto-segmentation on TCS images. Efficacy of the developed models were comprehensively evaluated in multiple application facets, including segmentation agreements, model stability, time efficiency, as well as model robustness against noise-corrupted TCS images. Findings of this study are anticipated to provide the community with enhanced understanding regarding the clinical usefulness of the developed auto-segmentation models, and to pave the way for further investigations on substantia nigra auto-segmentation towards objective, effective and fully automated PD assessment using TCS images in the long run.

2 Methods

2.1 Data Acquisition, Annotation and Image Pre-processing

TCS images of 584 patients, who were suspected or diagnosed with PD, were retrospectively enrolled from Beijing Tiantan Hospital between 2021 and 2023. Patient consent was waived because of the retrospective nature of this study. Ethical approval for this study was obtained from the Human Subject Ethics Sub-committee (HSESC) of the Hong Kong Polytechnic University (HSEARS20231102004). All the original TCS images were acquired by a Canon Aplio i900 i-series ultrasound system (Canon Medical Systems Corporation, Otawara, Tochigi, Japan) with i8CX1 convex array (center frequency = 2.6 MHz). An experienced physician maneuvered the ultrasound probe over the temporal region, followed by capturing the ultrasound imaging frames that display the butterfly-shaped midbrain structure.

The ground-truth midbrain annotation was manually delineated on the captured imaging frames using the Canon Aplio i900 i-series ultrasound system by experienced physicians with over 10 years of experience.

In this study, several pre-processing procedures were performed on the original TCS images prior to downstream analyses. First, all the TCS images together with the corresponding ground-truth midbrain annotations were cropped from a resolution of 1280x960 to 400x320, in order to effectively localize the midbrain region within the region-of-interest (ROI) meanwhile minimizing irrelevant information from other structures for achieving effective learning during subsequent model development. Second, a binary mask was generated by extracting the region within the ground-truth midbrain annotations for each patient; this procedure enabled an accurate representation of the segmentation target for downstream model development and evaluation.

2.2 Deep-learning Neural Networks

In the present study, three U-Net based neural networks, including U-Net, U-Net+++, and nnU-Net, were deployed for developing auto-segmentation models. Architectures of these three networks are illustrated in **Supplementary Figure 1A-1D**. Details on settings of each of these networks were presented as follows:

2.2.1 U-Net

The U-Net model is a state-of-the-art architecture designed dedicatedly for medical image segmentation [18]. It is one of the most widely adopted segmentation network within the medical imaging community. The U-Net network exhibits a U-shaped architecture, it strategically combines encoders and decoders, enabling information fusion throughout the network, for achieving precise segmentation outcomes. In this present experiment, the U-Net model contains four layers with 64, 128, 256 and 512 feature channels, respectively. In the encoder path, the input undergoes four layers convolution blocks for feature extraction, followed by four max pooling operations; each layer contains two 3x3 convolutions and same padding, followed by a 2x2 max pooling with strides of two in each dimension. In the decoder path, the extracted features are subsequently processed through four layers of convolution blocks and undergo four rounds of deconvolutions; each layer includes a 2x2 up-convolution with strides of two in each dimension, followed by two 3x3 convolutions and the same padding. For model development, a maximum of 100 epochs was specified, starting with an initial learning rate of 1e-4. The momentum was set to 0.9 to aid in faster convergence and stability, while the batch size of 4 was chosen to determine the number of samples processed before updating the model parameters.

2.2.2 U-Net+++

The U-Net+++ model is a full-scale skip connection and deep supervision based on the U-Net architecture [19]. The Unet+++ introduces a novel full-size skip connection, which strengthens the interconnections between the encoder and decoder, as well as the intra-connections among decoder subnetworks, allowing each decoder layer to effectively incorporate both small-scale and same-scale feature maps from the encoder. UNet+++ is designed to aggregate features across all scales, incorporating skip connections between the top and bottom layers of both the encoder and decoder. In this study, the U-Net+++ network contains four layers with 64, 128, 256 and 512 feature channels, respectively. In the encoder path, each layer comprises two 3x3 convolutions and the same padding, followed by a 2x2 max pooling operation with strides of two in each dimension. In the decoder path, each layer of the U-Net+++ model comprises a 2x2 up-convolution with strides of two in each dimension, followed by two 3x3 convolutions and the same padding. The model includes skip connections at the same feature map level within the encoder, skip connections from higher to lower feature map levels, and decoder skip connections. The parameters of the U-Net+++ model remained the same as those of the U-Net network, including a maximum of 100 epochs, an initial learning rate of 0.0001, a momentum value of 0.9, and a batch size of 4.

2.2.3 nnU-Net

The nnU-Net model, an adaptive and unified extensible training framework, was introduced for the segmentation of medical images. [20]. It is an automatic deep learning-based segmentation method that configures itself for new tasks, encompassing preprocessing, network architecture, training, and post-processing. It is built on an adaptive framework based on 2D and 3D U-Net, and automatically tunes hyperparameters based on the specific properties of the given dataset, such as the exact patch size, batch size, and inference settings based on the size and characteristics of the input images. In this research work, the 2D version of nnU-Net was utilized. The model consists of an encoder path and a decoder path. Within the nnU-Net architecture used in this study, there are seven layers, each with 32,

64, 128, 256, 512, 512 and 512 feature channels, respectively. In the encoder path, each layer employs two 3x3 convolutions to extract hierarchical features from the input data. This process allows for progressive capture of intricate details and patterns presented in the midbrain of the TCS images. Conversely, in the decoder path, each layer consists of a 2x2 up-convolution with strides of two in each dimension, followed by two 3x3 convolutions. The nnU-Net model was trained with a maximum of 100 epochs. Several other parameters were set to their initial values in the experiment. The learning rate was set to 0.0001. A momentum value of 0.9 was used to enhance the stability of the training process. The training was performed using a batch size of 4.

2.3 Model Development and Evaluation

To comprehensively compare the performance of the three models, the model Development involves dataset partitioning, image pre-processing, and training with 5-fold cross-validation using 3 random seeds. The evaluation of the models includes two essential aspects: assessing model segmentation performance and evaluating model robustness under various noise conditions.

2.3.1 Model Development

Figure 1 showcases a schematic diagram of the entire experimental workflow. First, the entire dataset (n=584) into was randomly into training, validation, and testing sets via a randomization seed, at an approximate ratio of 70% (n=416), 20% (n=104), 10% (n=64), respectively, taking reference from previous literature [33]. All the analyzed images at this stage were processed according to the image pre-processing procedures depicted in **Section 2.1.3**.

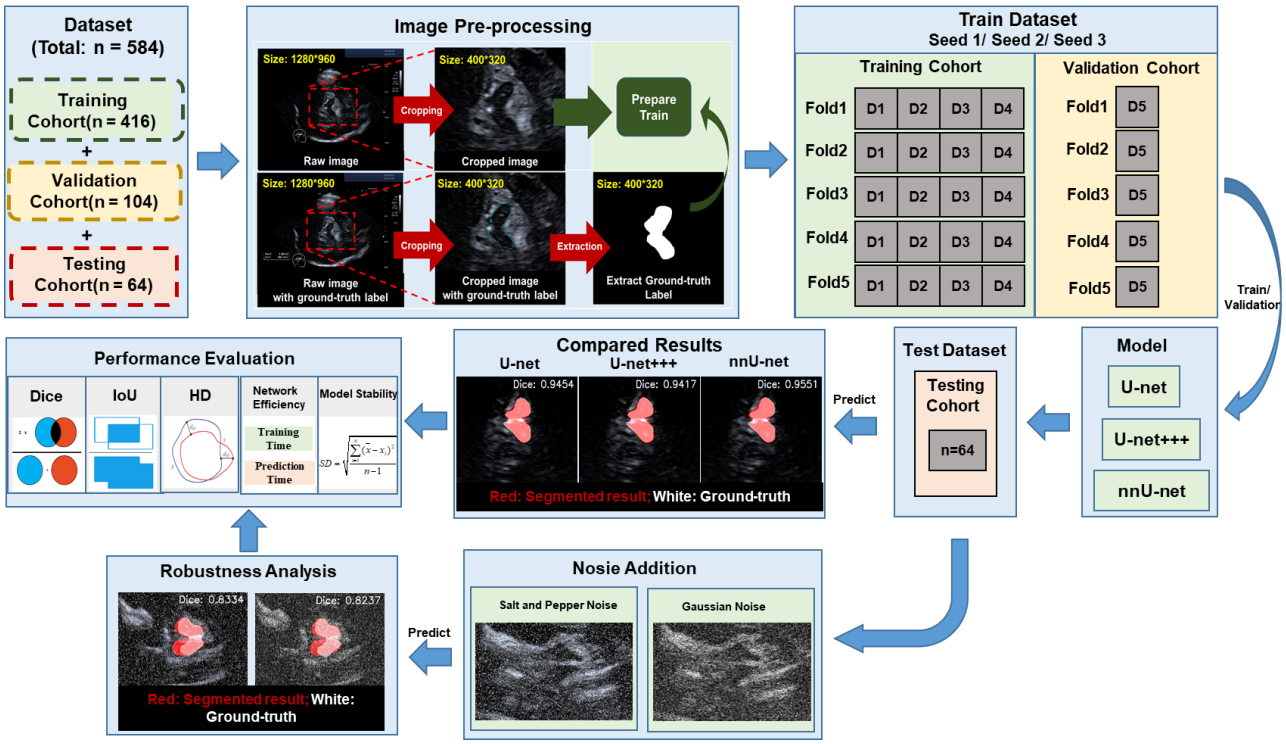

Fig. 1. Overall study workflow from dataset randomization, image pre-processing, to model training, robustness analysis after noise addition on the original images, and model performance evaluation

In the training set, three deep-learning networks (U-Net, U-Net+++, and nnU-Net) were separately utilized to develop three segmentation models under 5-fold cross-validation. In this experiment, three independent randomization seeds were deployed to obtain three sets of results for

the sake of harvesting a fair model performance across different patient sub-populations. The approach of multiple randomized stratification was commonly adopted [16, 34, 35]. An averaged performance of each model was reported in this study to assess model stability and to improve validity of the findings. The training process was carried out on an in-house service machine equipped with an Nvidia RTX A6000 GPU card. Training time for each model was recorded.

2.3.2 Model Evaluation

The developed models were evaluated in the testing set in two scenarios: (a) Model segmentation performance on original images; and (b) Model robustness assessment on images corrupted by two types of typical ultrasound imaging noise (after implementing noise addition on the original images). Details of the model evaluation procedures are described below:

(a) Model Segmentation Performance on Original TCS Images

All the developed models were evaluated in the testing set in aspects of model segmentation agreement, model stability and prediction time efficiency. For model segmentation agreement, DICE coefficient (DICE), Intersection over Union (IoU), and Hausdorff Distance (HD), were calculated for evaluation. These evaluating metrics have been widely adopted in the field of medical image segmentation [36, 37]. Relevant equations and definitions are shown below:

DICE is a measure of similarity between two segmentations. It provides a value between 0 and 1, where a higher value indicates better, X and Y represent the area of predict image and ground truth respectively:

$$DICE = \frac{2|X \bigcap Y|}{|X| + |Y|} \quad \textbf{(Eq. 1)}$$

IoU is a metric that measures the overlap between two regions. It is calculated by dividing the intersection area of the predicted and ground truth by their union area. The IoU value ranges from 0 to 1, where a higher value indicates better, X and Y represent the area of predict image and ground truth respectively:

$$IoU = \frac{|X \bigcap Y|}{|X \bigcup Y|} \quad \textbf{(Eq. 2)}$$

HD measures the maximum dissimilarity between two regions. A smaller HD indicates better agreement between segmentations, quantifying the maximum discrepancy between predicted and ground truth boundaries, A and B are point sets, a and b are points in A and B respectively, and *d* (*a*, *b*) represents the distance between point a and point b:

$$HD(A,B) = max(\underset{a \in A}{max}(\underset{b \in B}{min}(d(a,b))), \underset{b \in B}{max}(\underset{a \in A}{min}(d(b,a)))) \quad \textbf{(Eq. 3)}$$

For model stability, standard deviations (SD) of each of the above segmentation evaluating metrics (DICE, IoU, HD) under 5-fold cross-validation across the three randomization seeds were adopted for evaluation.

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n-1}} \quad \textbf{(Eq. 4)}$$

For prediction time efficiency, the time (in second) required for the developed models to execute prediction on the testing set was calculated for assessment.

(b) Model Robustness Assessment on Noise-corrupted TCS Images

In addition to segmentation agreement, model stability and time efficiency, the developed models were evaluated in aspects of their robustness under a varying degree of noise-corrupted ultrasound images. Salt-and-pepper noise and Gaussian noise are two common types of ultrasound imaging noise [38], and they have been adopted in AI studies of ultrasound-based auto-segmentation [39]. Salt-and-pepper noise is a random impulse noise where certain pixels have different intensities compared to their neighbors. Gaussian noise, on the other hand, is a statistical noise caused by random signal fluctuations and follows a Gaussian distribution. In this study, the quality of the images in the testing set of 64 testing images was corrupted by addition of the salt-and-pepper noise and the Gaussian noise.

For the salt-and-pepper noise, three noise levels reflected by the signal-to-noise ratio (SNR) of 0.95, 0.85, and 0.75 were introduced and analyzed. For the Gaussian noise, three noise levels in terms of the sigma values (σ) of 0.1, 0.3, and 0.5, representing the standard deviation of a Gaussian distribution, were employed and analyzed.

To analyze model robustness, segmentation agreements of the three comparing networks (U-Net, U-Net+++, and nnU-Net) were first assessed in terms of DICE, IoU, and HD, under a mild level of noise corruption on the salt-and-pepper noise corrupted images (SNR=0.95) and the Gaussian noise corrupted images (σ=0.1). Following the above analysis, the best-performing network was benchmarked. Finally, robustness of the benchmarked model was subsequently further analyzed on images with increasing degrees of noise corruption on the salt-and-pepper noise corrupted images (SNR=0.95, 0.85, 0.75) and the Gaussian noise corrupted images (σ=0.1, 0.3, 0.5).

3 Results

After completing the comprehensive experiments mentioned above, this study primarily presents results in two parts: the model segmentation performance, and the robustness of the model under noise-corrupted conditions.

3.1 Model Segmentation Performance on Original TCS Images

The section mainly discusses the model segmentation agreement, model stability, and prediction time efficiency. Furthermore, extensive quantitative comparisons were conducted.

Model Segmentation Agreement and Model Stability

From quantitative perspectives, **Table 1A** summarizes the quantitative comparisons among the U-Net, U-Net+++, and nnU-Net networks for midbrain segmentation agreement (in terms of averaged DICE, IoU, and HD) and model stability (in terms of SD) on the testing dataset (relevant results for the

validation set can be found in **Supplementary Table 1A**). It showcases the detailed results of the 3 randomization seeds under 5-fold cross validation. It is worth noting that the nnU-Net model yielded the best segmentation agreement, achieving the top-ranked averaged scores of DICE (0.910, 0.907, 0.910 in randomization seed 1 to 3, respectively), IoU (0.838, 0.832, 0.837 in randomization seed 1 to 3, respectively), HD (2.727 mm, 2.801 mm, 2.852 mm in randomization seed 1 to 3, respectively). With respect to model stability, the nnU-Net model outperformed the other two comparing networks, achieving the top-ranked values of SD in DICE (0.001 in all the three randomization seeds), IoU (0.002, 0.001, 0.002 in randomization seed 1 to 3, respectively), and HD (0.095 mm, 0.065 mm, 0206 mm in randomization seed 1 to 3, respectively).

Table 1A. Results of segmentation agreement and stability of the three comparing neural networks in each of the three randomization seeds under 5 cross-validation on the testing dataset. Averaged (AVG) scores of the segmentation agreement metrics in DICE, IoU, and HD are presented for each model in each randomization seed. Standard deviation (SD) was calculated to illuminate the underlying model stability. The top-ranked scores are bolded.

		Randomization Seed 1			Randomization Seed 2			Randomization Seed 3		
Network	**Fold**	**DICE**	**IoU**	**HD (mm)**	**DICE**	**IoU**	**HD (mm)**	**DICE**	**IoU**	**HD (mm)**
U-Net	Fold1	0.905	0.829	3.375	0.899	0.820	3.394	0.903	0.825	3.359
	Fold2	0.901	0.823	3.321	0.901	0.823	3.207	0.904	0.826	3.194
	Fold3	0.904	0.828	3.223	0.892	0.808	3.504	0.899	0.818	3.541
	Fold4	0.900	0.822	3.337	0.899	0.820	3.271	0.903	0.824	3.556
	Fold5	0.903	0.826	3.704	0.894	0.811	3.566	0.898	0.817	3.830
	AVG	0.903	0.826	3.392	0.897	0.816	3.389	0.901	0.822	3.496
	SD	0.002	0.003	0.184	0.004	0.006	0.152	0.003	0.004	0.239
U-Net+++	Fold1	0.889	0.806	4.385	0.898	0.817	3.680	0.898	0.817	3.452
	Fold2	0.888	0.806	3.889	0.898	0.817	3.899	0.893	0.811	3.417
	Fold3	0.894	0.813	3.520	0.892	0.808	4.571	0.894	0.811	3.887
	Fold4	0.893	0.811	3.776	0.893	0.810	3.987	0.894	0.810	3.637
	Fold5	0.886	0.802	3.804	0.895	0.813	4.034	0.893	0.810	3.734
	AVG	0.890	0.807	3.875	0.895	0.813	4.034	0.894	0.812	3.626
	SD	0.003	0.004	0.317	0.003	0.004	0.329	0.002	0.003	0.197
nnU-Net	Fold1	0.909	0.836	2.882	0.908	0.833	2.791	0.910	0.837	2.723
	Fold2	0.912	0.841	2.646	0.907	0.831	2.743	0.910	0.836	2.723
	Fold3	0.911	0.838	2.654	0.908	0.833	2.832	0.913	0.841	2.669
	Fold4	0.909	0.836	2.731	0.908	0.832	2.896	0.910	0.837	3.026
	Fold5	0.911	0.839	2.721	0.906	0.830	2.743	0.909	0.835	3.120
	AVG	**0.910**	**0.838**	**2.727**	**0.907**	**0.832**	**2.801**	**0.910**	**0.837**	**2.852**
	SD	**0.001**	**0.002**	**0.095**	**0.001**	**0.001**	**0.065**	**0.001**	**0.002**	**0.206**

Table 1B. Results of segmentation agreement (in terms of DICE, IoU, and HD) and stability (in terms of SD) of the three deep-learning networks across all the three randomization seeds under 5-fold cross validation on the testing dataset. The top-ranked scores are bolded.

Network	DICE (SD)	IoU (SD)	HD (mm) (SD)
U-Net	0.900 (0.004)	0.821 (0.006)	3.425 (0.187)
U-Net+++	0.893 (0.003)	0.811 (0.004)	3.845 (0.318)
nnU-Net	**0.910 (0.002)**	**0.836 (0.003)**	**2.793 (0.137)**

Table 1B presents the averaged results of the three comparing networks across all the randomization seeds under 5-fold cross-validation on the testing dataset (relevant results for the validation set can be found in **Supplementary Table 1B**). The nnU-Net network outperformed U-Net and U-Net+++ networks, achieving the best segmentation agreement in terms of DICE (0.910), IoU (0.836), and HD (2.793 mm), and model stability in terms of SD in DICE (0.002), IoU (0.003), and HD (0.137 mm).

From qualitative aspects, **Figure 2** shows the segmentation agreement in terms of the DICE score in 4 representative patients, with the red regions representing the predicted segments and the white regions indicating the ground-truth annotation. Notably, the nnU-Net network consistently outperformed the other two networks in these 4 patients. By contrast, the U-Net and U-Net+++ networks were only capable of performing comparable segmentation agreements compared to the nnU-Net model in Patient 1 and 2; while their capabilities were considerably under-performed in Patient 3 and 4.

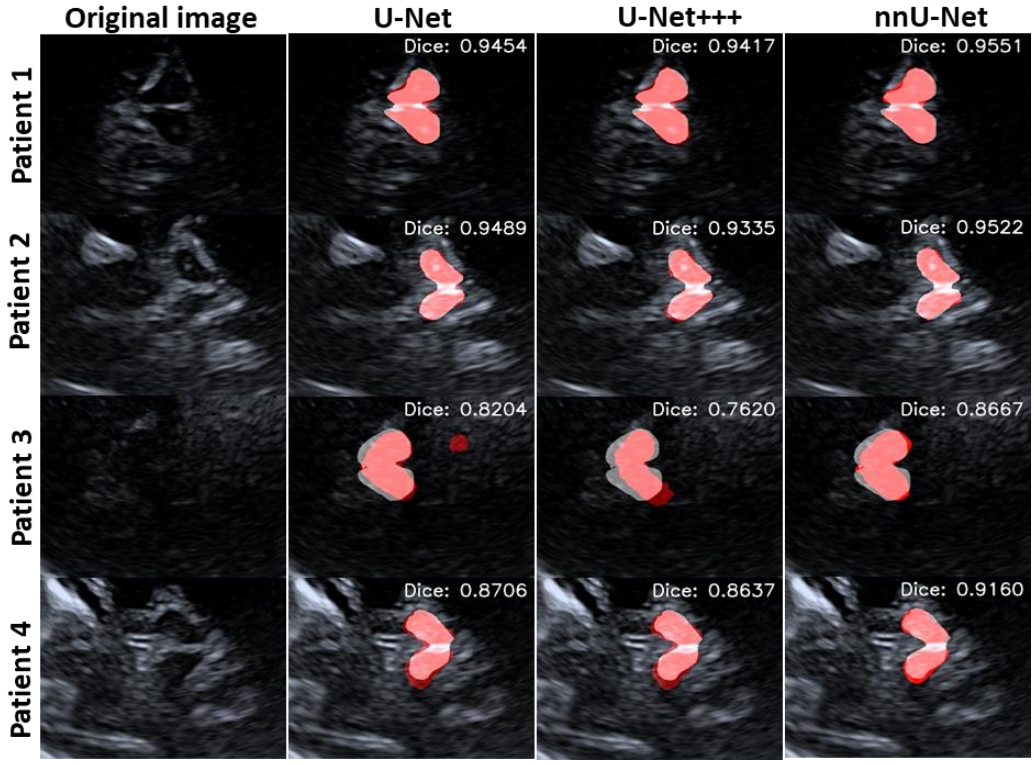

Fig 2. Qualitative visualization of the segmentation agreement in terms of the averaged DICE score in four representative patients, with the red region representing the predicted segments and the white region indicating the ground-truth annotation.

Prediction Time Efficiency

Table 2 presents the time efficiency of the three studied networks, in terms of training and prediction time in the training and testing sets, respectively. It is worth noting that the nnU-Net network was determined to be the most efficient model, requiring the least amount of time (1.456 second) in generating the predictions in the testing set, despite that it took the longest duration of time during model training (114 minute).

Table 2. Results of the training time and prediction time of the three studied networks. The shortest time are bolded.

Network	Training Time	Prediction Time
U-Net	**34 minutes**	2.427 second
U-Net+++	36 minutes	3.083 second
nnU-Net	114 minutes	**1.456 second**

Based on the above evaluations, the nnU-Net demonstrated the best-performing performance among the three neural networks in aspects of segmentation agreement, model stability, and time efficiency. Details of the segmentation agreement of the three comparing networks for each testing

case under randomization seed-1 are presented in DICE and visualized in **Supplementary Figure 2A-2C.**

3.2 Model Robustness Assessment on Noise-corrupted TCS Images

The section primarily discusses the impact of noise addition on the performance of the benchmarked model and the impact of incremental noise addition on the performance of the nnU-Net Model.

Impact of Noise Addition on Performance of the Benchmarked Model

Table 3 summarizes the segmentation agreement of the three studied networks in both original and noise-corrupted ultrasound images under mild degree of noise (i.e., noise level of SNR=0.95 for the salt-and-pepper noise, and σ=0.1 for the Gaussian noise). The nnU-Net was the best-performing model, demonstrating a dramatically greater capacity of midbrain segmentation in both noise-corrupted images compared to the other two comparing networks, and the underlying segmentation agreements were approximate to that on the original images. By contrast, the U-Net+++ network was the most underperforming model in the salt-and-pepper noise corrupted images (SNR=0.95: DICE=0.032, IoU=0.017, HD=37.827-mm) and in the Gaussian noise corrupted images (σ=0.1: DICE=0.157, IoU=0.089, HD=41.802-mm); and the underlying segmentation agreements on the noise-corrupted images were far lower than those of the nnU-Net model, incurring up to 28 times degradation in DICE, 48 folds in IoU and 14 times in HD.

Of note, it is intriguing to point out that both the U-Net and U-Net+++ segmentation models were highly sensitive to the two studied types of ultrasound imaging noise, even at a low intensity level of noise corruption, resulting in tremendous drops in the segmentation agreements. On the contrary, the segmentation power of the nnUNet network remained robust in the same noise addition setting.

Figure 3 and Figure 4 visualize the segmentation capacity of the three neural networks in three representative patients when the images were corrupted with the salt-and-pepper noise (SNR=0.95) and the Gaussian noise (σ=0.1), respectively. The results were found consistent with those illustrated in **Table 3**.

Table 3. Segmentation agreement of the three studied networks (in terms of averaged DICE, IoU and HD) in both original and noise-corrupted ultrasound images (noise level of SNR=0.95 for the salt-and-pepper noise, and σ=0.1 for the Gaussian noise), across the three randomization seeds under 5-fold cross validation on the testing set.

Network	Original image			Salt-and-pepper noise corrupted image (SNR = 0.95)			Gaussian noise corrupted image (σ = 0.1)		
	DICE	IoU	HD (mm)	DICE	IoU	HD (mm)	DICE	IoU	HD (mm)
nnU-Net	0.909	0.836	2.793	**0.904**	**0.827**	**2.941**	**0.906**	**0.830**	**2.967**
U-Net	0.900	0.821	3.425	0.389	0.268	41.946	0.535	0.389	38.241
U-Net+++	0.893	0.811	3.845	0.032	0.017	37.837	0.157	0.089	41.802

The nnU-Net network illuminated the greatest segmentation agreement and robustness both qualitatively in terms of the matching between the red region (i.e., the predicted segment) and the white region (i.e., the ground-truth annotation), and quantitatively in aspects of averaged DICE ranging from 0.900 to 0.955 **(Figure 3 and Figure 4)**. Consistent with the findings in **Table 3**, the UNet+++ exhibited the worst segmentation agreement and robustness on the noise-corrupted images in both

qualitative and quantitative perspectives **(Figure 3 and Figure 4)**, with averaged DICE scores ranging from 0.010 to 0.338. It may also be worth noting that the predicted segments in the U-Net and U-Net+++ models were of small size and were sparsely distributed inside and outside the midbrain region on the noise-corrupted images.

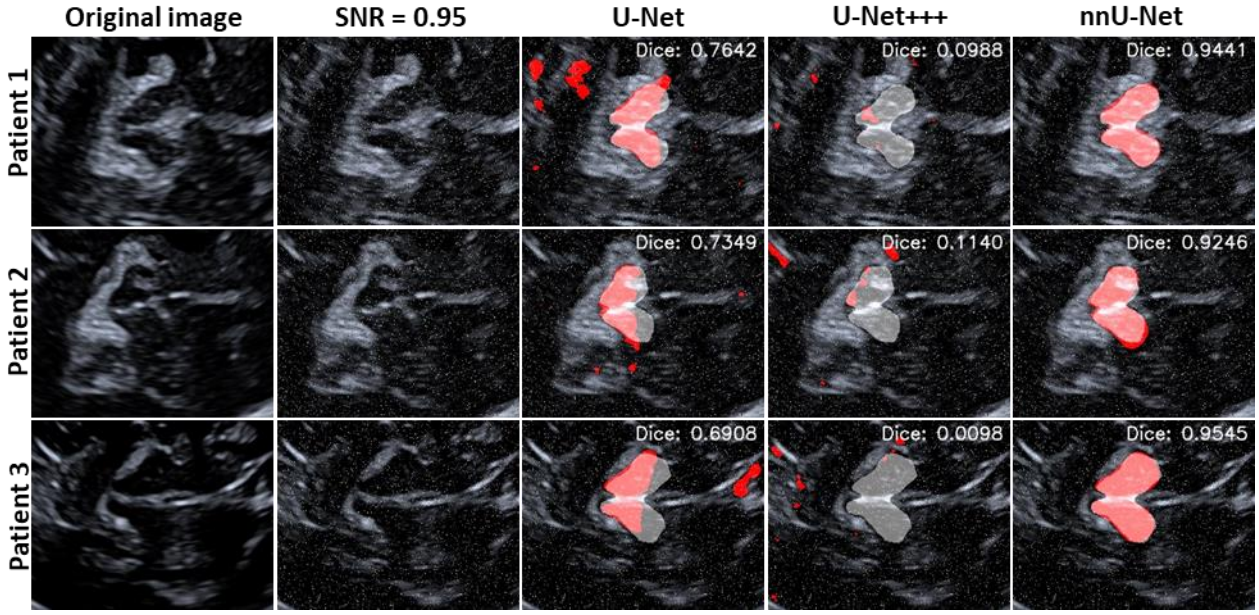

Fig. 3. Qualitative visualization of the segmentation agreement in terms of the averaged DICE score in three representative patients, on the salt-and-pepper noise corrupted images (SNR=0.95). The red region representing the predicted segments and the white region indicating the ground-truth annotation.

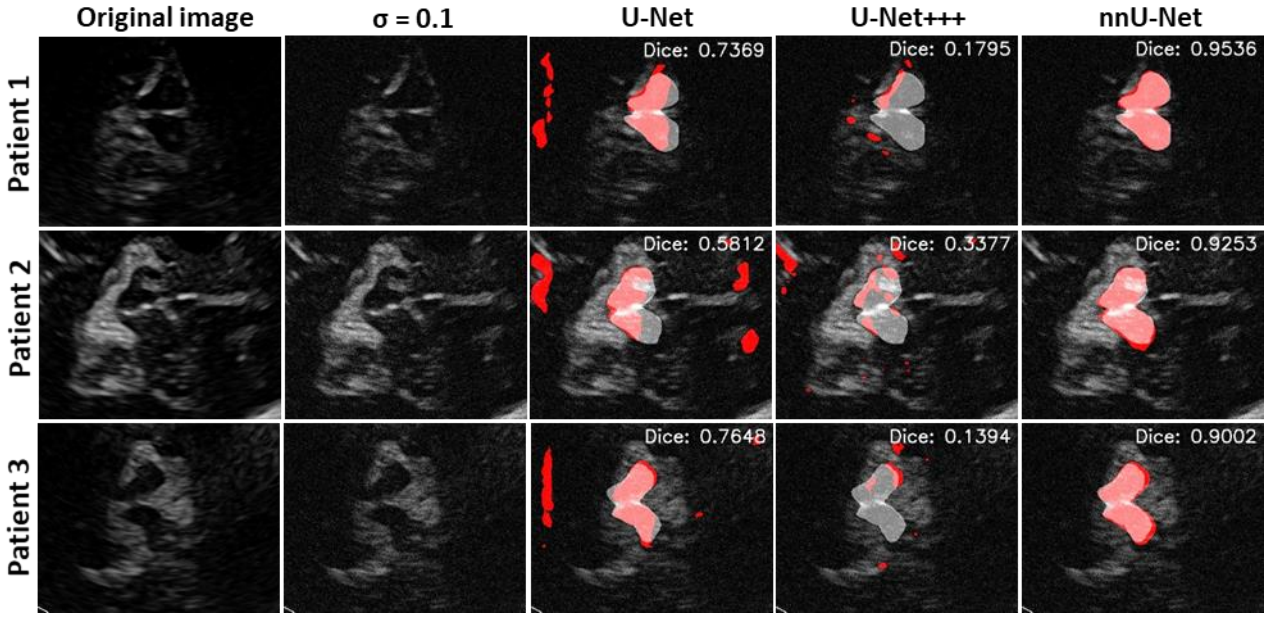

Fig. 4. Qualitative visualization of the segmentation agreement in terms of the averaged DICE score in three representative patients, on the Gaussian noise corrupted images (σ=0.1). The red region representing the predicted segments and the white region indicating the ground-truth annotation.

Impact of Incremental Noise Addition on Performance of nnU-Net Model

Table 4 summarizes the segmentation power of the benchmarked nnU-Net model on the original images and corrupted-images of varying degrees of noise intensity (SNR ranged from 0.75 to 0.95 for the salt-and-pepper noise; σ values ranged from 0.10 to 0.50 for the Gaussian noise); the results were averaged across the three randomization seeds under 5-fold cross validation on the testing set.

Table 4 summarizes the segmentation power of the benchmarked nnU-Net model on the original images and corrupted-images of varying degree of noise intensity; the results were averaged across the three randomization seeds under 5-fold cross validation on the testing set.

Noise level	Segmentation Agreement		
	DICE	IoU	HD (mm)
Original Image	0.909	0.836	2.793
SNR = 0.95	0.904	0.827	2.941
SNR = 0.85	0.888	0.805	3.552
SNR = 0.75	0.882	0.796	4.040
σ = 0.10	0.906	0.830	2.967
σ = 0.30	0.891	0.809	3.501
σ = 0.50	0.874	0.784	4.183

Figure 5 and **Figure 6** demonstrate a representative case for visually illuminating the impact of varying intensity levels of the salt-and-pepper noise and Gaussian noise, respectively, on both the image appearance and the segmentation agreement of the nnU-Net network; additionally, histograms are presented for each of the studied noise intensity levels.

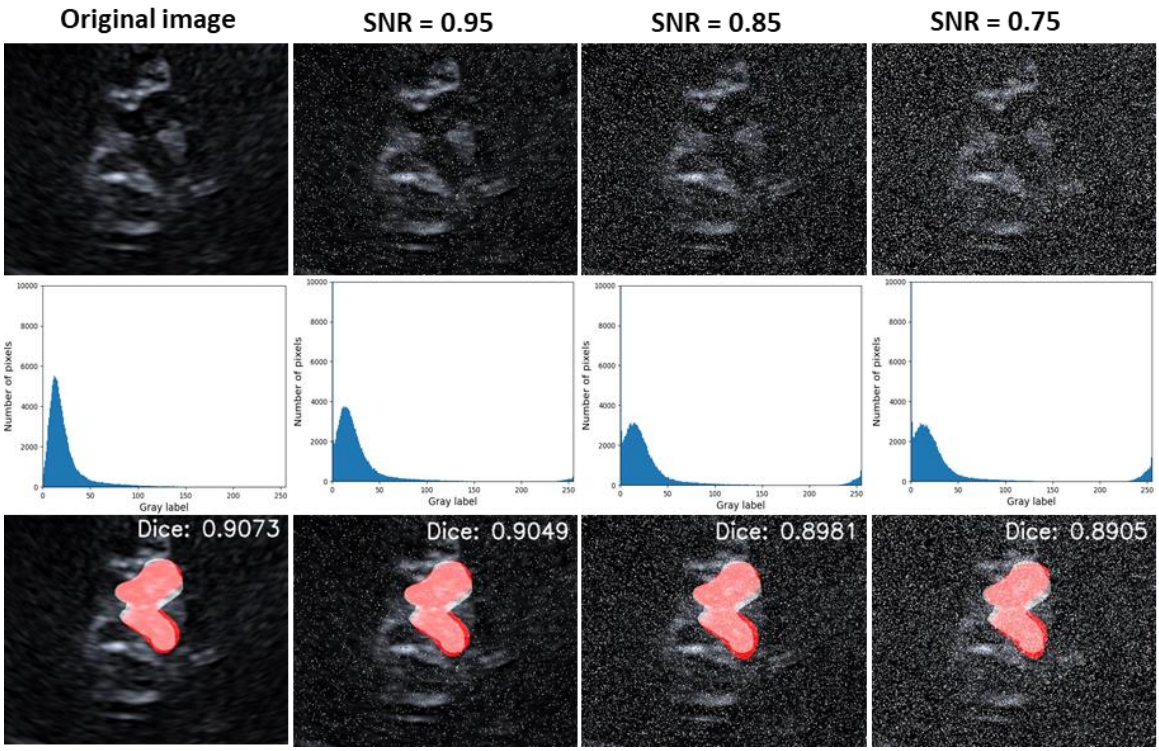

Fig. 5. demonstrate a representative case for visually illuminating the impact of varying intensity levels of the salt-and-pepper noise (SNR= 0.75, 0.85, and 0.95), on both the image appearance (first row), the underlying histograms (middle row), and the segmentation agreement of the nnU-Net network (bottom row), where the red regions and white regions representing the predicted and ground-truth annotation, respectively.

Two important findings were noted. Firstly, additions of imaging noise and increasing degree of noise corruption to the original ultrasound images rendered a gradual drop of the segmentation power of the nnU-Net network in aspects of DICE, IoU and HD, irrespective of the types of imaging noise studied **(Table 4, Figure 5 and Figure 6)**. Secondly, in spite of the reduction in segmentation agreement under the increasing intensity of imaging noise, the nnU-Net network was still capable of producing satisfactory segmentation power, even at the highest level of noise intensity **(Table4, Figure 5 and Figure 6)**. For instance, in the setting SNR of 0.75 resulting from the addition of the salt-and-pepper noise, the averaged scores of DICE, IoU and HD were 0.882, 0.796, and 4.040-mm respectively, corresponding to approximately 2%, 4.5%, and 45% reductions compared to those on the original images **(Table 4)**.

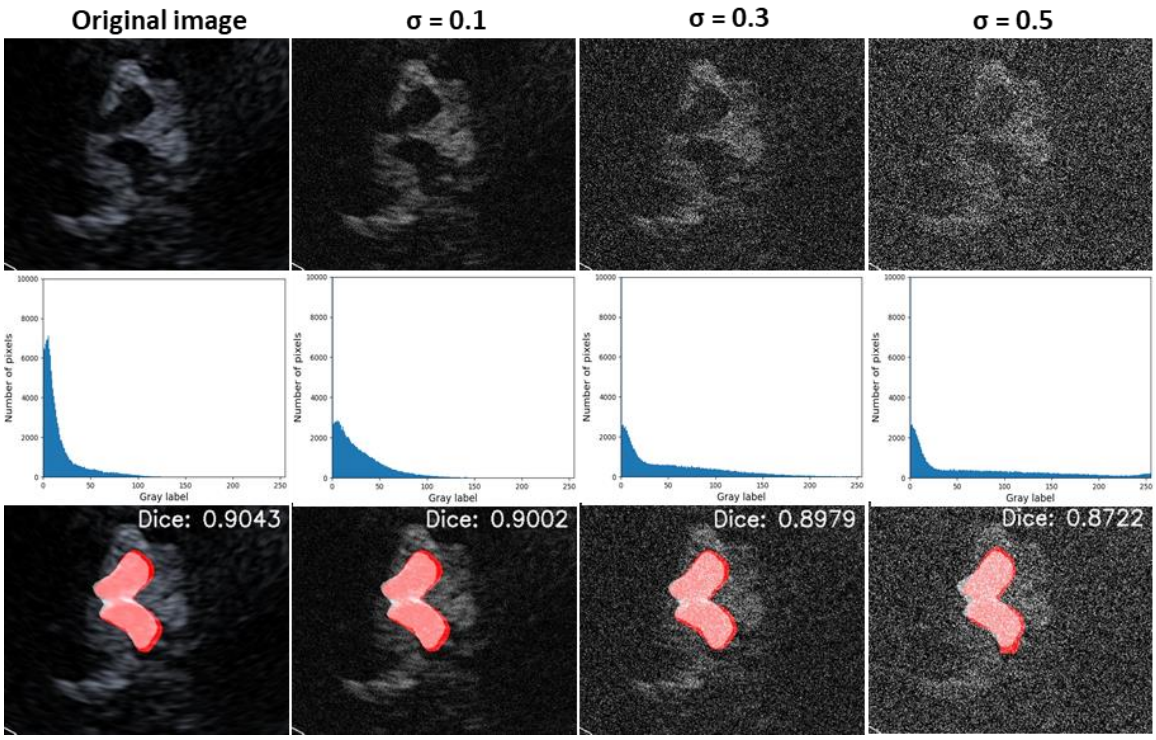

Fig. 6. demonstrate a representative case for visually illuminating the impact of varying intensity levels of the salt-and-pepper noise (σ = 0.1, 0.3, and 0.5), on both the image appearance (first row), the underlying histograms (middle row), and the segmentation agreement of the nnU-Net network (bottom row), where the red regions and white regions representing the predicted and ground-truth annotation, respectively.

4 Discussion

Midbrain segmentation is a key initial step in TCS-based PD grading for purposes of diagnosis, severity assessment, and treatment response surveillance. Yet, the current practice of midbrain segmentation is a manual, time-consuming and subjective procedure, putting a huge demand for automated, effective and objective alternatives in the contemporary paradigm of AI. Despite the superiority of TCS in offering a non-invasive, radiation-free, rapid, low cost, highly accessible, high patient compliance, easy-to-operate tool in a stark contrast to other imaging modalities [8], relevant studies on AI for midbrain segmentation are desperately wanting. The only relevant research in the literature by Milletari et al. was deficient in providing sufficient value of their developed models from clinical implementation perspectives, on top of the inadequacy of sample size of 34 subjects [32]. Given the challenges outlined above, our objective was to develop deep learning methods and benchmark best-performing state-of-the-art U-Net based neural networks using a substantial cohort of 584 subjects.

The evaluation focused on segmentation agreement, model stability, time efficiency, and model robustness. Results indicated that the nnU-Net model outperformed the other two networks in all these evaluating aspects, potentially providing the community with a favorable and automated alternative for midbrain segmentation towards TCS-based PD assessment in the future.

Aligned with the present study, all the abovementioned studies reflected the importance of TCS-based automated midbrain segmentation over the past decade, which are all valuable to the research community. However, there exists a growing demand on auto-segmentation works for an adoption of large sample size and comprehensive model evaluation beyond segmentation accuracy. In contrast to these previous works, the present study benchmarked the state-of-the-art nnU-Net network as the best-performing midbrain segmentation model in aspects of segmentation agreement (in DICE, IoU, HD, and AD) **(Table 1A-1B, Figure 2)**, model stability **(Table 1A-1B)**, time efficiency **(Table 2)**, as well as model robustness against noise-corrupted TCS images **(Table 3-4, Figure 5-6)**, when compared to the U-Net and U-Net+++ models; these findings were obtained using a larger cohort of 584 subjects. For segmentation agreement, the present nnU-Net model yielded a higher averaged scores of DICE (91.0%), IoU (83.6%), HD (2.793 mm), and AD (79%) as shown in **Table 1A-1B**, which were superior over the abovementioned previous works [32]. For model stability, the nnU-Net model also achieved the top-ranked values of SD in DICE, IoU, HD, and AD, compared to the U-Net and U-Net+++ models **(Table 1A-1B)**. For time efficiency, the nnU-Net network required the least amount of time (1.456 second) in generating the predictions in the testing set **(Table 2)**. For model robustness, the nnU-Net demonstrated a dramatically greater capacity of midbrain segmentation in images corrupted by the salt-and-pepper noise (SNR=0.95) and the Gaussian noise (σ=0.10), and the underlying segmentation agreements were approximate to that on the original images **(Table 3)**. Notably, such assessments of model robustness against noise corruption have not been previously revealed in the current body of literature. The overall performance of the benchmarked nnU-Net network outperformed the traditional and AI-based approached reported in previous studies [16, 20, 40, 41]. Based on the findings in the present study, it is believed that the above superiorities of the nnU-Net network would further bring added values to the community for midbrain segmentation, as well as TCS-based PD assessment, in the future.

For our clinical objectives, we have meticulously selected these three models (U-Net, U-Net+++ and nnU-Net) to conduct a thorough and comprehensive benchmarking study [18-20, 42, 43]. First, for our specific clinical task, we selected the U-Net model for its classic, well-established, and effective performance in supervised segmentation tasks, as acknowledged in literature. Given our fully supervised segmentation task and the meticulous annotation by experienced clinicians with over a decade of practice, we possess a valuable dataset. Furthermore, the U-Net+++ model was chosen for its extensive skip connections and deep supervision, enhancing supervision quality. Additionally, the nnU-Net integrates the complete training pipeline, from pre-processing to post-processing, based on the U-Net architecture. This choice optimizes the training process for our benchmark tasks. Secondly, the three selected networks are relatively more mature and readily applicable in clinics than some emerging networks, considering the length of time they have been developed, evaluated, and applied for a wide spectrum of segmentation tasks over the past decade. We believe that performing a deeper and more comprehensive analysis of their performance via the submitted work, from the perspectives of segmentation agreement, model stability and robustness analysis etc, would provide enlightening insights for their applications when it comes to clinical implementation.

Intriguingly, results of this study indicated that although both the U-Net and U-Net+++ models achieved good segmentation agreements even they were inferior to the nnU-Net model **(Table 1A-1B)**, their segmentation capabilities dramatically degraded when it comes to the images corrupted by the

salt-and-pepper noise (SNR=0.95) and the Gaussian noise (σ=0.1) even at a low intensity level of noise addition **(Table 3)**. Conversely, the nnU-Net network not only outperformed on TCS images at the same level of noise corruption **(Table 3)**, its segmentation power remained even at greater degree of noise corruption (SNR=0.85, 0.75; σ=0.30, 0.50), as illustrated in **Table 4**. We speculated that the exceptional performance of the nnU-Net model in midbrain segmentation may be attributed to its unique and comprehensive training architecture, distinguishing it from the U-Net and U-Net+++ models. Although the network structure of the nnU-Net was not modified, its adaptive nature enables nnU-Net to automatically adjust the underlying hyper-parameters, such as the exact patch size, batch size, and inference settings, to adapt to the input dataset, while integrating image preprocessing strategies and image augmentation techniques seamlessly into the training workflow, crucially contribute to nnU-Net's superior performance.

In addition, it is worth noting that the present work contained multiple layers of strengths. First, we enrolled a remarkably larger cohort of data (n=584) compared to the previous studies (n=34-130). The inclusion of a larger sample size enhanced the validity of the findings from this study. Second, the findings were presented in a wide spectrum of application scopes beyond segmentation agreement. Particularly, the noise-corruption test was applied for the first time on TCS-based midbrain auto-segmentation for securing model robustness; this would hopefully provide the community with more valuable insights into clinical contextualization of the benchmarked model from application point-of-view. Third, three independent randomization seeds for train-test splits and 5-fold cross-validation were applied in this study in the hope of generating more reliable findings of the developed networks; these techniques have been proven to be effective in previous studies [34, 35, 44], and is believed to enhance the validity of the findings in this study.

From a public health standpoint, TCS exhibits a great application potential for cost-effective population-wide screening and frequent surveillance for timely PD management owing to its advantages of being radiation-free, high accessibility and high affordability [9, 45]. Several research groups have also explored the utility of methods based on other imaging modalities as valuable tools for early diagnosis of PD [46, 47]. However, such as MRI due to their low accessibility and affordability to the general public, particularly in developing or under-developed countries [48-50], as well as Positron Emission Tomography and Single-Photon Emission Computed Tomography due to their low accessibility, low affordability, and the potential radiation hazards [51, 52]. As such, we hope that results of this study would provide enlightening insights and stimulate researchers in the field to steer more focuses on AI-assisted TCS-based PD assessment in the long run, providing a cost-effective alternative to both the sufferers and medical practitioners in PD management in the smart-ageing era.

In spite of the encouraging results, the study presents several shortcomings that worth further investigations in the future. First, results of this study were generated with the dataset collected from a single institution and the same scanner vendor, which may limit the generalizability of the benchmarked model in real-world clinical settings. A multi-center study would improve and validate model generalizability, so a multi-center study with TCS data acquired from different vendors are warranted in the future in order to further validate findings of this work. Second, although the sample size of 584 in this study was considerably larger than previous studies where the sample sizes ranged from 40 to 130, further investigations using a larger cohort are preferred in the context of deep learning to improve and validate model generalizability. Third, for the robustness analysis, we chose salt-and-pepper noise and Gaussian noise in the study, the two most common types of noise in medical imaging, effectively simulating the majority of ultrasound noise scenarios. Additionally, other speckle noise and poisson noise are two types of noises that medical professionals may encounter. We anticipate further detailed discussions on noise impact in future research. Finally, due to the retrospective nature of this

study, although this study recruited highly experienced physicians (>10 years) in ultrasound imaging for ground-truth annotation generation, there may exist intra- or inter-rater variabilities in the dataset. Therefore, addressing this issue in a prospective study is warranted in the future to further facilitate widespread adoption of the developed model in clinical practice.

5 Conclusions

The nnU-Net model was benchmarked to be the best-performing model for midbrain segmentation in terms of segmentation agreement, model stability, prediction time efficiency, and model robustness, compared to the U-Net and U-Net+++ models. Despite the discussed limitations, results of this study presented multiple layers of superiorities in comparison to previous works, in terms of sample size, comprehensive scope of model evaluation, as well as the enhanced reliability owing to the adoption of multiple independent randomization seeds together with 5-fold cross-validation. Results of this study would provide enlightening insights and stimulate researchers in the field to steer more focuses on cost-effective AI-assisted TCS-based PD assessment in the long run, potentially benefiting millions of sufferers as well as medical practitioners worldwide for PD management in the smart-ageing era. Moving forward, a multi-center multi-vendor study is warranted in the future when it comes to clinical implementation.

6 Acknowledgement

This study was partially supported by Research Institute for Smart Ageing of The Hong Kong Polytechnic University (1-CD5B).

7 Ethical statement

Ethical approval for this study was obtained from the Human Subject Ethics Sub-committee (HSESC) of the Hong Kong Polytechnic University (HSEARS20231102004).

8 Declaration of interests

All the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Maserejian, L. Vinikoor-Imler, A. Dilley, Estimation of the 2020 Global Population of Parkinson's Disease (PD), Movement Disord, 35 (2020) S79-S80.

[2] G.B.D.N. Collaborators, Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet Neurol, 18 (2019) 459-480.

[3] B.R. Bloem, M.S. Okun, C. Klein, Parkinson's disease, Lancet, 397 (2021) 2284-2303.

[4] S. Behnke, D. Berg, M. Naumann, G. Becker, Differentiation of Parkinson's disease and atypical parkinsonian syndromes by transcranial ultrasound, J Neurol Neurosurg Psychiatry, 76 (2005) 423-425.

[5] L. Zecca, D. Berg, T. Arzberger, P. Ruprecht, W.D. Rausch, M. Musicco, D. Tampellini, P. Riederer, M. Gerlach, G. Becker, In vivo detection of iron and neuromelanin by transcranial sonography: A new of substantia approach for early detection nigra damage, Movement Disord, 20 (2005) 1278-1285.

[6] Y.J. Bae, J.M. Kim, C.H. Sohn, J.H. Choi, B.S. Choi, Y.S. Song, Y. Nam, S.J. Cho, B. Jeon, J.H. Kim, Imaging the Substantia Nigra in Parkinson Disease and Other Parkinsonian Syndromes, Radiology, 300 (2021) 260-278.

[7] D. Berg, K. Seppi, S. Behnke, I. Liepelt, K. Schweitzer, H. Stockner, F. Wollenweber, A. Gaenslen, P. Mahlknecht, J. Spiegel, J. Godau, H. Huber, K. Srulijes, S. Kiechl, M. Bentele, A. Gasperi, T. Schubert, T. Hiry, M. Probst, V. Schneider, J. Klenk, M. Sawires, J. Willeit, W. Maetzler, K. Fassbender, T. Gasser, W. Poewe, Enlarged Substantia Nigra Hyperechogenicity and Risk for Parkinson Disease A 37-Month 3-Center Study of 1847 Older Persons, Arch Neurol-Chicago, 68 (2011) 932-937.

[8] Y.L. Mei, J. Yang, Z.R. Wu, Y. Yang, Y.M. Xu, Transcranial Sonography of the Substantia Nigra for the Differential Diagnosis of Parkinson's Disease and Other Movement Disorders: A Meta-Analysis, Parkinsons Dis-Us, 2021 (2021).

[9] P. Bartova, D. Skoloudik, M. Bar, P. Ressner, P. Hlustik, R. Herzig, P. Kanovsky, Transcranial Sonography in Movement Disorders, Biomed Pap, 152 (2008) 251-258.

[10] D. Skoloudik, T. Fadrna, P. Bartova, K. Langova, P. Ressner, O. Zapletalova, P. Hlustik, R. Herzig, P. Kannovsky, Reproducibility of sonographic measurement of the substantia nigra, Ultrasound Med Biol, 33 (2007) 1347-1352.

[11] A. Berardelli, G.K. Wenning, A. Antonini, D. Berg, B.R. Bloem, V. Bonifati, D. Brooks, D.J. Burn, C. Colosimo, A. Fanciulli, J. Ferreira, T. Gasser, F. Grandas, P. Kanovsky, V. Kostic, J. Kulisevsky, W. Oertel, W. Poewe, J.P. Reese, M. Relja, E. Ruzicka, A. Schrag, K. Seppi, P. Taba, M. Vidailhet, EFNS/MDS-ES recommendations for the diagnosis of Parkinson's disease, Eur J Neurol, 20 (2013) 16-+.

[12] D. Monaco, D. Berg, A. Thomas, V. Di Stefano, F. Barbone, M. Vitale, C. Ferrante, L. Bonanni, M. Di Nicola, T. Garzarella, L.P. Marchionno, G. Malferrari, R. Di Mascio, M. Onofrj, R. Franciotti, The predictive power of transcranial sonography in movement disorders: a longitudinal cohort study, Neurol Sci, 39 (2018) 1887-1894.

[13] S. Zhu, Y.X. Wang, Y.Y. Jiang, R.X. Gu, M. Zhong, X. Jiang, B. Shen, J. Zhu, J. Yan, Y. Pan, L. Zhang, Clinical Features in Parkinson's Disease Patients with Hyperechogenicity in Substantia Nigra: A Cross-Sectional Study, Neuropsych Dis Treat, 18 (2022) 1593-1601.

[14] P. Singh, A neutrosophic-entropy based clustering algorithm (NEBCA) with HSV color system: A special application in segmentation of Parkinson's disease (PD) MR images, Comput Meth Prog Bio, 189 (2020).

[15] D. Basukala, R. Mukundan, A. Lim, M.A. Hurrell, R.J. Keenan, J.C. Dalrymple-Alford, T.J. Anderson, D.J. Myall, T.R. Melzer, Automated segmentation of substantia nigra and red nucleus using quantitative susceptibility mapping images: Application to Parkinson's disease, Comput Electr Eng, 91 (2021).

[16] S.A. Ahmadi, M. Baust, A. Karamalis, A. Plate, K. Boetzel, T. Klein, N. Navab, Midbrain Segmentation in Transcranial 3D Ultrasound for Parkinson Diagnosis, Medical Image Computing and Computer-Assisted Intervention, Miccai 2011, Pt Iii, 6893 (2011) 362-+.

[17] A. Sakalauskas, A. Lukosevicius, K. Lauckaite, D. Jegelevicius, S. Rutkauskas, Automated segmentation of transcranial sonographic images in the diagnostics of Parkinson's disease, Ultrasonics, 53 (2013) 111-121.

[18] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Lect Notes Comput Sc, 9351 (2015) 234-241.

[19] H.M. Huang, L.F. Lin, R.F. Tong, H.J. Hu, Q.W. Zhang, Y. Iwamoto, X.H. Han, Y.W. Chen, J. Wu, Unet 3+: A Full-Scale Connected Unet for Medical Image Segmentation, Int Conf Acoust Spee, (2020) 1055-1059.

[20] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat Methods, 18 (2021) 203-211.

[21] Y. Hiasa, Y. Otake, M. Takao, T. Ogawa, N. Sugano, Y. Sato, Automated Muscle Segmentation from Clinical CT Using Bayesian U-Net for Personalized Musculoskeletal Modeling, Ieee T Med Imaging, 39 (2020) 1030-1040.

[22] K.B. Chen, Y. Xuan, A.J. Lin, S.H. Guo, Lung computed tomography image segmentation based on U-Net network fused with dilated convolution, Comput Meth Prog Bio, 207 (2021).

[23] S.Q. Li, J.Y. Zheng, D.J. Li, Precise segmentation of non-enhanced computed tomography in patients with ischemic stroke based on multi-scale U-Net deep network model, Comput Meth Prog Bio, 208 (2021).

[24] H. Dong, G. Yang, F.D. Liu, Y.H. Mo, Y.K. Guo, Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks, Comm Com Inf Sc, 723 (2017) 506-517.

[25] H.F. Cui, Y.W. Chang, L. Jiang, Y. Xia, Y.N. Zhang, Multiscale attention guided U-Net architecture for cardiac segmentation in short-axis MRI images, Comput Meth Prog Bio, 206 (2021).

[26] A.M.G. Allah, A.M. Sarhan, N.M. Elshennawy, Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information, Expert Syst Appl, 213 (2023).

[27] J. Yang, M. Faraji, A. Basu, Robust segmentation of arterial walls in intravascular ultrasound images using Dual Path U-Net, Ultrasonics, 96 (2019) 24-33.

[28] M. Amiri, R. Brooks, B. Behboodi, H. Rivaz, Two-stage ultrasound image segmentation using U-Net and test time augmentation, Int J Comput Ass Rad, 15 (2020) 981-988.

[29] M. Amiri, R. Brooks, H. Rivaz, Fine-Tuning U-Net for Ultrasound Image Segmentation: Different Layers, Different Outcomes, Ieee T Ultrason Ferr, 67 (2020) 2510-2518.

[30] G.P. Chen, L. Li, Y. Dai, J.X. Zhang, M.H. Yap, AAU-Net: An Adaptive Attention U-Net for Breast Lesions Segmentation in Ultrasound Images, Ieee T Med Imaging, 42 (2023) 1289-1300.

[31] A. Sakalauskas, K. Lauckaite, A. Lukosevicius, D. Rastenyte, Computer-Aided Segmentation of the Mid-Brain in Trans-Cranial Ultrasound Images, Ultrasound in Medicine and Biology, 42 (2016) 322-332.

[32] F. Milletari, S.A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Botzel, N. Navab, Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound, Comput Vis Image Und, 164 (2017) 92-102.

[33] M. Weinreich, J.J. Chudow, B. Weinreich, T. Krumerman, T. Nag, K. Rahgozar, E. Shulman, J. Fisher, K.J. Ferrick, Development of an Artificially Intelligent Mobile Phone Application to Identify Cardiac Devices on Chest Radiography, Jacc-Clin Electrophy, 5 (2019) 1094-1095.

[34] S.K. Lam, Y. Zhang, J. Zhang, B. Li, J.C. Sun, C.Y. Liu, P.H. Chou, X. Teng, Z.R. Ma, R.Y. Ni, T. Zhou, T. Peng, H.N. Xiao, T. Li, G. Ren, A.L. Cheung, F.K. Lee, C.W. Yip, K.H. Au, V.H. Lee, A.T. Chang, L.W. Chan, J. Cai, Multi-Organ Omics-Based Prediction for Adaptive Radiation Therapy Eligibility in Nasopharyngeal Carcinoma Patients Undergoing Concurrent Chemoradiotherapy, Front Oncol, 11 (2021) 792024.

[35] J. Zhang, S.K. Lam, X.Z. Teng, Z.R. Ma, X.Y. Han, Y.P. Zhang, A.L.Y. Cheung, T.C. Chau, S.C.Y. Ng, F.K.H. Lee, K.H. Au, C.W.Y. Yip, V.H.F. Lee, Y. Han, J. Cai, Radiomic feature repeatability and its impact on prognostic model generalizability: A multi-institutional study on nasopharyngeal carcinoma patients, Radiother Oncol, 183 (2023).

[36] M.V. Sherer, D.N. Lin, S. Elguindi, S. Duke, L.T. Tan, J. Cacicedo, M. Dahele, E.F. Gillespie, Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review, Radiother Oncol, 160 (2021) 185-191.

[37] K. Harrison, H. Pullen, C. Welsh, O. Oktay, J. Alvarez-Valle, R. Jena, Machine Learning for Auto-Segmentation in Radiotherapy Planning, Clin Oncol-Uk, 34 (2022) 74-88.

[38] A.C. Bovik, Handbook of image and video processing, Academic press2010.

[39] T. Peng, C.Y. Tang, Y.Y. Wu, J. Cai, H-SegMed: A Hybrid Method for Prostate Segmentation in TRUS Images via Improved Closed Principal Curve and Improved Enhanced Machine Learning, Int J Comput Vision, 130 (2022) 1896-1919.

[40] C. Gonzalez, K. Gotkowski, A. Bucher, R. Fischbach, I. Kaltenborn, A. Mukhopadhyay, Detecting When Pre-trained nnU-Net Models Fail Silently for Covid-19 Lung Lesion Segmentation, Medical Image Computing and Computer Assisted Intervention - Miccai 2021, Pt Vii, 12907 (2021) 304-314.

[41] L. Huo, X.X. Hu, Q. Xiao, Y.J. Gu, X. Chu, L. Jiang, Segmentation of whole breast and fibroglandular tissue using nnU-Net in dynamic contrast enhanced MR images, Magn Reson Imaging, 82 (2021) 31-41.

[42] D. Gut, Z. Tabor, M. Szymkowski, M. Rozynek, I. Kucybala, W. Wojciechowski, Benchmarking of Deep Architectures for Segmentation of Medical Images, IEEE Trans Med Imaging, 41 (2022) 3231-3241.

[43] R. Azad, E.K. Aghdam, A. Rauland, Y. Jia, A.H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J.P. Cohen, E. Adeli, D. Merhof, Medical image segmentation review: The success of u-net, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2024).

[44] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat Comput, 21 (2011) 137-146.

[45] D. Skoloudik, M. Jelinkova, J. Blahuta, P. Cermak, T. Soukup, P. Bartova, K. Langova, R. Herzig, Transcranial Sonography of the Substantia Nigra: Digital Image Analysis, Am J Neuroradiol, 35 (2014) 2273-2278.

[46] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, S. Tangaro, Complex networks reveal early MRI markers of Parkinson's disease, Med Image Anal, 48 (2018) 12-24.

[47] J. Zhang, Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease, Npj Parkinsons Dis, 8 (2022).

[48] B. Heim, F. Krismer, R. De Marzi, K. Seppi, Magnetic resonance imaging for the diagnosis of Parkinson's disease, J Neural Transm, 124 (2017) 915-964.

[49] S. Lehericy, M.A. Sharman, C.L. Dos Santos, R. Paquin, C. Gallea, Magnetic resonance imaging of the substantia nigra in Parkinson's disease, Movement Disord, 27 (2012) 822-830.

[50] M. Brammerloh, M. Morawski, I. Friedrich, T. Reinert, C. Lange, P. Pelicon, P. Vavpetic, S. Jankuhn, C. Jager, A. Alkemade, R. Balesar, K. Pine, F. Gavriilidis, R. Trampel, E. Reimer, T. Arendt, N. Weiskopf, E. Kirilina, Measuring the iron content of dopaminergic neurons in substantia nigra with MRI relaxometry, Neuroimage, 239 (2021).

[51] T. Brucke, S. Djamshidian, G. Bencsits, W. Pirker, S. Asenbaum, I. Podreka, SPECT and PET imaging of the dopaminergic system in Parkinson's disease, J Neurol, 247 (2000) 2-7.

[52] G. Pagano, F. Niccolini, M. Politis, Imaging in Parkinson's disease, Clin Med, 16 (2016) 371-375.

Supplementary Materials

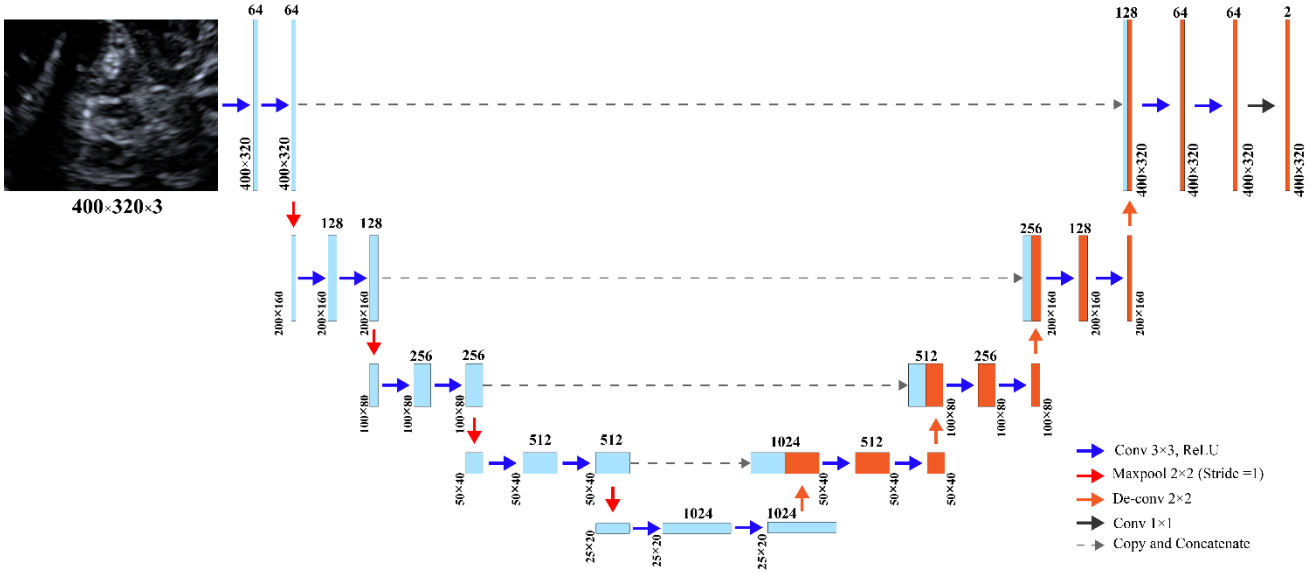

Fig. 1A. U-Net architecture

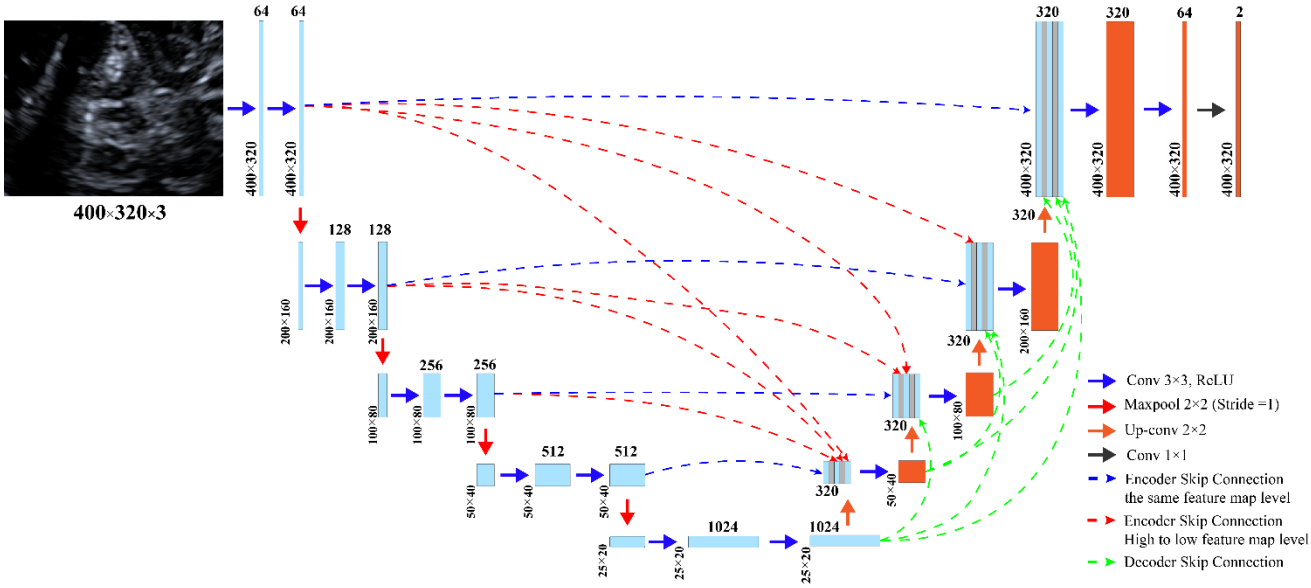

Fig. 1B. U-Net+++ architecture

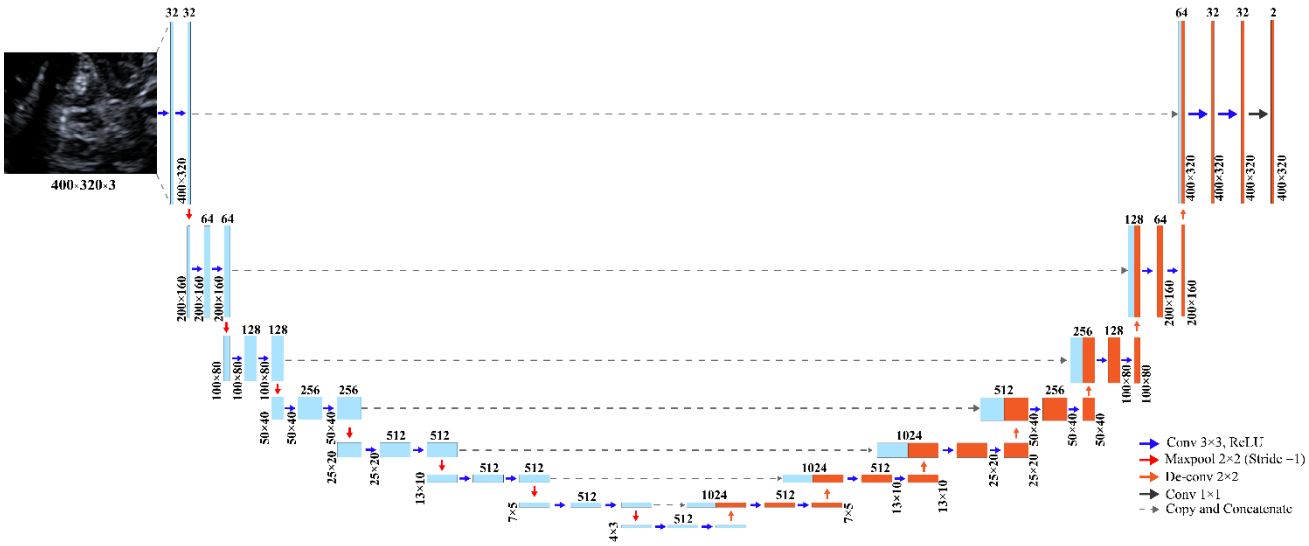

Fig. 1C. nnU-Net architecture

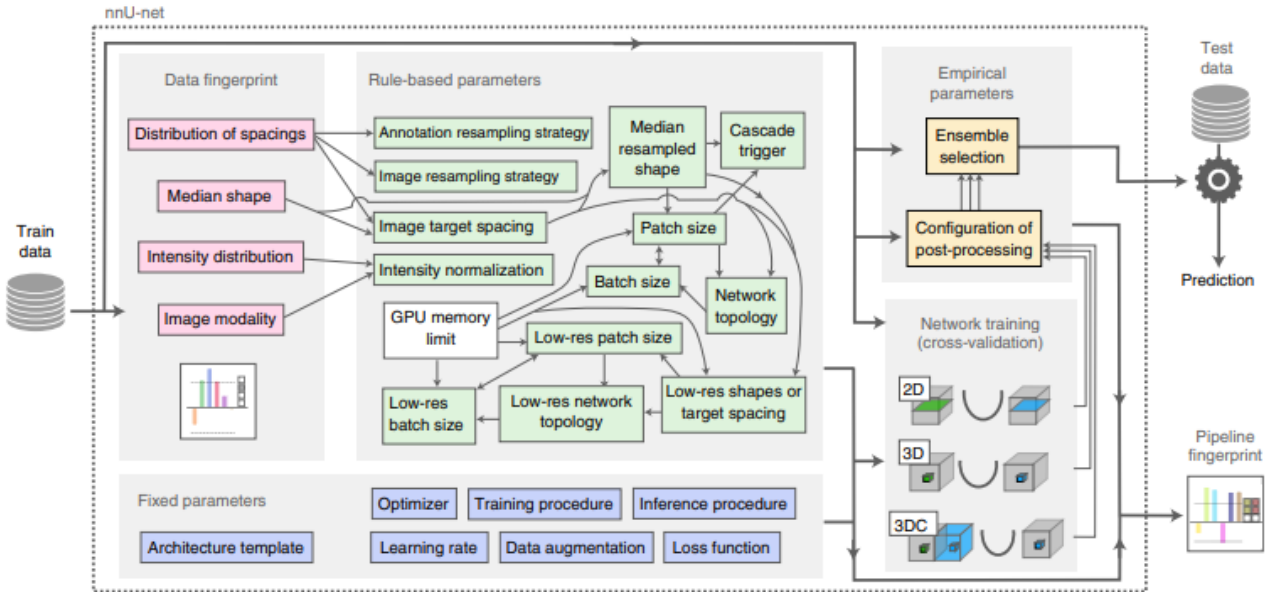

Fig. 1D. Flowchart of the nnU-Net architecture [20]

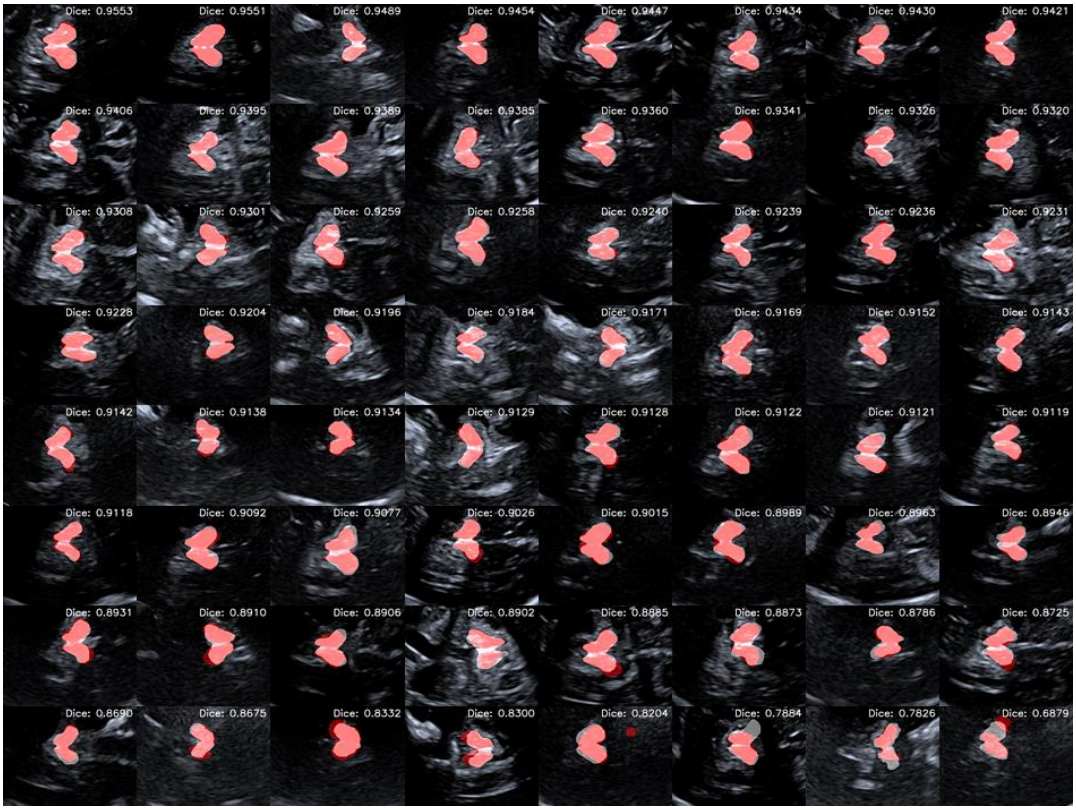

Fig. 2A. Qualitative visualization of the U-Net segmentation agreement with quantitative DICE score for all the 64 testing cases under the randomization seed 1, where the red regions represent the predicted segments and the white regions indicate the ground-truth annotations.

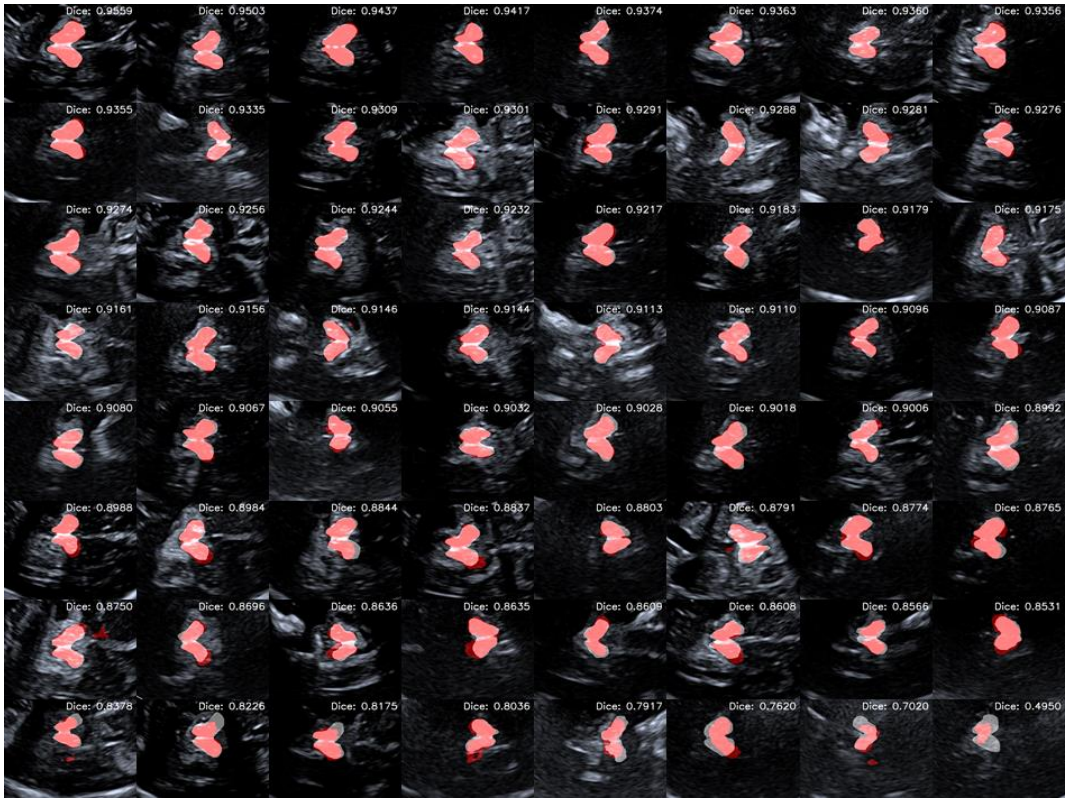

Fig. 2B. Qualitative visualization of the U-Net+++ segmentation agreement with quantitative DICE score for all the 64 testing cases under the randomization seed 1, where the red regions represent the predicted segments and the white regions indicate the ground-truth annotations.

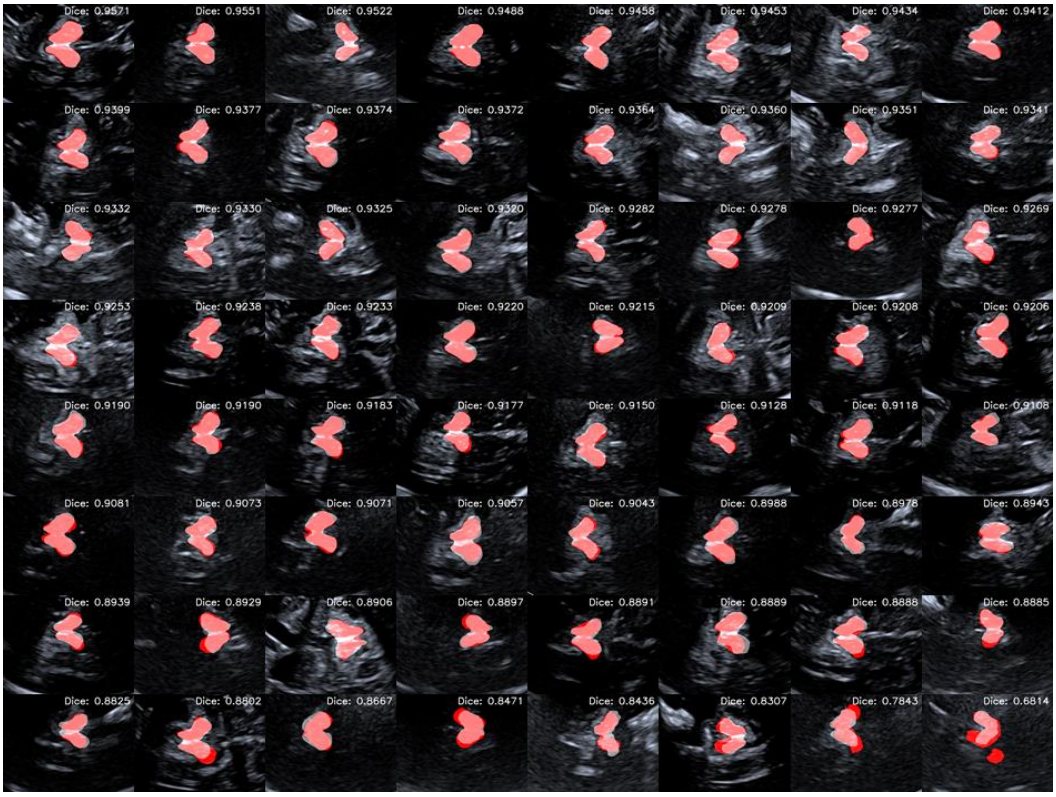

Fig. 2C. Qualitative visualization of the nnU-Net segmentation agreement with quantitative DICE score for all the 64 testing cases under the randomization seed 1, where the red regions represent the predicted segments, and the white regions indicate the ground-truth annotations.

Table 1A. Results of segmentation agreement and stability of the three comparing neural networks in each of the three randomization seeds under 5 cross-validation on validation dataset. Averaged (AVG) scores of the segmentation agreement metrics in DICE, IoU, and HD are presented for each model in each randomization seed. Standard deviation (SD) was calculated to illuminate the underlying model stability. The top-ranked scores are bolded.

		Randomization Seed 1			Randomization Seed 2			Randomization Seed 3		
Network	**Fold**	**DICE**	**IoU**	**HD (mm)**	**DICE**	**IoU**	**HD (mm)**	**DICE**	**IoU**	**HD (mm)**
U-Net	Fold1	0.915	0.848	2.710	0.912	0.840	2.804	0.919	0.850	2.499
	Fold2	0.905	0.828	2.567	0.899	0.818	2.921	0.900	0.819	2.854
	Fold3	0.914	0.842	2.775	0.909	0.834	2.766	0.913	0.845	2.834
	Fold4	0.918	0.849	2.719	0.920	0.854	2.515	0.918	0.850	2.555
	Fold5	0.922	0.855	2.534	0.926	0.862	2.407	0.915	0.845	2.783
	AVG	0.915	0.845	2.661	0.913	0.842	2.683	0.913	0.842	2.705
	SD	0.006	0.011	0.105	0.010	0.017	0.214	0.008	0.013	0.166
U-Net+++	Fold1	0.909	0.835	3.126	0.908	0.832	3.198	0.909	0.836	2.898
	Fold2	0.892	0.807	3.069	0.893	0.810	3.097	0.893	0.810	3.189
	Fold3	0.910	0.836	2.906	0.901	0.824	3.419	0.906	0.831	3.016
	Fold4	0.914	0.843	3.084	0.913	0.840	2.961	0.914	0.842	2.958
	Fold5	0.915	0.845	2.890	0.918	0.848	2.859	0.918	0.849	2.881
	AVG	0.908	0.833	3.015	0.906	0.831	3.107	0.908	0.834	2.988
	SD	0.010	0.015	0.109	0.010	0.015	0.217	0.009	0.015	0.124
nnU-Net	Fold1	0.967	0.939	1.158	0.963	0.934	1.308	0.972	0.949	0.932
	Fold2	0.975	0.954	0.887	0.965	0.936	1.051	0.969	0.943	1.121
	Fold3	0.969	0.943	1.053	0.963	0.934	1.124	0.968	0.941	1.151
	Fold4	0.975	0.954	0.940	0.966	0.939	1.092	0.970	0.946	1.130
	Fold5	0.966	0.939	1.123	0.961	0.931	1.727	0.964	0.935	1.287

AVG	0.970	0.946	1.032	0.964	0.935	1.261	0.969	0.943	1.124
SD	0.004	0.007	0.116	0.002	0.003	0.279	0.003	0.005	0.127

Table 1B. Results of segmentation agreement (in terms of DICE, IoU, and HD) and stability (in terms of SD) of the three deep-learning networks across all the three randomization seeds under 5-fold cross validation on validation dataset. The top-ranked scores are bolded.

Network	DICE (SD)	IoU (SD)	HD (mm) (SD)
U-Net	0.914 (0.008)	0.843 (0.013)	2.683 (0.156)
U-Net+++	0.908 (0.009)	0.833 (0.014)	3.037 (0.155)
nnU-Net	**0.968 (0.004)**	**0.941 (0.007)**	**1.139 (0.200)**