Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-22-55

Reviewer A

Summary:

This paper presents a method for efficiently labeling volumetric images. The motivation is that while Cardiac MRI (CMR) generates large amounts of images, the lack of high-volume labeled data prevents the effective use of these images to train machine learning models. Labeling is usually provided by domain experts, which has a high labor cost. To address this challenge, this paper proposes a method for ground-truth labeling of CMR image data by leveraging the ranking of multiple images. It eliminates the need of labeling one image at a time and improves labeling efficiency. The proposed method is evaluated on a CMR dataset. The label agreement between the proposed method and a conventional labeling strategy is compared. The performance of machine learning models trained on these two labeling strategies is evaluated. However, before publication in this or other journals, the script would further be improved with some additional work.

Strength:

1, This paper is well-motivated. An efficient labeling strategy is important for providing labeled images for training machine learning models, and the conventional labeling strategy has low efficiency.

2, The paper is well-written and easy to follow.

3, The experimental results show superior accuracy of machine learning models trained with the proposed labeling strategy over the conventional strategy.

Weakness and expected modifications:

1, While the proposed multiple-image-ranking labeling strategy is more efficient than the conventional one-image-at-a-time, there was only moderate agreement between the labels from the two strategies as shown in Figure 4. It is not clear which labeling strategy is more close to gold standard labels.

Thank you for this interesting point. This is a situation in which although the *position* of each slice has a 'gold standard' ground truth (the relative position of each slice is known because of DICOM header information written in by the scanner), there is no agreed 'gold standard' for whether or not an individual slice has intersected the LV myocardium or not. This is because the definition of 'in the LV' is somewhat subjective and different clinicians will judge it differently compared to others. This is illustrated by the inter-observer variability for the ‘one-image-at-a-time’ strategy in our paper (Cohen’s kappa = 0.77).

The two labelling strategies have only moderate agreement (Cohen’s kappa = 0.67) because the ‘one-image-at-a-time’ strategy is based only on one person’s opinion, whereas the ‘multiple-image-ranking’ strategy is a composite of *three* people’s rankings. On one hand, the ‘one-image-at-a-time’ strategy should be regarded the ‘gold standard’ because it is the current albeit imperfect strategy (before our innovation). On the other hand, the ‘multiple-image-ranking’ strategy could be considered the ‘gold standard’ because it is a consensus opinion and results in more precise labels. We have chosen to regard the ‘one-image-at-a-time’ labels as the current ‘gold standard’, because the ‘multiple-image-ranking’ strategy is a novel approach that is being presented for the first time in this study. In future, perhaps the ‘multiple-image-ranking’ strategy could be regarded to be the ‘gold standard’ on the grounds of superior precision.

We have clarified this in the revised Methods section:

“Performance of both versions was evaluated on the same test set, using labels assigned by the one-image-at-a-time strategy. We regarded ‘one-at-a-time’ labels to be the definitive ground truth because the one-image-at-a-time strategy is the currently established labelling paradigm for this task.”

2, In Table 2, the model trained using labels provided by the proposed technique show better performance than using the conventional labeling strategy. However, since the performance metrics are evaluated against two different groups of labels, a higher accuracy does not necessarily represent better performance. To have a fair comparison, it could be better to evaluate the trained models by two labeling strategies on a public CMR dataset.

Thank you. We apologise for any inclarity regarding the test set. To clarify, even though the training and validation labels were different for each model version, the test set labels were *exactly the same* for both versions (see table below). In both cases, the test set labels came from the ‘one-image-at-a-time’ strategy, because we regard this to be the closest to ‘gold standard’, as explained in the response to Comment 1 above. We deliberately set up the analyses in this way so that the two models’ performances can be compared to each other. Your point about testing on an external public CMR dataset is very helpful, and we have added this to the revised Limitations section.

“Our future work will test the reproducibility of these results by evaluating the models’ performance on large, public CMR datasets.”

	Model trained using one-image-at-a-time strategy	Model trained using multiple-image-ranking strategy
Training set labels	One-image-at-a-time	Multiple-image ranking
Validation set labels	One-image-at-a-time	Multiple-image ranking
Test set labels	**One-image-at-a-time**	**One-image-at-a-time**

3, Some related works on label-efficient machine learning with volumetric medical images are missing from the discussion [1][2]. Please discuss them in the related works.

[1] Chaitanya, Krishna, et al. "Contrastive learning of global and local features for medical image segmentation with limited annotations." Advances in Neural Information Processing Systems 33 (2020): 12546-12558.

[2] Zeng, Dewen, et al. "Positional contrastive learning for volumetric medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.

Thank you for drawing our attention to this important work, which is now included to strengthen the Discussion section.

“A third approach is to use contrastive learning, a variant of self-supervised learning, based on the intuition that transformations of an image should have similar representations to each other and the original image, but dissimilar representations to different images. These approaches have shown some success to get the most algorithmic training value from limited clinical images (20,21).”

Reviewer B

Major concerns:

1. I have a concern about the labeling system of the “multiple image ranking”. Without seeing the context of the slide, will the rater be able to rate whether the slide is basal or apical relative to the LV? Also, the “ranking” and “rating” have scales of 0 to 3000 (or higher). This is a too large scale for a human-based

qualitative review. Normally, a qualitative review will have a scale of 5. When the range is 0~3000, it will be too hard to grade qualitatively. Can the authors clarify how were the raters reviewing those data?

Thank you. When considering whether a single slice is 'in the LV' or not, the rater did not need to see the other slices in the stack for context, because they are make a judgement based on how much of the left ventricular muscle they can see around the blood in the LV cavity in that specific image. This means that during the 'multiple-image-ranking' task, even though each image from the batch of 8 will typically come from different patients, this does not prevent raters from making a judgment about each individual image within the batch.

During the multiple-imaging-ranking tasks, the raters only saw images and were asked to drag them into order (most basal image toward the top left of the screen, most apical image toward the bottom right of the screen). They did *not* have to make any manual rating themselves. The 'ranking' and 'rating' were functions defined behind-the-scenes and not available to operators to ensure that each image had a unique spot in the ranking at any one point. Every time a rater submitted a batch, these variables updated in the band-end of the ranking algorithm. Users were not aware of this happening and they and never consciously ranked or rated these variables. The user-view is best shown in Figure 3, which shows our purposefully simple UI honed for high throughput and efficient use of clinician time.

We have clarified this in the revised Methods section:

"These variables were defined and updated in the back-end of the ranking algorithm, and were never shown to the raters. The only thing the raters saw were batches of 8 images which they could drag from most basal (top left of screen) to most apical (bottom right of screen), and a button to submit the current batch and load the next batch. (Figure 3)"

2. The authors use the labels from one-image-at-a-time strategy as the definitive ground truth for testing, which means they assume those are the most reliable labels. However, according to the Cohen's kappa test result (k=0.67), the proposed strategy has only moderate agreement with the ground truth. Does that mean the proposed labeling strategy is not accurate?

Thank you please see response to Reviewer A Point 1 above – in essence, there is no universally agreed 'ground truth' for this task because the definition of 'in the LV' is subjective and different clinicians will judge it differently compared to others. We chose to use the one-image-at-a-time strategy as the de facto ground truth in this study because it is the currently established way to label these data and assign discrete class labels. Since we are presenting a new way to do this (the "multiple-image-ranking" way), we can only compare its efficacy to the current paradigm (the "one-image-at-a-time" way). In future, perhaps the 'multiple-image-ranking' strategy could be regarded to be the 'gold standard' on the grounds of superior precision.

We have clarified this in the revised Methods section:

"Performance of both versions was evaluated on the same test set, using labels assigned by the one-image-at-a-time strategy. We regarded these labels to be the definitive ground truth because the one-image-at-a-time strategy is the currently established labelling paradigm for these type of data."

3. The authors should conduct cross-validation given the relatively small dataset. Besides, they should report confidence intervals of the metrics values to demonstrate the significance of their method.

Thank you for your suggestion. The validation method was by having a separate validation set which was used to evaluate the model's performance *during* training and adjust hyperparameters, and a held-out test set which was used to evaluate performance only after training was complete. We confirmed that both validation and training sets were representative of the class balance of the training set.

We now provide 95% confidence intervals for the metrics of recall, precision, accuracy and ROC AUC. These have been calculated by applying binominal probability functions. The only way to derive 95% CIs for F1-scores would be by bootstrapping. This would require repeating the experiment many more times than we did, in order to get robust estimates for the upper and lower bounds of the confidence interval. For the reason, we have not provided 95% CIs for F1 scores.

The CIs are available in revised Table 2 and revised Table 4

4. It is hard to believe that the model trained based on “multiple image ranking” labels can have better accuracy on the testing set than the model training with “one-image-at-a-time”, even the testing set’s ground truth is based on the “one-image-at-a-time” strategy. Can the authors explain it in the discussion?

Thank you for this interesting observation. We agree that this is a surprising finding, since one would expect the one-image-at-a-time strategy to succeed because the test set was labelled in exactly the same way as the training and validation sets. We hypothesise this is due to the reframing of a classification task into a regression task.

This has at least two advantages. First, a slice which is borderline basal is no longer forced into a dichotomy of acceptable versus too basal, with the model punished for getting it wrong - instead, the model is rewarded for placing the slice appropriately on the decision boundary. The influence of 'grey zone' labels and the problems it causes for classification tasks is a well recognised phenomenon, and indeed, our ranking strategy could be viewed of as a novel extension of categorical label smoothing: https://towardsdatascience.com/what-is-label-smoothing-108debd7ef06 - a strategy where the classification boundaries are blurred and performance improvements are significant, especially in datasets where there is controversy at the boundaries.

Second, our approach has in built robustness to labelling errors, whilst allowing collaborative labelling with pooling of experience. The multiple-ranking strategy relies on rapid repeated relative labelling - rare stochastic labelling errors, or unrepresentative expert opinions, have only a modest influence over the final rating of the result.

We propose these two phenomena may be responsible for the improved performance of using the multiple image ranking strategy for model training. However, further experiments will need to be performed in this area, and given this, we have chosen to leave out these unproven theories from the manuscript at this stage.

We have discussed these important points in the revised Discussion section:

“For the same human labelling time investment, the version trained on the labels from the 'multiple-image-ranking' strategy outperformed the 'one-image-at-a-time' strategy, demonstrating it is a more efficient way to harness human time. This is quite a surprising finding since the test set was labelled using the one-image-at-a-time strategy. We hypothesise this is because the multiple-image-ranking strategy reframes the labelling task from classification to regression. This has at least two potential advantages. First, a slice which is borderline between "too basal" and "in the LV" is no longer forced into a dichotomous label, with the model being *punished* for getting it wrong. Instead, the model is *rewarded* for placing the slice appropriately on the decision boundary. The influence of 'grey zone' labels and the problems it causes for classification tasks is a well recognised phenomenon. Indeed, our multiple-image-ranking strategy could be regarded to be a novel extension of categorical label smoothing (22). In this type of strategy, the classification boundaries are blurred and performance improvements are significant. This may be particularly useful for datasets where there is controversy at the boundaries between classes, such as LV slice level as illustrated by the inter-observer variability.

Second, our multiple-image-ranking strategy has in built robustness to labelling errors, whilst allowing collaborative labelling with pooling of experience. The multiple-image-ranking strategy relies on rapid repeated relative labelling, therefore rare stochastic labelling errors, or unrepresentative expert opinions, have only a modest influence over the final rating for a given image. We propose these two phenomena may be responsible for the improved performance of using the multiple-image-ranking strategy for model training. Further work is required to test and prove these hypotheses formally."

5. Again, using the comparison method as the ground truth of testing set is not reliable enough. In "one-image-at-a-time" strategy, each image will be only seen by one clinician, which will bring lots of bias from one person. Will the authors consider create a "golden" ground truth by using consensus from multiple experts?

Thank you. We agree about the relative unreliability of the one-image-at-a-time strategy. Yet despite this, it is the current commonest way to label these type of data in real-life clinical AI research. The single-rater-bias is a major downside of this approach and is one of the reasons that motivated us to develop better methods such as the multiple-image-ranking strategy.

However, since the multiple-image-ranking is the *novel* approach, and the one-image-at-a-time is the *current* approach, we have to use the latter to benchmark performance. Using different test from each approach to test each respective strategy would prevent the direct comparison that is required to be able to say which one is better. Furthermore, evaluating the multiple-image-ranking approach on a test set *also* labelled by the multiple-image-ranking approach would expectedly give good performance, and researchers would not know whether this new approach is actually better than the one they currently use. Our choice of setting the one-image-at-a-time labels as the 'ground truth' actually gives an advantage to the one-image-at-a-time method, and a disadvantage to the multiple-image-ranking method. As a result, our comparison is conservative and the performance of multiple image ranking has likely been underestimated.

We considered attempting to create a 'golden ground truth' by averaging multiple experts' labels from the one-image-a-time strategy and are grateful for the Reviewer's suggestion. We found the challenge was that it was:

(i) much more time-consuming than the multiple-image-ranking approach since multiple passes through the data would be required. This could confound any differences seen between the two approaches since one could claim that the only reason one performed better than the other was because more time was spent on that method. In our analyses we take time out of the equation by fixing the amount of time spent on both approaches to the same amount.

(ii) not representative of current research practice labelling more data once is prioritised over labelling fewer data multiple times.

(iii) it would create a pseudo-multiple-image-ranking approach by taking into account a consensus of experts.

The fact that you suggested this approach itself shows that there is a need for a labelling paradigm that captures multiple experts' opinions and we believe that this is why the multiple-image-ranking strategy is important to introduce in the literature.

Minor concern:

1. For the McNemar's test in the "Comparison of model versions" section, the authors should describe their hypothesis and what are the two objects they are comparing?

Thank you. We have clarified this in the revised Methods section:

"The predictions could therefore be evaluated as a binary classification ('in the LV' vs. 'not in the LV') enabling comparison with McNemar's tests (15). The null hypothesis was that there was no difference in the overall accuracy between the model trained using the 'one-image-at-a-time' strategy and the model trained using the 'multiple-image-ranking' strategy. A p-value <0.05 was considered significant."

2. For the "intra-rater and inter rater variability" section, what labeling strategy are they testing for agreement? The authors should provide more details.

Thank you. The labelling strategy that was tested for inter and intra-rater agreement was the one-image-at-a-time strategy:

"The test set (333 images, 10% of the dataset) was double-labeled by the same clinician and labeled by a second experienced clinician at least two weeks apart, in a blinded manner, using the one-image-at-a-time labelling strategy."