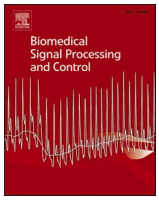
Biomedical Signal Processing and Control 97 (2024) 106667

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



Single annotator versus multi-annotator: Challenge of segmenting two neighboring hippocampus head and body with high precision

Hossein Arabi [a], Habib Zaidi [a,b,c,d,*]

[a] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva 4, Switzerland
[b] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, 9700 RB Groningen, Netherlands
[c] Department of Nuclear Medicine, University of Southern Denmark, DK-500 Odense, Denmark
[d] University Research and Innovation Center, Óbuda University, Budapest, Hungary

ARTICLE INFO

Keywords:
MRI
Hippocampus
Segmentation
Deep learning
Multi-annotator

ABSTRACT

Background and objective: This work investigates multi-annotator segmentation framework for semantic image segmentation using deep learning approaches.
Methods: The framework consists of various deep learning models which perform independently, and then their decisions are fused by a combiner network. The underlying idea of the multi-annotator (ensemble) segmentation framework is to take advantage of the synergistic effect created by the complementary information provided by the different annotators to reach a peak performance superior to each of the annotators when performing alone. The performance of the multi-annotator framework was assessed for the challenging task of the hippocampus head and body segmentation from MR images. Atlas-based segmentation method, deep learning-based multi-view segmentation approach, as well as different decision fusion algorithms, such as STAPLE and shape-based averaging, were assessed versus the proposed multi-annotator segmentation framework.
Results: Among the different deep learning architectures and loss functions, ResNet model with cross-entropy loss function (ResNet-CE) exhibited the best performance (best performing single model) with a Dice of 88.6 ± 1.8 % and 88.6 ± 1.8 % for the hippocampus head and body, respectively. However, the multi-annotator segmentation framework (which also involved Resnet-CE as one of the annotators) exhibited superior performance with Dice of 91.1 ± 1.3 % (head) and 91.0 ± 1.3 % (body) compared to the atlas-based method with Dice of 88.5 ± 1.5 % (head) and 88.4 ± 1.5 % (body) and multi-view approach with Dice of 89.0 ± 1.4 % (head) and 88.9 ± 1.5 % (body).
Conclusion: The multi-annotator segmentation framework exhibits superior performance over each annotator alone which could be employed almost for all machine learning-related problems to enhance the overall effectiveness of the deep/machine learning algorithms.

1. Introduction

The hippocampus, which converts short-term memory to long-lasting memory, is one of the earliest subcortical regions that undergo remarkable tissue loss/degradation within the progression of Alzheimer's Disease. In this regard, the shape and volume of the hippocampi have been considered to be early and effective biomarkers for the prediction and diagnosis of Alzheimer's Disease [1], schizophrenia [2], and epilepsy [3]. Hence, the development of automated algorithms for the identification and delineation of hippocampi from anatomical images, such as magnetic resonance images (MRI), is of great interest [4]. The hippocampal atrophy is the early clinical manifestation of various neurological diseases, wherein clinicians need to (manually) segment this tissue and perform morphological analysis [5]. Manual delineation is regarded as the reference/gold standard for the morphological analysis of the hippocampus, yet is highly time-consuming, tedious, and suffers from intra- and inter-observer variability [6].

Automated hippocampal delineation, which is a pixel-wise semantic tissue/organ segmentation task, has remarkably progressed owing to the recent development of deep learning approaches and their efficient capabilities to extract meaningful/discriminative image features [7,8]. The hippocampus, as a gray matter structure, bears relatively very low contrast with the neighboring structures/tissues in MR images (for instance T1-weighted MPRAGE sequence). In addition, its irregular

* Corresponding author at: Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).

https://doi.org/10.1016/j.bspc.2024.106667
Received 3 January 2024; Received in revised form 30 April 2024; Accepted 25 July 2024

shape, small volume, and large variations across the different subjects have rendered the automated hippocampus segmentation highly challenging [9]. In this regard, a number of segmentation models based on convolutional neural networks have been proposed to reliably delineate the hippocampus structures across different individuals presenting with either normal or affected brain morphology due to neurodegenerative diseases [9–11].

To cope with the challenge of semantic segmentation of complex tissues/organs, the focus has been on the development of fully convolutional networks providing an effectively large receptive field to guarantee high-level semantic feature/information extraction/detection [12]. The fusion of features obtained from the multi-scale processing of input images proved to be a highly efficient way of aggregating complementary information to create comprehensive feature maps [13,14]. In addition, in multi-channel semantic segmentation, for instance multi-sequence MR image segmentation, not only every single channel should create an effective feature map, but the model should also establish efficient connections between the different feature maps from the different channels [15].

Besides, to further improve the performance of neural networks for the task of semantic segmentation, the attention technique, mimicking the cognitive attention, is introduced which tends to enhance the importance of the more relevant/determining features/regions while fading out the rest [16]. This would allow the models to devote more computing power on a sub-section but an important part of the data [17]. Depending on the size/structure of the target organ/tissue, a hierarchical attention mechanism could be employed to create a comprehensive feature map using a context-aware feature extraction module. Different stages of such a network would concentrate to detect/extract low-level (voxel-level) to high-level semantic information from the important part of the input data to efficiently guide the tissue/organ segmentation [18].

Most studies in the literature set out to develop a single but powerful semantic segmentation model which enables reliable and accurate delineation of the target organs/tissues [19]. However, several works have demonstrated that a combination of multiple annotators or ensemble learning (though each of which has a sub-optimal performance) is able to outperform a single powerful annotator [20–22]. In this regard, a number of deep learning-based multi-view solutions have been proposed to deal with computer-aided detection problems [22,23]. Though the decisions by the different annotators would not be uncorrelated, a sort of latent complementary and interrelated information exists within the different annotators' decisions regarding the true delineation of the target tissue/organ [20,24]. In other words, an efficient combination of weak/sub-optimal annotations would create a synergistic impact to outperform a single powerful/comprehensive annotator.

The idea of employing multiple annotators to delineate a target structure has been investigated for hippocampus segmentation from a single MR input image [25]. To this end, the different neural network annotators (with the same architecture) estimated the boundary of the target organ from different perspectives (views of the input image) [26]. The target structure was segmented from 9 different views of the input MR images, wherein a three-dimensional network aggregated/combined the label decisions from the multiple views.

In this study, we aimed to conduct a thorough evaluation of the performance of a single annotator compared to multiple annotators in the intricate task of hippocampus body and head segmentation [27]. Various neural networks, each characterized by distinct structures and loss functions, will be employed as multiple annotators. These will be juxtaposed against the top-performing single annotator. Instead of presenting varied perspectives in a 2-dimensional mode to a single neural network [25], we will leverage different network architectures and loss functions in a 3-dimensional (3D) mode, fostering a diverse array of label decisions for a synergistic multi-annotator segmentation approach. Additionally, we will implement atlas-based methods [28] to establish a performance baseline for evaluating the segmentation frameworks of both single and multi-annotators.

The main contributions of this study lie in several key areas. Firstly, a novel approach was introduced to ensemble deep learning segmentation by diversifying the deep learning architectures used for training sub-models within the ensemble. Unlike previous studies that typically rely on a single architecture, this work thoroughly explored various architectures to identify the most effective ones, thus enhancing segmentation performance. Secondly, the number of sub-models within the ensemble was optimized, providing flexibility across different datasets and applications, and maximizing segmentation accuracy. Additionally, an exhaustive analysis of a wide range of loss functions was conducted in conjunction with different deep learning models, optimizing the selection to tailor the ensemble framework for specific datasets and tasks, thereby improving segmentation accuracy and robustness. Finally, our comprehensive comparative analysis included evaluations against shape-based averaging, atlas-based methods, and multi-view segmentation, highlighting the superiority of ensemble deep learning in terms of accuracy, robustness, and outlier reduction, particularly in medical imaging applications.

2. Materials and methods

2.1. Datasets

Two datasets consisting of MR images with manually defined hippocampus labels were employed in this study. The first dataset was obtained from Alzheimer's Disease Neuroimaging Initiative (http://www.loni.usc.edu/ADNI) containing 50 T1-weighted brain MR images together with the corresponding hippocampus masks. Since the hippocampus only occupies a small area of the brain in MR images, the MR images were cropped roughly to contain separately left and right hippocampus. In total, 100 training samples (considering left and right hippocampus separately) were created from the 50 subjects. This dataset was used only for pretraining of the models.

The second dataset belongs to the medical segmentation Decathlon challenge (http://medicaldecathlon.com) containing in total (considering left and right hippocampus separately) 260 3D T1-weighted MPRAGE sequence (TI/TE/TR, 860/3.7/8.0 ms and a voxel size of 1.0 mm^3) MR images. The binary masks of the hippocampus head and body were manually defined on the MR images [6]. This dataset was used for final training (fine-tuning) and evaluation of the models. Prior to the implementation of the segmentation algorithms, the MR images underwent denoising and N4 bias corrections [29,30]. Then, the intensity normalization on the MR images was carried out using intensity levels at 90 % cumulative histogram of each image independently.

2.2. Multi-annotator (ensemble) seminal segmentation

Multi-annotator seminal segmentation leverages the synergistic benefits of integrating label decisions from multiple annotators to yield a solution superior to each annotator's individual outcome. Each annotator addresses the segmentation problem uniquely, showcasing varied perspectives or sensitivities to underlying features/information. An effective framework capitalizes on the complementary information present in the sub-optimal annotators' decisions. For the multi-annotator segmentation, diverse deep learning architectures with distinct loss functions were employed to autonomously segment the hippocampus from the input MR image (Fig. 1). Each network was individually optimized for peak performance in 3D mode, delineating the hippocampus head and body. The best-performing network was chosen to integrate label decisions from different networks using the combiner network (depicted as "Combiner" in Fig. 1). The combiner network fuses decisions based on 3D probability maps independently generated by diverse networks/annotators, in addition to incorporating the input image.

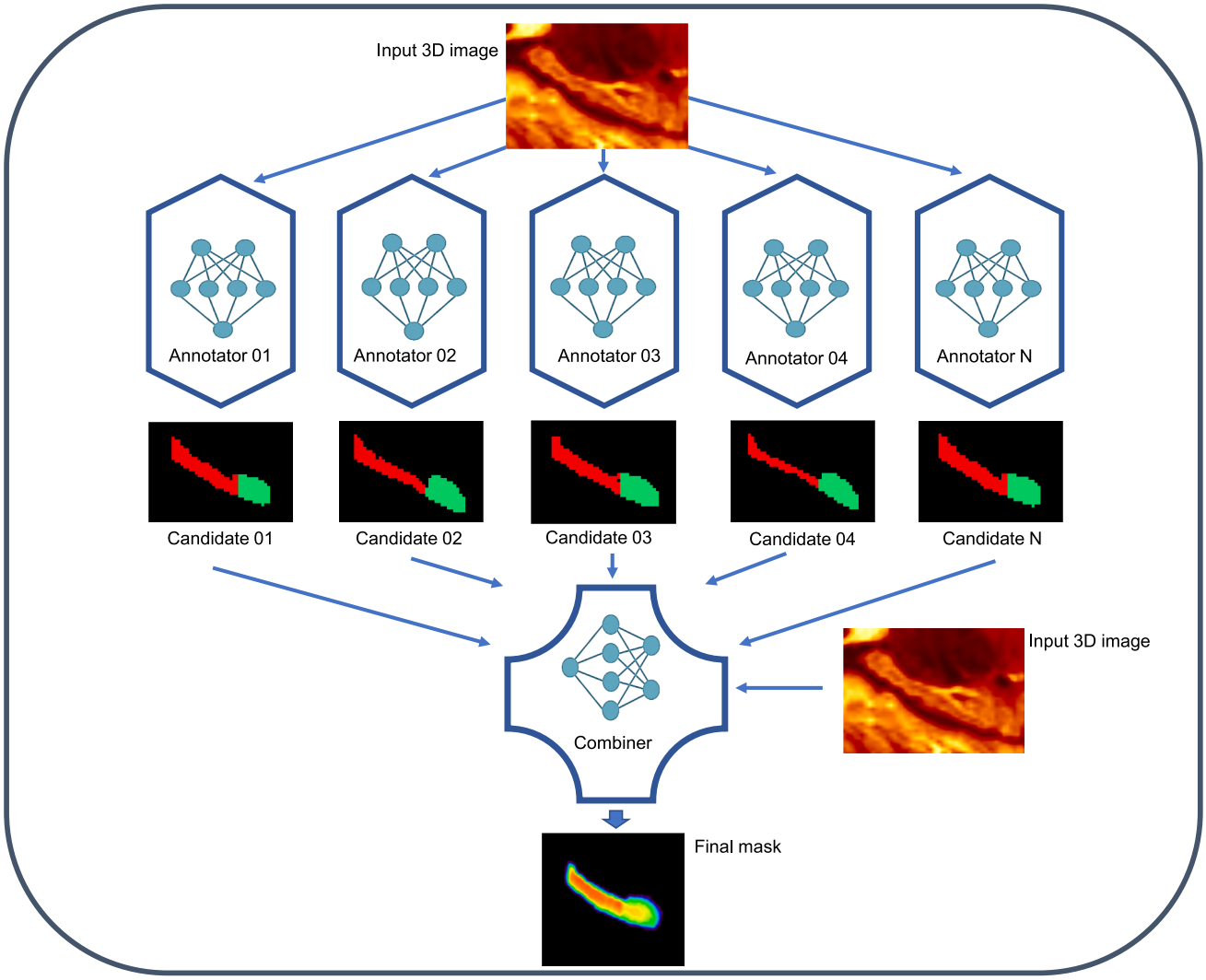


Fig. 1. illustration of the multi-annotator segmentation framework, wherein independently generated 3D probability maps by distinct annotators (networks) are integrated by the combiner network to produce the ultimate organ/tissue mask.

2.3. Alternative approaches

2.3.1. Multi-view segmentation

Alternative to the multi-annotator segmentation approach, the identification, and delineation of the target organ/tissue could be performed from multiple views (such as sagittal, coronal, and axial) using an identical neural network followed by an aggregator network to combine the label decisions [25,26]. Fig. 2 illustrates the multi-view segmentation approach proposed in [26], wherein the input MR image is processed from nine different views including sagittal, coronal, axial, and six diagonal views. In the next step, the different label decisions are fused (by the combiner network) in 3D mode. The input of the combiner network are the 9 probability maps generated from the 9 different views plus the input 3D MR image. To implement this approach, the best performing neural network architecture was employed to serve as the combiner as well as the annotator from the different views.

2.3.2. Atlas-based

Despite the exceptional capability/performance of the deep learning algorithms to decode the complex patterns/relationships of the input data [31], sub-optimal performance may be observed regarding capturing/modeling of the shape/morphology characteristics of the target structure [32]. In this regard, atlas-based methods are able to properly account for the shape/morphology of the target structure owing to the presence of prior knowledge in the form of template/atlas images [33]. In this light, two voxel-wise weighting atlas-based approaches were implemented for comparison with the deep learning-based methods. These methods rely on the phase congruency maps (PCMs) [34] and local normalized cross-correlation (LNCC) similarity indices [35], calculated between the target MR image and each of the template/atlas images, to define voxel-wise weighting factors for the fusion of the template/atlas images registered to the target MR image. These approaches are referred to as Atlas-PCM and Atlas-LNCC, respectively.

2.3.3. Decision fusion techniques

Given the label decisions generated by the different annotators (Fig. 1), the final label map can be also calculated using the popular label fusion techniques, including majority voting [28], simultaneous truth and performance level estimation (STAPLE) [36], and shape-based averaging [37,38]. For the sake of comprehensive investigation, the performance of the deep learning-based multi-annotator segmentation framework, as well as the atlas-based methods, was compared with the above-mentioned label fusions techniques.

2.4. Implementation details

To realize the multi-annotator segmentation framework, a number of

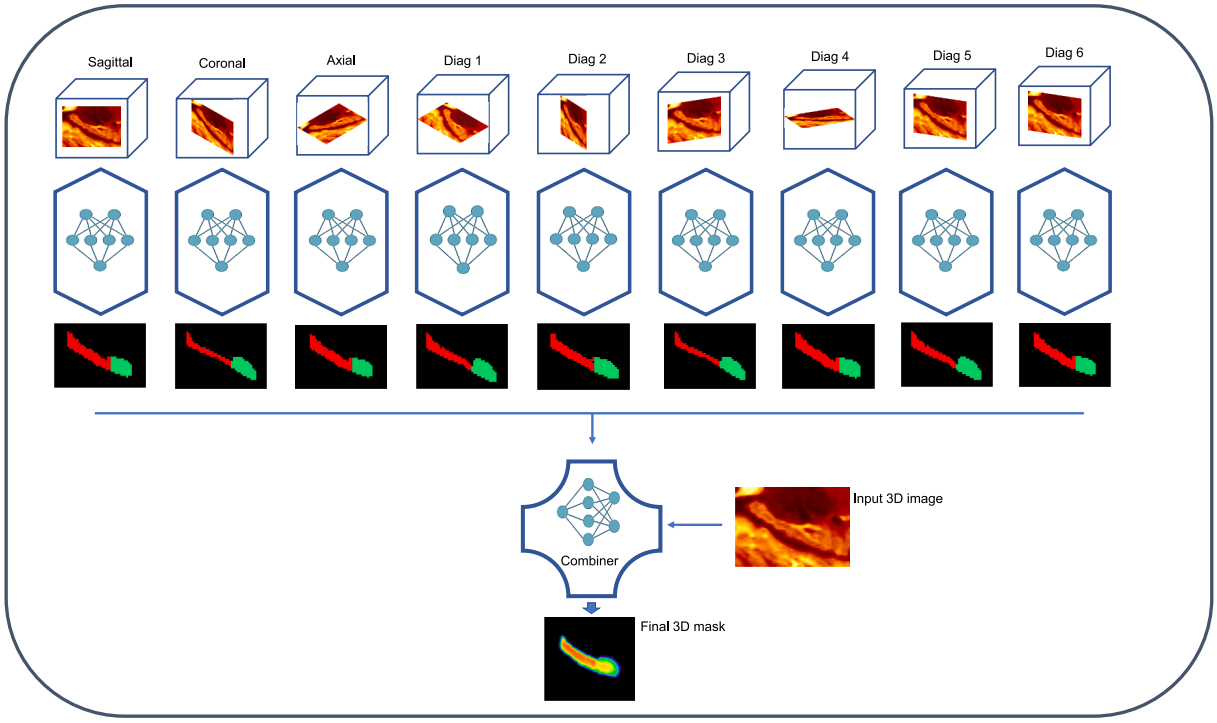


Fig. 2. Multi-view segmentation where 2D segmentations from the nine different views are combined to generate the final segmentation mask [26].

neural network architectures with different loss functions were implemented to serve as the different annotators in Fig. 1. These networks include Dense-Vnet [39], ResNet [40], no-new-net (NNnet) [10], eCNN [41], HolisticNet [42], SimGAN [43,44], and 3D_Unet [45]. In addition to the different neural network architectures, different loss functions [46], including cross-entropy (CE), Dice, Dice no-square (Dice_NS), generalized Wasserstein Dice (GW_Dice), sensitivity_specificity (SS), and Tversky, were examined for each of these networks to determine the best combination of the models (for the task of multi-annotator segmentation) as well as the best performing network and loss function. These models were trained using a 10 fold cross-validation scheme, then the different combinations of the models were examined to determine the optimal configuration of the multi-annotator segmentation network (which networks with which loss functions). To increase the size of the training dataset, data augmentation technique including rotation within a range of (−15° to 15°), ±15 % scaling, elastic deformation with sigma = 1, and random flipping was exploited.

The best performing single model (turned out to be ResNet with CE loss function (Resnet-CE)) was selected for the combiner network in the multi-annotator framework (Fig. 1) as well as realization of the multi-view segmentation approach (Fig. 2).

2.5. Evaluation strategy

The evaluation of the different semantic segmentation methods was performed versus the manually defined label masks of the hippocampus body and head using standard image segmentation metrics, including Dice cooefficient (Eq. (1), Jaccard coefficient (JC) (Eq. (2), relative volume difference (RVD) (Eq. (3), sensitivity (S) (Eq. (4), mean absolute surface distance (MASD) (Eq. (5), and Hausdorff distance (HD) (Eq. (6).

$$Dice(Ref, Tst) = \frac{2|Ref \cap Tst|}{|Ref| + |Tst|} \tag{1}$$

$$RVD(Ref, Tst) = 100 \times \frac{|Tst| - |Ref|}{|Ref|} \tag{2}$$

$$JC(Ref, Tst) = \frac{|Ref \cap Tst|}{|Ref \cup Tst|} \tag{3}$$

$$S(Ref, Tst) = \frac{|Ref \cap Tst|}{|Tst|} \tag{4}$$

$$MASD(Ref, Tst) = \frac{d_{ave}(S_{Ref}, S_{Tst}) + d_{ave}(S_{Tst}, S_{Ref})}{2} \tag{5}$$

$$HD(Ref, Tst) = \max_{Ref}\{\min_{Tst}\{d(Ref, Tst)\}\} \tag{6}$$

where *Ref* is the reference binary mask, *Tst* denotes the estimated binary mask of the target structures from the MR images. $d_{ave}(S_{Ref},S_{Tst})$ stands for the average of the direct distances from all points on the surface of the reference mask S_{Ref} to the estimated surface mask S_{Tst}. The Hausdorff distance indicates the maximum distance between the boundaries of the reference volume (*Ref*) and the estimated structure volume (*Tst*).

The paired *t*-test analysis was performed to investigate the significance of the differences between the results obtained from the different segmentation approaches, considering p-values smaller than 0.05 as statistically significant.

3. Results

To realize the multi-annotator segmentation framework (Fig. 1), the different neural network architectures, along with a number of commonly used loss functions, were implemented to determine the most efficient combination of the models as well as the best performing single model. Overall, the ResNet model with cross-entropy (CE) loss function (Resnet-CE) exhibited superior performance with Dice = 88.7 ± 1.6 and 88.6 ± 1.8 for the hippocampus head and body, respectively, among all network architectures and loss functions. Hence, the Resnet-CE model, as the best performing model, was selected to implement the multi-view segmentation framework (Fig. 2) as well as the combiner network in the multi-annotator framework (Fig. 1). Among the various combinations of the network architectures and loss functions, the multi-annotator

segmentation framework reached its peak performance with the following models: Resnet-CE, Dense-Vnet with sensitivity–specificity loss function (DVnet-SS), ResNet with Tversky loss function (Resnet-Tv), no-new-net with Dice loss function (NNnet-Dice), HolisticNet with cross-entropy loss function (Holnet-CE), and Dense-Vnet with generalized Wasserstein Dice loss function (DVnet-WD). These six models are not necessarily the best performing among the several architectures and loss functions implemented in this work. However, the combination of these models in the form of multi-annotator led to the best overall outcome. Other models, such as Dense-Vnet with Tversky loss function (DVnet-Tv), ResNet with Dice, and generalized Wasserstein Dice loss functions (Resnet-Dice and Resnet-WD, respectively), and no-new-net with cross-entropy loss function (NN-CE) exhibited even better results compared to the selected models. Nevertheless, the combination of these models led to a sub-optimal performance supposedly due to their highly correlated outcomes. Table 1 compares Dice indices obtained from these models.

In this segmentation task focusing on the hippocampus head, our investigation revealed that the optimal number of annotators was determined to be six. Surprisingly, the addition of more annotators did not yield improvements in the overall segmentation quality, as evidenced in Supplemental Fig. 1. On the contrary, augmenting the number of annotators actually led to a degradation in model performance. This deterioration can be attributed to the heightened complexity introduced by an increased number of annotators, which posed challenges for the combiner model in effectively integrating diverse segmentation outputs. This crucial observation underscores the importance of optimizing the number of annotators for each specific segmentation task. Moving forward, it becomes imperative to carefully consider and tailor the number of annotators to maximize segmentation efficacy in future endeavors.

Fig. 3 depicts representative views of the hippocampus body and head segmented by the different approaches. The visual inspection revealed some differences between the binary masks obtained from the different approaches as a result of the vague hippocampus boundary in the input MR images. To meticulously investigate the performance of these methods, the overall quantitative metrics should be considered.

Table 2 summarizes the segmentation accuracy of the different deep learning models, atlas-based, as well as commonly used decision fusion algorithms, such as STAPLE and shape-based averaging (SBA). It should be noted that majority voting, STAPLE, and SBA algorithms were applied to the outcomes of the six deep learning models, which were employed in the multi-annotator segmentation framework. Regarding the quantitative metrics, the multi-annotator method (Dice = 91.1 and 91.0) outperformed the Resnet-CE model as the best performing deep learning model (Dice = 87.7 and 88.6), multi-view method (Dice = 89.0 and 88.9), and the Atlas-LNCC approach (Dice = 88.5 and 88.4) for the hippocampus head and body segmentation, respectively. The p-values calculated between the multi-annotator segmentation approach and other methods are presented in Table 3 which demonstrates that these differences are statistically significant.

Figs. 4 to 7 depict the box plots of the Dice and HD metrics achieved by the different segmentation frameworks for the hippocampus body and head. The Resnet-CE model, as the best performing single deep learning network, exhibited a comparable performance to atlas-based and STAPLE methods. However, this model failed to properly identify the hippocampus from the input MR images in two cases which are visible in Figs. 3 to 6 as outliers. Fig. 7 illustrates one of these outliers wherein the other approaches could properly detect the hippocampus body and head, while the Resnet-CE model led to a gross error.

4. Discussion

4.1. Multi-annotator performance

The primary aim of this work was to introduce/investigate the multi-annotator segmentation framework in comparison to the conventional deep learning approach (wherein a single network is trained and evaluated), atlas-based method, and deep learning-based multi-view segmentation strategy. The primary idea behind the multi-annotator segmentation framework relies on the fact that different deep learning architectures as well as loss functions would capture certain features/signals/characteristics from the input image (or bear particular sensitivity to certain image features) in order to identify the target organs/tissues. In this light, a multi-annotator segmentation framework would benefit from the complementary information provided by the different annotators to create a synergistic effect beyond the capacity of each of the annotators alone [20].

The objective of this study was not to conduct an exhaustive search to find the optimal combination of the deep learning architectures and loss functions (for the realization of the multi-annotator segmentation framework) through the implementation of a large number of models. Rather, we focused on the demonstration of the superiority of the multi-annotator segmentation framework over each of the annotators alone provided they could offer complementary information. For the implementation of the multi-annotator segmentation framework, only deep learning-based annotators were employed, whereas other segmentation approaches, such as atlas-based methods, could also participate in this framework.

4.2. Atlas-based methods

Atlas-based methods are still of special interest since they could offer realistic solutions, even for complex structures, owing to the access to prior knowledge in the form of templates or atlases [28,47]. In this regard, deep learning approaches may fail to capture/estimate the correct morphology of the input subject leading to gross errors and/or outliers [32]. In essence, atlas-based methods are sort of multi-annotator segmentation approaches wherein each template or atlas image provides a candidate segmentation of the target structure to the decision fusion

Table 1
Comparison of hippocampus body and head segmentation accuracy (mean ± SD) between the deep learning models using the selected evaluation metrics, including Dice coefficient, relative volume distance (RVD), Jaccard similarity (JC), sensitivity (S), mean absolute surface distance (MASD), Hausdorff distance (HD), and distance error (DE).

		Resnet-CE	DVnet-SS	Resnet-Tv	NNnet-Dice	Holnet-CE	DVnet-WD	DVnet-Tv	Resnet-Dice	NN-CE
Dice	Head	88.7 ± 1.6	77.6 ± 1.6	85.3 ± 1.7	82.2 ± 1.6	79.6 ± 1.7	76.2 ± 1.7	84.9 ± 1.7	85.5 ± 1.7	82.9 ± 1.7
	Body	88.6 ± 1.8	77.0 ± 1.7	84.6 ± 1.8	81.9 ± 1.7	79.0 ± 1.7	76.0 ± 01.7	84.1 ± 1.7	84.9 ± 1.8	82.2 ± 1.7
RVD (%)	Head	−8.8 ± 1.8	−10.3 ± 2.0	9.6 ± 1.9	9.9 ± 1.9	−10.1 ± 2.0	−10.8 ± 2.0	−9.9 ± 1.9	9.6 ± 1.9	−9.9 ± 1.9
	Body	−8.9 ± 1.8	−10.5 ± 2.0	9.9 ± 1.9	10.0 ± 2.0	−10.3 ± 2.0	−11.0 ± 2.1	−10.0 ± 1.9	9.9 ± 1.9	−10.0 ± 2.0
JC	Head	0.82 ± 0.01	0.71 ± 0.03	0.80 ± 0.02	0.78 ± 0.02	0.74 ± 0.03	0.70 ± 0.03	0.80 ± 0.02	0.80 ± 0.02	0.79 ± 0.02
	Body	0.82 ± 0.02	0.70 ± 0.03	0.80 ± 0.02	0.78 ± 0.02	0.73 ± 0.03	0.96 ± 0.04	0.79 ± 0.02	0.80 ± 0.02	0.78 ± 0.02
S	Head	0.82 ± 0.03	69.9 ± 0.05	0.79 ± 0.03	0.75 ± 0.04	0.73 ± 0.04	68.8 ± 0.05	0.79 ± 0.03	0.79 ± 0.03	0.75 ± 0.04
	Body	0.81 ± 0.03	69.1 ± 0.05	0.78 ± 0.03	0.74 ± 0.04	0.73 ± 0.04	68.3 ± 0.05	0.78 ± 0.03	0.78 ± 0.03	0.74 ± 0.04
MASD (mm)	Head	1.33 ± 0.08	2.97 ± 0.18	2.19 ± 0.11	2.68 ± 0.11	2.90 ± 0.17	3.10 ± 0.20	2.21 ± 0.11	2.20 ± 0.10	2.69 ± 0.11
	Body	1.49 ± 0.09	3.00 ± 0.19	2.20 ± 0.11	2.70 ± 0.12	2.99 ± 0.18	3.13 ± 0.20	2.22 ± 0.12	2.21 ± 0.11	2.71 ± 0.12
HD (mm)	Head	3.86 ± 0.21	6.33 ± 0.49	4.50 ± 0.29	4.91 ± 0.32	5.88 ± 0.39	6.18 ± 0.44	4.59 ± 0.30	4.47 ± 0.28	4.85 ± 0.31
	Body	4.11 ± 0.22	6.10 ± 0.50	4.71 ± 0.31	5.01 ± 0.35	5.94 ± 0.42	6.22 ± 0.56	4.78 ± 0.32	4.64 ± 0.30	4.91 ± 0.33

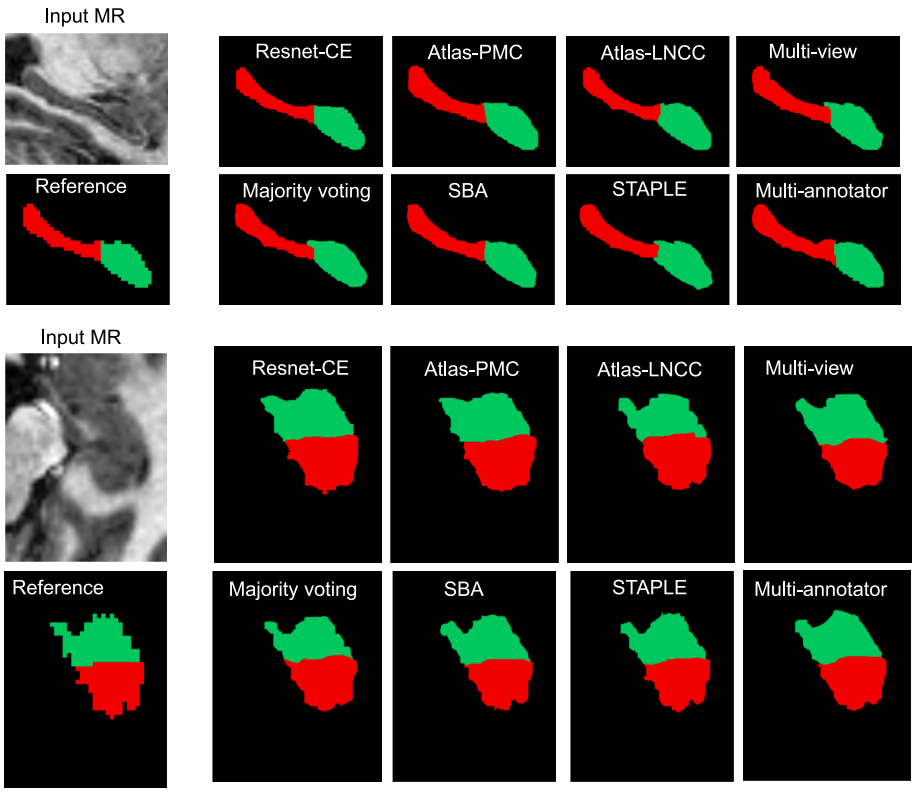


Fig. 3. Representative views of the delineated hippocampus using the different methods. The hippocampus head and body are depicted in green and red, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of hippocampus body and head segmentation accuracy (mean ± SD) between the different approaches using the selected evaluation metrics, including the Dice, relative volume distance (RVD), Jaccard similarity (JC), sensitivity (S), mean absolute surface distance (MASD), Hausdorff distance (HD), distance error (DE). The best model is highlighted in **bold font** and gray highlight, while the second-best model is highlighted solely in gray.

		Dice	RVD (%)	JC	S	MASD (mm)	HD (mm)
Single best model (Resnet-CE)	Head	88.7 ± 1.6	−8.8 ± 1.8	0.82 ± 0.01	0.82 ± 0.03	1.33 ± 0.08	3.86 ± 0.21
	Body	88.6 ± 1.8	−8.9 ± 1.8	0.82 ± 0.02	0.81 ± 0.03	1.49 ± 0.09	4.11 ± 0.22
Atlas-PCM	Head	87.9 ± 1.5	9.1 ± 1.7	0.81 ± 0.01	0.81 ± 0.03	1.99 ± 0.09	4.15 ± 0.23
	Body	87.5 ± 1.6	9.2 ± 1.8	0.80 ± 0.01	0.80 ± 0.03	2.09 ± 0.10	4.23 ± 0.24
Atlas-LNCC	Head	88.5 ± 1.5	8.4 ± 1.7	0.82 ± 0.01	0.82 ± 0.02	1.81 ± 0.08	3.93 ± 0.21
	Body	88.4 ± 1.5	8.5 ± 1.8	0.82 ± 0.01	0.81 ± 0.03	1.90 ± 0.09	4.18 ± 0.22
Multi-view	Head	89.0 ± 1.4	8.1 ± 1.7	0.83 ± 0.01	0.82 ± 0.02	1.40 ± 0.08	3.01 ± 0.20
	Body	88.9 ± 1.5	8.3 ± 1.8	0.82 ± 0.01	0.82 ± 0.03	1.46 ± 0.09	3.22 ± 0.21
Majority voting	Head	86.2 ± 1.6	9.5 ± 1.8	0.79 ± 0.01	0.79 ± 0.03	2.17 ± 0.10	4.93 ± 0.26
	Body	85.8 ± 1.6	9.9 ± 1.9	0.78 ± 0.01	0.78 ± 0.03	2.24 ± 0.11	5.02 ± 0.27
SBA	Head	87.9 ± 1.5	−8.9 ± 1.8	0.82 ± 0.01	0.81 ± 0.03	2.00 ± 0.09	4.05 ± 0.22
	Body	87.8 ± 1.7	−8.9 ± 1.8	0.82 ± 0.02	0.80 ± 0.03	2.01 ± 0.10	4.15 ± 0.23
STAPLE	Head	87.9 ± 1.4	−8.9 ± 1.7	0.82 ± 0.01	0.81 ± 0.02	1.98 ± 0.09	4.07 ± 0.22
	Body	87.9 ± 1.6	−8.9 ± 1.7	0.82 ± 0.01	0.80 ± 0.03	2.00 ± 0.10	4.10 ± 0.23
Multi-annotator	Head	**91.1 ± 1.3**	**−7.2 ± 1.6**	**0.85 ± 0.01**	**0.85 ± 0.02**	**1.05 ± 0.07**	**2.26 ± 0.16**
	Body	**91.0 ± 1.3**	**−7.3 ± 1.7**	**0.85 ± 0.01**	**0.85 ± 0.02**	**1.06 ± 0.07**	**2.33 ± 0.17**

Table 3
The p-values calculated between the multi-annotator approach and the other methods for the different quantitative metrics.

		Dice	**RVD (%)**	**JC**	**S**	**MASD (mm)**	**HD (mm)**
(Resnet-CE) vs Multi-annotator	Head Body	0.020.02	0.030.03	0.020.02	0.020.02	0.030.03	0.030.03
Atlas-PCM vs Multi-annotator	Head Body	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01
Atlas-LNCC vs Multi-annotator	HeadBody	0.020.02	0.020.02	0.020.02	0.020.02	0.020.02	0.020.02
Multi-view vs Multi-annotator	HeadBody	0.030.03	0.040.04	0.030.03	0.030.03	0.050.05	0.040.04
Majority voting vs Multi-annotator	HeadBody	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01	<0.01<0.01
SBA vs Multi-annotator	HeadBody	0.010.01	0.020.02	0.020.02	0.010.01	0.020.02	0.010.01
STAPLE vs Multi-annotator	HeadBody	0.020.02	0.030.03	0.030.03	0.030.03	0.030.03	0.030.03

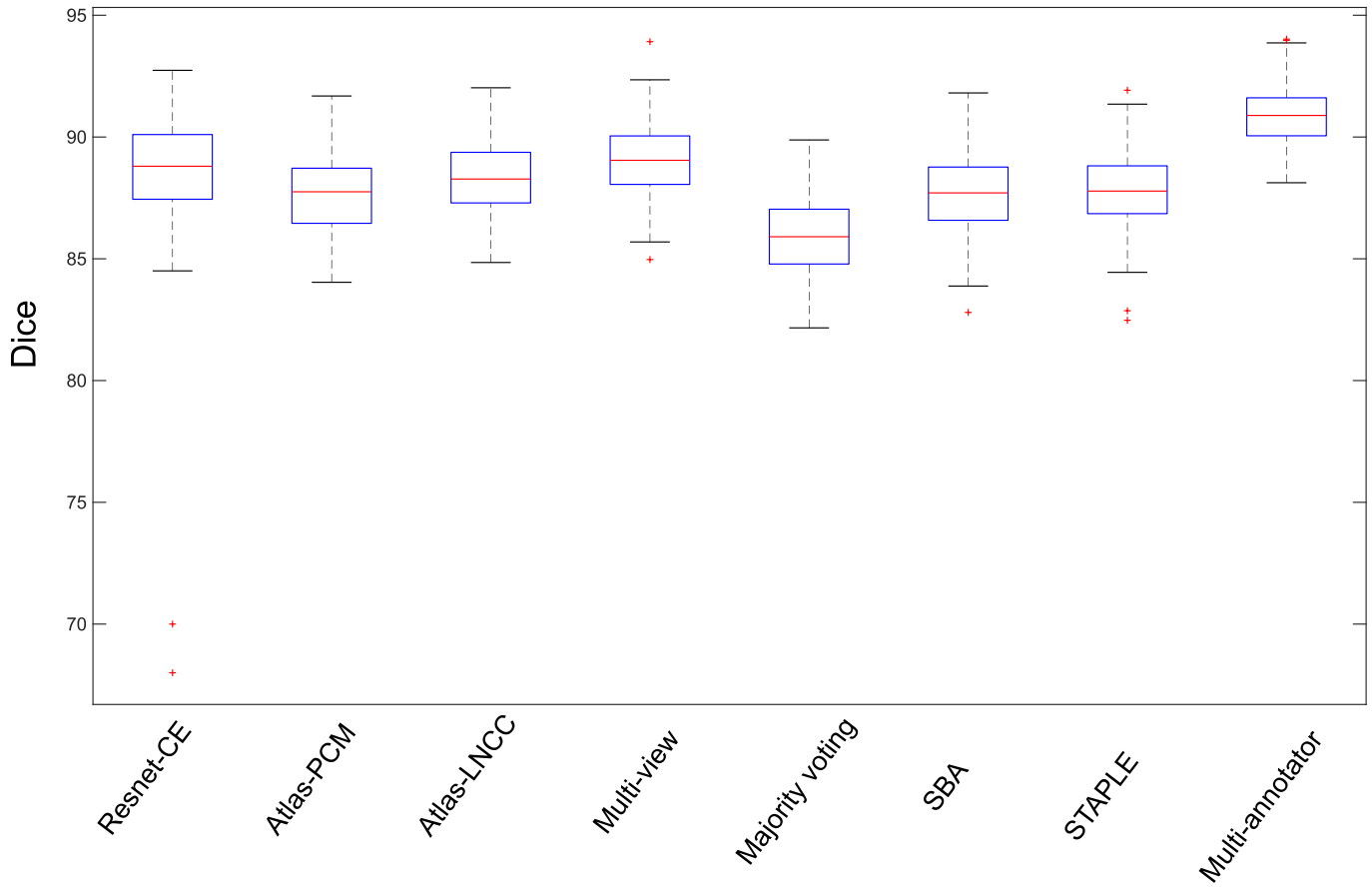


Fig. 4. Box plot of Dice coefficient obtained by the different segmentation algorithms for the hippocampus body.

algorithm. Incorporation of atlas-based methods in multi-annotator framework might offer highly complementary information to the other deep learning-based annotators, which would lead to overall better performance. The information in the atlas dataset could also be included in the multi-annotator framework through principal component analysis to avoid the computational burden of atlas-based methods [48].

4.3. Key parameters

The primary factor which determines the efficiency of the multi-annotator segmentation framework is the level/amount of complementary information provided by the different annotators. In this light, combination of a number of best-performing annotators would not necessarily lead to peak performance of the multi-annotator approach. Though each of the annotators may exhibit high accuracy, the strong correlations and considerable overlap between their outcomes would lead to suboptimal performance of the multi-annotator approach. Fig. 9A depicts such an example wherein the heatmap illustrates the overlap/correlation between the six best performing annotators while their combination in the form of the multi-annotator framework led to relatively sub-optimal performance. Fig. 9B depicts the overlap heatmap for the six annotators leading to the overall peak performance of the multi-annotator approach. The levels of overlap between the different annotators in Fig. 9B are noticeably lower than those in Fig. 1A; however, due to the presence of effective complementary information, these annotators resulted in overall superior outcomes. On the other hand, Fig. 9C depicts the overlap heatmap of six annotators which led to sub-optimal performance of the multi-annotator approach due to inaccurate outcomes of the different participating annotators and the presence of the less effective complementary information.

The ensemble learning approach showcased superior performance compared to individual models, underscoring its efficacy in enhancing segmentation accuracy. However, the optimization process employed in this study for efficient ensemble implementation remains specific to the segmentation task at hand. For diverse tasks, datasets, or problems, dedicated optimization of factors, such as the number and architecture of sub-models, as well as the choice of loss function, becomes imperative. Notably, the selection of loss function emerges as a critical determinant, exerting a profound impact on the outcomes of sub-models relative to network architecture. Thus, the incorporation of multiple loss functions capable of providing complementary predictions or estimations is paramount for the implementation of an efficient ensemble learning model. To this end, for novel tasks, it is advisable to meticulously optimize ensemble learning models by identifying the optimal combination of loss functions, ensuring maximal complementary information for the combiner model, thereby facilitating enhanced segmentation accuracy and robustness.

4.4. Outliers

In addition to the overall superior performance of the multi-annotator (ensemble) approach over the best performing single annotator, this approach did not exhibit gross errors (outliers such as Fig. 8) while every single annotator resulted in a few outliers when they were performing independently. In this study, the best performing model was used as the combiner to fuse the different decision maps (Fig. 1); however, specific network architectures and/or dedicated loss functions could be developed/employed to combine more effectively the different decision maps.

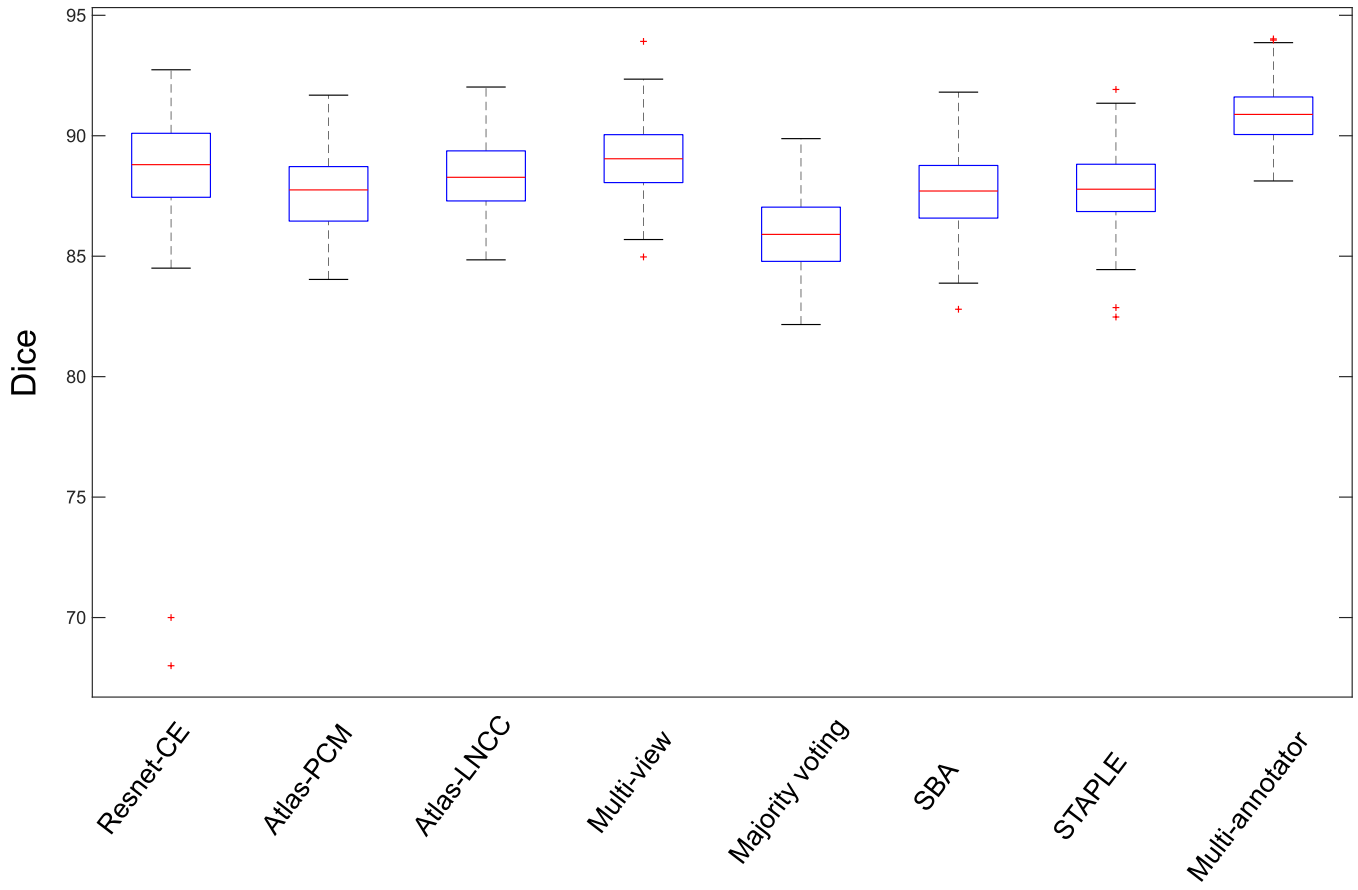


Fig. 5. Box plot of Dice coefficient obtained by the different segmentation algorithms for the hippocampus head.

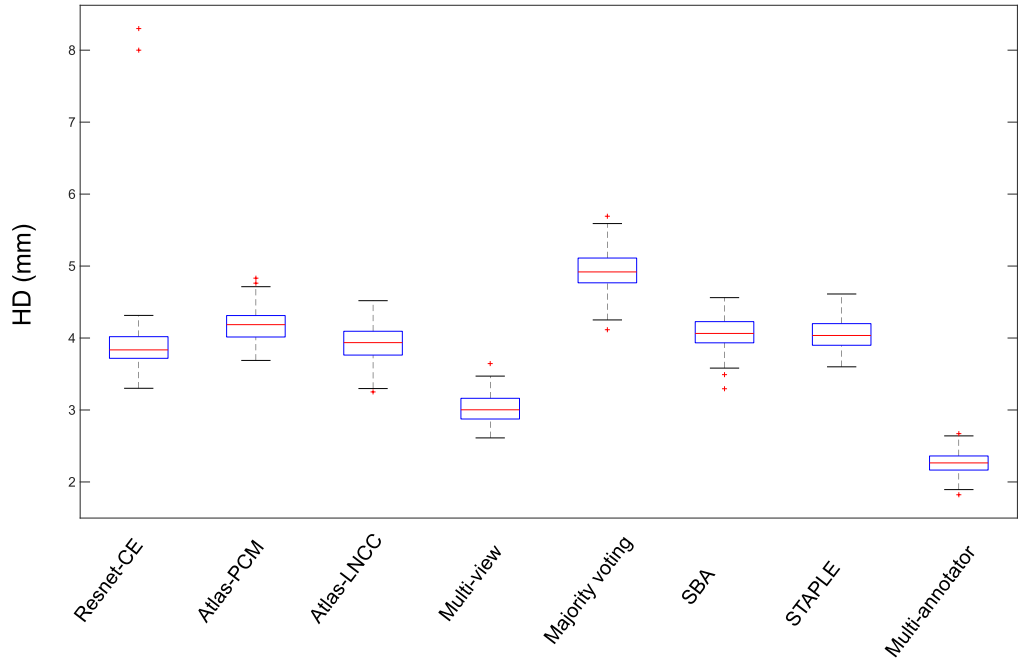


Fig. 6. Box plot of HD coefficient obtained from the different segmentation algorithms for the hippocampus body.

5. Conclusion

In this work, multi-annotator segmentation framework was investigated wherein a number of deep learning models were combined into a single pipeline to create a synergistic effect superior to the performance of each of the participating models alone. The multi-annotator approach exhibited superior performance over the atlas-based methods, STAPLE, as well as the single best performing deep learning model in the

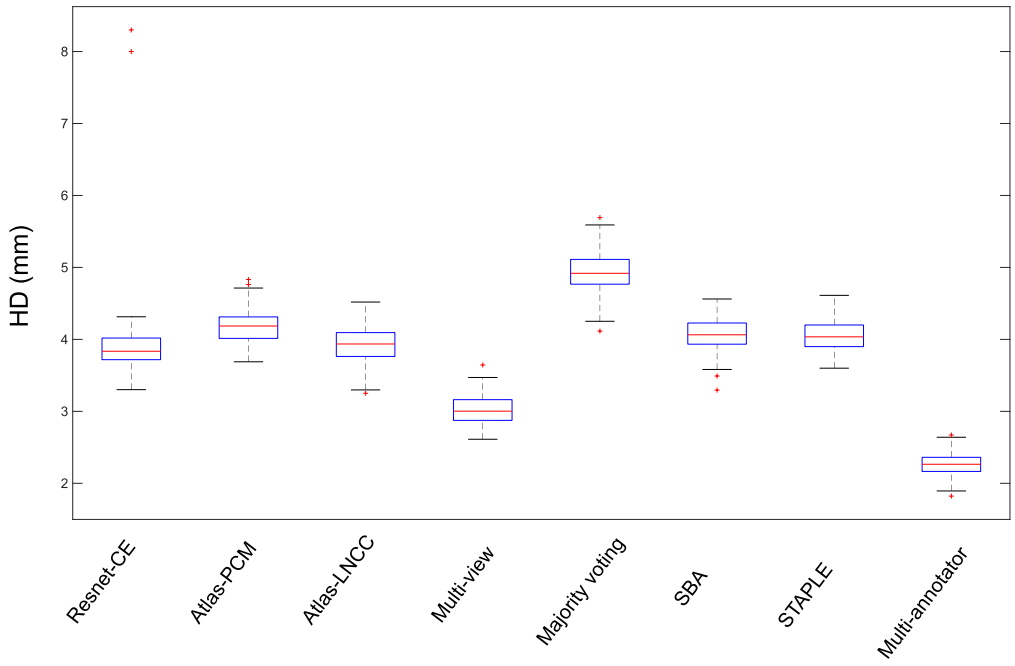


Fig. 7. Box plot of HD coefficient obtained from the different segmentation algorithms for the hippocampus head.

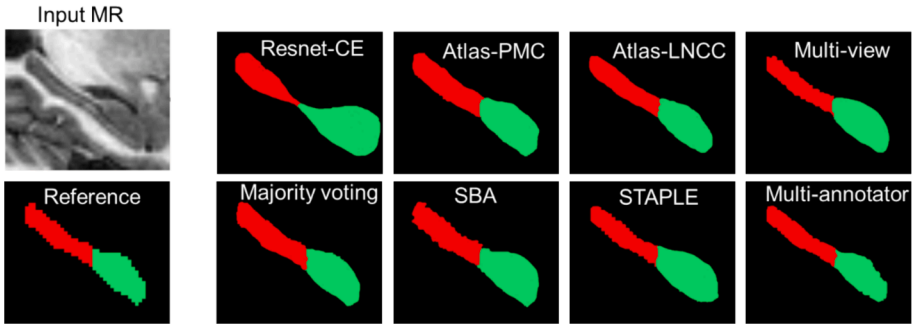


Fig. 8. Outlier report: The Resnet-CE network failed to properly identify the hippocampus head and body compared to the other methods.

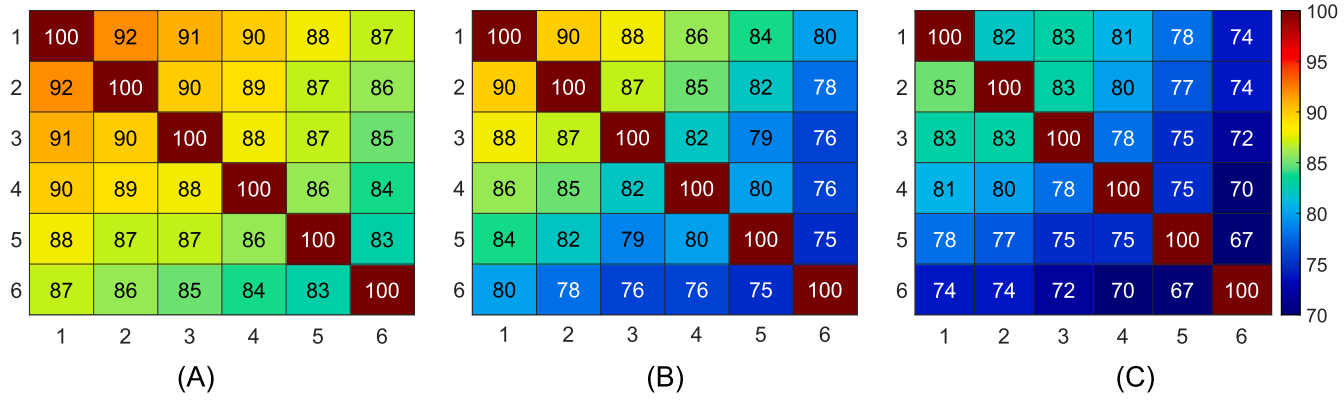


Fig. 9. The overlap head maps of (A) the six best performing deep learning-based annotators (which led to overall sub-optimal performance of the multi-annotator approach), (B) the six deep learning-based annotators which led to peak performance of the multi-annotator approach, and (C) the six deep learning-based annotators with poor individual and overall performance.

challenging task of the hippocampus body and head segmentation. The proposed framework could be employed almost for all machine learning-related problems to enhance the overall effectiveness of the deep/machine learning algorithms.

CRediT authorship contribution statement

Hossein Arabi: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Habib**

Zaidi: Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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.

Data availability

The authors do not have permission to share data.

Acknowledgment

This work was supported by the Swiss National Science Foundation under grant SNRF 320030_176052, the Private Foundation of Geneva University Hospitals under Grant RC-06-01.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2024.106667.

References

[1] K.H. Hobbs, et al., Quad-mesh based radial distance biomarkers for Alzheimer's disease, in: IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 19–23.
[2] M.D. Nelson, et al., Hippocampal volume reduction in schizophrenia as assessed by magnetic resonance imaging: a meta-analytic study, Arch. Gen. Psychiatry 55 (1998) 433–440.
[3] U.C. Wieshmann, et al., Development of hippocampal atrophy: a serial magnetic resonance imaging study in a patient who developed epilepsy after generalized status epilepticus, Epilepsia 38 (1997) 1238–1241.
[4] N. Nogovitsyn, et al., Testing a deep convolutional neural network for automated hippocampus segmentation in a longitudinal sample of healthy participants, Neuroimage 197 (2019) 589–597.
[5] P. Coupé, et al., Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease, Neuroimage 59 (2012) 3736–3747.
[6] A.L. Simpson, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, arXiv preprint arXiv:1902.09063, (2019).
[7] L.C. Chen, et al., DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 834–848.
[8] H. Arabi, H. Zaidi, Applications of artificial intelligence and deep learning in molecular imaging and radiotherapy, Eur J Hybrid Imaging 4 (2020) 17.
[9] J. Shi, et al., Discriminative feature network based on a hierarchical attention mechanism for semantic hippocampus segmentation, IEEE J. Biomed. Health Informatics (2020).
[10] F. Isensee, et al., nnu-net: Self-adapting framework for u-net-based medical image segmentation, arXiv preprint arXiv:1809.10486, (2018).
[11] Y. Yang, et al., Automatic Brain Mask Segmentation for Mono-modal MRI, in: Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics, 2020, pp. 124-128.
[12] A.E. Kavur, et al., CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation, Med. Image Anal. 69 (2020) 101950.
[13] P. Zhang, et al., Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202-211.
[14] B. Thyreau, et al., Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing, Med. Image Anal. 43 (2018) 214–228.
[15] D.M. Vigneault, et al., Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks, Med. Image Anal. 48 (2018) 95–106.
[16] F. Liu, et al., Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography, Med. Image Anal. 67 (2021) 101873.
[17] B. Lei, et al., Self-co-attention neural network for anatomy segmentation in whole breast ultrasound, Med. Image Anal. 64 (2020) 101753.
[18] Y. Wang, et al., Abdominal multi-organ segmentation with organ-attention networks and statistical fusion, Med. Image Anal. 55 (2019) 88–102.
[19] J. Sourati, et al., Active Deep learning with fisher information for patch-wise semantic segmentation, deep learning in medical image analysis and multimodal learning for clinical decision support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, hel, 11045 (2018) 83-91.
[20] D. Karimi, et al., Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Med. Image Anal. 65 (2020) 101759.
[21] M. Hatt, et al., The first MICCAI challenge on PET tumor segmentation, Med Image Anal 44 (2018) 177–195.
[22] H.R. Roth, et al., Improving computer-aided detection using convolutional neural networks and random view aggregation, IEEE Trans. Med. Imaging 35 (2016) 1170–1181.
[23] A.A. Setio, et al., Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks, IEEE Trans. Med. Imaging 35 (2016) 1160–1169.
[24] N. Tajbakhsh, et al., Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693.
[25] D. Ataloglou, et al., Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning, Neuroinformatics 17 (2019) 563–582.
[26] Y. Chen, et al., Hippocampus segmentation through multi-view ensemble ConvNets, in: IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 192–196.
[27] A. Arabi, H. Zaidi, Segmentation of the hippocampus head and body: Comparison of single annotator and multi-annotator, IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) (2021) 1–3.
[28] H. Arabi, H. Zaidi, Comparison of atlas-based techniques for whole-body bone segmentation, Med Image Anal 36 (2017) 98–112.
[29] N.J. Tustison, et al., N4ITK: improved N3 bias correction, IEEE Trans Med Imaging 29 (2010) 1310–1320.
[30] H. Arabi, H. Zaidi, Improvement of image quality in PET using post-reconstruction hybrid spatial-frequency domain filtering, Phys. Med. Biol. 63 (2018) 215010.
[31] H. Arabi, H. Zaidi, Deep learning-guided estimation of attenuation correction factors from time-of-flight PET emission data, Med Image Anal 64 (2020) 101718.
[32] I.S. Klyuzhin, et al., Testing the Ability of Convolutional Neural Networks to Learn Radiomic Features, medRxiv, (2020).
[33] H. Arabi, H. Zaidi, One registration multi-atlas-based pseudo-CT generation for attenuation correction in PET/MRI, Eur J Nucl Med Mol Imaging 43 (2016) 2021–2035.
[34] H. Arabi, et al., Atlas-guided generation of pseudo-CT images for MRI-only and hybrid PET-MRI-guided radiotherapy treatment planning, Phys Med Biol 61 (2016) 6531–6552.
[35] H. Arabi, H. Zaidi, Magnetic resonance imaging-guided attenuation correction in whole-body PET/MRI using a sorted atlas approach, Med Image Anal 31 (2016) 1–15.
[36] S.K. Warfield, et al., Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE Trans Med Imaging 23 (2004) 903–921.
[37] T. Rohlfing, C.R. Maurer Jr., Shape-based averaging, IEEE Trans Image Process 16 (2007) 153–161.
[38] H. Arabi, H. Zaidi, Whole-body bone segmentation from MRI for PET/MRI attenuation correction using shape-based averaging, Med Phys 43 (2016) 5848.
[39] E. Gibson, et al., Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks, IEEE Trans. Med. Imaging 37 (2018) 1822–1834.
[40] W. Li, et al., On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 348–360.
[41] A. Bahrami, et al., A new deep convolutional neural network design with efficient learning capability: application to CT image synthesis from MRI, Med. Phys. 47 (2020) 5158–5171.
[42] L. Fidon, et al., Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks, International MICCAI Brainlesion Workshop, Springer, 2017, pp. 64-76.
[43] Y. Hu, et al., Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks, Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment, Springer, 2017, pp. 105–115.
[44] H. Arabi, et al., Novel adversarial semantic structure deep learning for MRI-guided attenuation correction in brain PET/MRI, Eur. J. Nucl. Med. Mol. Imaging 46 (2019) 2746–2759.
[45] Ö. Çiçek, et al., 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 424-432.
[46] E. Gibson, et al., NiftyNet: a deep-learning platform for medical imaging, Comput. Methods Programs Biomed. 158 (2018) 113–122.
[47] H. Arabi, et al., Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region, Med. Phys. 45 (2018) 5218–5233.
[48] S. Kaur, et al., Diagnosis of Parkinson's disease using principle component analysis and deep learning, J. Med. Imaging Health Informatics 9 (2019) 602–609.