Peer Review File

Article information: https://dx.doi.org/10.21037/jtd-23-470

Review comments-Reviewer A

Artificial intelligence methods for localization and tracking cardiac structures, especially in disease states, are fascinating. This will not only help with interventional procedures, but will also be of great value for diagnostic imaging and observation of the condition. We applaud it.

Comment 1: The presentation of the data set in lines 138-141 should be more comprehensive.
Reply 1: Thank you very much for your valuable and insightful comment. We have described the local dataset and the two public datasets in detail regarding data type, data volume, types of diseases included, and ethical review in the dataset. This will provide a better picture of the data used in the study. This is very necessary.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 5, lines 148 to 190).

Comment 2: A discussion of the value of AI-assisted ultrasound image interpretation should be included.
Reply 2: Thank you very much for asking this question. Artificial intelligence-assisted ultrasound image interpretation can help reduce the technical difficulty of ultrasound-guided interventional procedures while shortening the training period for the medical staff involved. We have included a more detailed discussion in the paper.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 11, lines 338 to 347).

Comment 3: The latest applications of ultrasound-guided interventional techniques should be presented.
Reply 3: Thank you for your thoughtful suggestion. It is essential to present the latest applications of ultrasound-guided interventional techniques, which will further highlight the significance of our work. And it will help to deepen the reader's understanding of this research. We will present it in more detail in the paper.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 4, lines 103 to 117).

Comment 4: The discussion should introduce the disadvantages of AI-assisted medicine.
Reply 4: Thank you very much for asking this relevant question. As with various AI-assisted medical research and applications, AI can assist human experts but not replace them, which is ethically impermissible. Ultimately, the quality of medical care still needs to be controlled by human experts.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 13, lines 404 to 407).

Review comments-Reviewer B

Performing structural cardiac interventions under ultrasound guidance avoids the risk of radiation exposure to patients and healthcare professionals during traditional radiation-guided interventions. This is a great innovation that is very appealing. The use of artificial intelligence to promote this technology will be of great interest. There are a few more issues that need to be clarified before publication.

Comment 1: What are the disadvantages of ultrasound-guided surgery?
Reply 1: Thank you very much for asking this question. The main disadvantage of ultrasound-guided surgery is the high demand on the ability of the surgeon and sonographer to quickly and accurately determine the heart's structure under ultrasound. At the same time, the operator is required to reconstruct the 3D structural model in his mind quickly. All of these have limited the diffusion of this technology.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 4, lines 118 to 127).

Comment 2: What is the innovation of the artificial intelligence model for heart structure recognition?
Reply 2: Thank you very much for asking this relevant question. The model used in the study incorporates a spatial attention module for calculating spatial attention and a channel attention module for calculating channel attention in response to the fact that ultrasound images tend to detect a small percentage of the entire frame. The combination of the two dramatically improves the model's effectiveness and is the main innovation of this artificial intelligence model.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 7, lines 193 to 212).

Comment 3: It is recommended that the 15 experts involved in the human-machine comparison validation be categorized and described.
Reply 3: Thank you very much for your valuable and insightful comment. We have described the units, work experience, and sections of all 15 experts. This is very important.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 7, lines 215 to 221).

Comment 4: Is the model code and data used in this study open source?
Reply 4: Thank you very much for asking this question. The published data sets EchoNet-Dynamic and CAMUS involved in the study are available from the original authors upon request. The artificial intelligence model code, ultrasound images, and video data used in the can be obtained by contacting the corresponding author (email: panxiangbin@fuwaihospital.org).
Changes in the text: We have added content regarding these points in the revised manuscript

(Manuscript file, page 14, lines 446 to 449).

Review comments-Reviewer C

The authors propose a deep learning method to recognise nine cardiac structures from echocardiography. The neural network is trained on data including sick patients and externally validated on the CAMUS and Echonet datasets.

The manuscript is difficult to follow and unclear (data, experiments, results and terminology). From what I understood, I found two concrete flaws in the work, one related to the training and one related to the evaluation on external datasets. In addition to that, there are many unprecise points that gives the impression that the authors lack understanding of the clinical, technical and statistical aspects of the topic. This is to my opinion not acceptable for a manuscript with 14 co-authors that are meant to have contributed to the design of the study, analysis, interpretation of the results and proof reading.
Reply: Thank you very much for asking this question. Your guidance is significant and we have revised it within the full text based on your guidance. Also, the author team contains several very prominent researchers who can guarantee the quality of the paper, and our explanations and additions can clear your doubts.
Changes in the text: The full text.

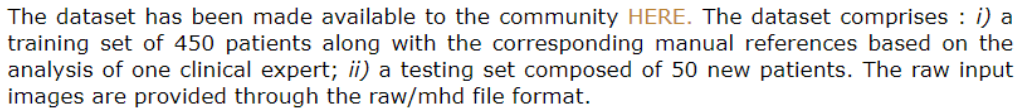
Below some more detailed comments on the manuscript:
The study has two main flaws:

l.153: “Forty views were randomly selected” to form the test dataset. The consequence of this is that data from the same patient can consequently be found in both the training and testing set. The rule number 1 of machine learning for medical image analysis is to split into training/validation/test at the patient level. If it is the case that the splitting is not done at the patient level, the results on the Fuwai hospital dataset are void.
Reply: Thank you very much for asking this question. Because ultrasound-guided intervention for structural heart disease is a very new technology, we could not collect enough data to construct an independent external validation dataset. This point we have added in the discussion section. However, this model performs equally well on public datasets, which proves that this model has good adaptability to ease your concerns. It is also acceptable for us to use internal validation (i.e., data from the same patient may appear in both the training and validation sets). For example, the internationally renowned dataset "CheXpert," published by Stanford University, consists of 224,316 chest films from 65,240 patients. (https://stanfordmlgroup.github.io/competitions/chexpert/) It is clear that one patient corresponds to multiple chest slices. However, this does not affect the use of internal validation methods for AI studies. Such studies are also published in top journals like Nature Medicine (DOI: 10.1038/s41591-021-01595-0). I think this is enough to put your doubts to rest.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 12, lines 400 to 404).

L.227: “The randomly selected views in both the datasets after application of the model were judged”: The judgment is basically not blinded if you provide to the physicians the output of the model. For the present study, I t would have been possible to do a blinded analysis by 1) manually annotating the datasets and 2) comparing with the yolo results. However, the authors do a non-blinded analysis and do not mention it in the limitations of the study. I consequently consider this as a major flaw which makes all results on the CAMUS and Echonet datasets void.
Reply: Thank you very much for asking this question. Since the external dataset does not provide a "gold standard" for heart structure annotation, the human expert panel's interpretation is the "gold standard" for our study. Due to a large number of external datasets, it is almost impossible to annotate them manually in advance. And there is no gold standard to compare with Yolo's results. Therefore, we had to ask a panel of additional human experts to evaluate the model's results. We have added this point to the discussion section. In fact, a similar approach can be found in top journal papers, such as "Dermatologist-level classification of skin cancer with deep neural networks".（DOI: 10.1038/nature21056）
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 12, lines 407 to 411).

It is also difficult to understand the data the authors use in the work.

Fuwai hospital: 3856 “views”. To my understanding, 80% are used for training and 20% for validation (l.176). Additionally, 40 “views” (approx 1%) are used for the test (l.153). This leads to a total of 101%.
CAMUS: The CAMUS dataset has 500 (not 450 as you write l.168) patients, and each of them have one Apical Four Chambers and one Apical Two Chambers standard view recording. It is unclear which standard view is used, I assume only the Apical Four Chambers. My understanding is that you used 10 frames to do the external test on this dataset.
Echonet: I understand that you use 200 frames to do the external test on this dataset.
Reply: Thank you very much for asking this question. Sorry, we didn't explain it clearly. We first randomly selected 40 images as the test set, and the rest of the data were divided into training and validation sets according to 8:2. We have provided additional explanations in the paper. Secondly, the CAMUS dataset has 500 cases of data. Still, the information related to 50 cases of the test set is not comprehensive, so we only used 450 cases of the training set data. This can be illustrated in the following figure on the official website. In addition, the section used in the study was the four-chambered heart section. We have added the above to the text and are very grateful.（https://www.creatis.insa-lyon.fr/Challenge/camus/index.html）
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 6, lines 192 to 193; page 6, lines 186 to 187).


The dataset has been made available to the community HERE. The dataset comprises : *i)* a training set of 450 patients along with the corresponding manual references based on the analysis of one clinical expert; *ii)* a testing set composed of 50 new patients. The raw input images are provided through the raw/mhd file format.

A figure representing the data available and the data used for training, validation, internal testing and external testing would definitely benefit the manuscript.
Reply: Thank you for your thoughtful suggestion, but due to space limitations and the fact that the corresponding data are described in the text, it may not be possible to add such a chart for the time being.

Additionally multiple points are questionable:

l.46: “Cardiac structures and lesions”. The method detects cardiac structures, but not lesions. Lesions are even not mentioned in the Results section of the abstract.
Reply: Thank you very much for asking this question. Thank you for your reminder. In fact, we have a label referring to "lesions," which is reflected in line 173, "439 for atrial and ventricular septal defects (nidus)". Unfortunately, the performance of the model in detecting lesions is not sufficiently prominent to be highlighted in the abstract. However, we have described it in both the methods and results.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 6, lines 173).

L.56: Reporting results on training data is not of interest.
Reply: Thank you for your suggestion. We also reported the test set results.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 2, lines 56 to 64).

l.143: “views”: It is unclear what the authors refer to with “view”. In the echocardiography field, a view is related to the transducer position and the so called “standard views” (for example apical four chambers). In the manuscript, I understand view as an ultrasound recording, composed of multiple frames. This seems to correspond with an average of 49 recordings by patient (3856 views / 79 patients)
Reply: Thank you very much for asking this question. Sorry for the lack of clarity, but the term "views" in the text refers to static ultrasound views, which may have caused some misunderstanding. We have added clarification as well.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 5, lines 158 to 159).

L.153: The authors select 40 out of 3856 “views” for the test data. This sounds definitely too small and not representative of the general population. The ROC curves and AUC values you obtain from this test set have consequently no value.
Reply: Thank you very much for asking this question. Thank you for your suggestion, but since the test set needs to take into account the possibility of human-machine comparisons with human experts, the dataset size needs to be designed to take into account the workload of highly qualified experts. But the model works so well that there is no need for additional human-machine comparisons. And there is no unified standard to specify the test set size. Therefore, this design is reasonable, and the results are meaningful.

L.167: To the best of my knowledge, the Echonet and CAMUS datasets do not have annotations of the heart structures. CAMUS has endocardium, epicardium and left-atrium tracing, whereas Echonet has endocardium tracing and Ejection Fraction values. Please do not write “labeled” when the labels are not related to the problem you aim to solve.
Reply: Thank you very much for asking this question. Thanks to your suggestion, we have completely rewritten this paragraph.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 6, lines 180 to 187).

L.223: “ratio of 20:1”. I assume this ratio comes from the number of Echonet recordings / the number of CAMUS recordings (10030/450 ~20/1). It is unclear which bias and which sampling method you write about. When one has the chance to have to large datasets for external evaluation and want to demonstrate the clinical value of the method, the way to go is to do the external testing on all the available data.
Reply: Thank you very much for asking this question. As you wisely described, we chose a sampling ratio of "20:1" based on the sample size ratio of the two datasets. However, due to the large size of the available dataset and our limited computing resources, random sampling was chosen to improve the feasibility of the study. Sampling methods are described in the statistical methods section.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 8, lines 262 to 265).

L.225: “27938 labels were detected”. This number should be related to the number of frames the neural network is applied on.
Reply: Thank you very much for asking this question. This number does relate to the number of frames applied to the neural network.

L.235: Is should be described how the accuracy is calculated.
Reply: Thank you very much for asking this question. We have added a description in the manuscript.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 8, lines 265 to 269).

L.248: Please do not report result on the training data. This gives the feeling that you do not know what you are doing.
Reply: Thank you very much for asking this question. We report the results for both the validation and test sets.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 9, lines 275 to 285).

L.255: You end up with AUC of 1 because the test dataset is too small.
Reply: Thank you very much for asking this question. Our test set may be small, but the model performs well in both the validation set and the external validation set, which may help allay

your concerns.

L.298: “t test show no significant difference”. This is not surprising as you have very few samples. The wording “no significant difference” is correct, but it does not imply equality.
Reply: Thank you very much for asking this question. We will consider your suggestions in our follow-up study. However, the current description of the results is correct.

L.330: “localising and tracking devices”: which devices? This is not mentioned earlier in the manuscript.
Reply: Thank you very much for asking this question. Sorry for the lack of clarity, but the term "devices" refers to the interventional instruments used in the procedure, also referred to as "sealing installations" in the methodology section. We have described this in the manuscript.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 11, lines 361; page 6, lines 174).

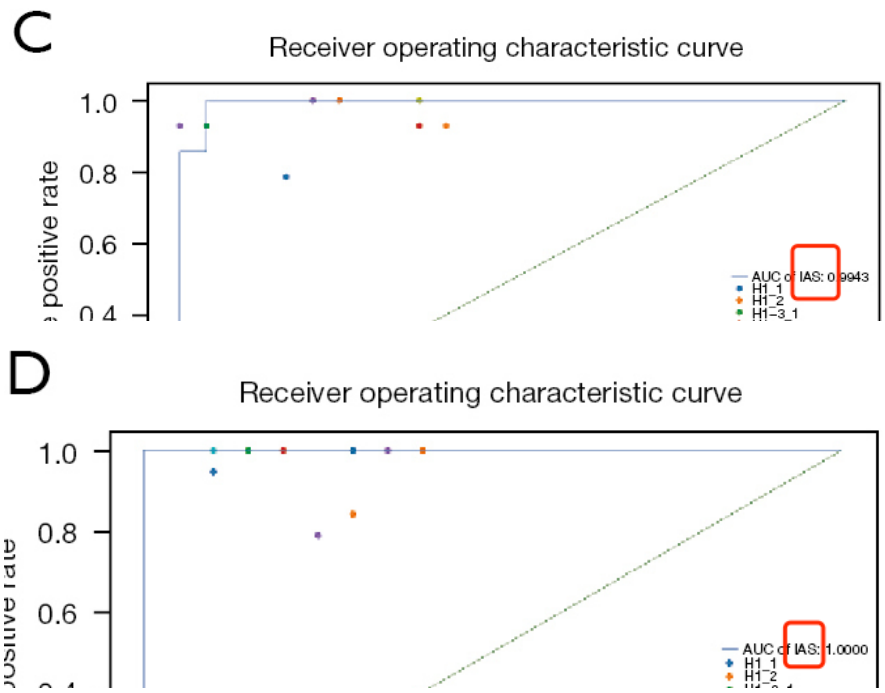
Finally, the authors propose a tailored neural network architecture that aims to maximize the accuracy for the task. However, no benchmark results are reported to demonstrate the technical value of using a tailored architecture. Further, the two main flaws mentioned above makes impossible to draw any conclusion on the clinical value of the tailored architecture.
Reply: Since this technology is so new that it is difficult for us to find references. But this innovation is also the driving force of scientific progress. Also, in the human-machine comparison section, we demonstrated the excellent performance of the model. We believe that this helps to demonstrate that our research will help to promote the diffusion of ultrasound-guided technology and improve the quality and efficiency of diagnostic ultrasound clinical workflow. We believe the shortcomings mentioned above can be remedied with your guidance.

Additionally, this reference could be useful: https://www.sciencedirect.com/science/article/pii/S2352914822002878
Reply: Thank you very much for your help. This literature is very informative.

Review comments-Reviewer D

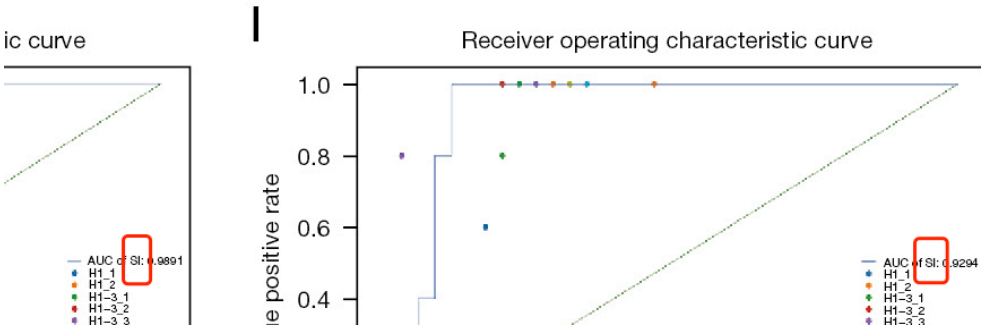
1. Figure 2:
a. Please define all abbreviations shown in figure 2 in the figure legends.
Reply: Thank you very much for your valuable and insightful comment. All abbreviations shown in Figure 2 are supplemented with definitions in the legend.
Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 18, lines 546-552).

2. Figure 3
a. Please check if figures C and D are correct.



Reply: Thank you for your guidance. Figure 3 needed to be corrected. We replaced the image in the manuscript and submitted a new one in the email attachment.

b. Please check if figures H and I are correct.



Reply: Thank you for your guidance. Figure 3 needed to be corrected. We replaced the image in the manuscript and submitted a new one in the email attachment.

3. Please define all abbreviations shown in Figure 5 in figure legends.

Reply: Thank you very much for your valuable and insightful comment. All abbreviations shown in Figure 5 are supplemented with definitions in the legend.

Changes in the text: We have added content regarding these points in the revised manuscript (Manuscript file, page 19, lines 569-571).

4. And it would be much appreciated if higher resolution Figure 5 would be resubmitted.

Reply: A higher resolution Figure 5 is attached to the email. The content of the image has remained the same. Please check it. Thank you.

5. Table 1: Check if table header should be completed.

Table 1 Accuracy of experts versus the AI model in identifying multiple cardiac structures and interventional devices

	AO (ACC)	AV (ACC)	IAS (ACC)	IVS (ACC)	MV (ACC)	Nidus (ACC)	PA (ACC)	SI (ACC)	TV (ACC)
H1									

Reply: Thank you for your thoughtful suggestion. The header of Table 1 was checked to be complete, and no additions were required.