Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-24-1

Review Comments

This paper investigated different machine learning approaches to predict all-cause mortality from multimodal data, including SP-CMR images, LGE images and various clinical features. They implemented a range of popular CNN models and compared their performance when used alone or in conjunction with clinical features in an HNN model. Their results showed that the HNN model performed the best, but it was followed closely by a statistical model using just clinical variables.

This study is interesting and novel in several ways, such as using an NLP model to extract clinical data. However, the following points should be improved:

1. Methods:

a. Line 191-195 "For the purpose of this study, and NLP model was …… Elastic Search.": Could you please elaborate on this and make it more accessible to readers with a medical background? Also, please align your text explanation with what is shown in Figure 1 – for example, what is MedCAT? What are the unsupervised and supervised learning shown in Figure 1?

Reply a: Thank you. We have added a paragraph explaining the steps of clinical data extraction in more details, including NLP training with terms explanation (such as MedCAT).

Changes in the text: Page 7 / Paragraph 2 / line 20.

b. Use of CNN: since you have multiple images as input, how exactly was each CNN applied to the images? Were the images somehow combined before applying the CNN?

Reply b: Thank you. Images were combined as a stack, a paragraph explaining input shape has been added.

Changes in the text: Page 8 / Paragraph 3/ Line 22.

c. Line 238-245: Could you please provide some rationale for the way you calculated the initial bias and class weights? Also please improve the format of the formulae.

Reply c: Thank you. The initial bias has been explained and the formula improved. For class weights, equation has been updated to make it clearer.

Changes in the text: Page 9 / Paragraph 2 / Line 15.

d. Could you provide more details on how each of the clinical predictors (shown in Figure 3) were defined? For example, what counts as an "cerebrovascular accident (CVA)" given that it is so broad? Is it a yes/no variable or do you distinguish different types of CVA?

Reply d: Thank you. More details have been added to explain the clinical predictors input values in Statistical Analysis section.

Changes in the text: Page 10 / Paragraph 2 / Line 11.

e. Is your prediction outcome all-cause mortality at any time or by a fixed time? If it is at any time, then it might be inappropriate, as death after 1 month and after 3 years are very different outcomes in nature.
Reply e: Thank you. Prediction was calculated at a fixed time period calculated as explained in our Statistical Analysis section: “Follow up was calculated as the mean time to all-cause mortality event, and all cases without events and with shorter duration from CMR date to collection date were excluded.”. In the Results section, this value was reported as mean follow up is 1090 days.
Changes in the text: N/A.

f. Line 260: if you are comparing binary variables such as smoking across the 3 age groups, you should use Chi-square test instead of Kruskal Wallis test, as the latter is designed to work with ordinal variables.
Reply f: Thank you. We have re-calculated p values based on Chi-square and updated the text and table. Changes in the text: Page 10 / Paragraph 1 / Line 8.

2. Results:
a. In Figure 4, do the p-values correspond to the significance of difference between age groups or gender groups? The current way it’s labelled is confusing.
Reply a: Thank you. We have re-labelled the figure with the corresponding comparison for p-values.
Changes in the text: Figure 4 re-labelled.

b. What Figure 5 plots are the receiver-operating curves (ROC) and the precision-recall curves, instead of AUC or F1 scores, which are summary numbers from these plots.
Reply b: Thank you. The text has been amended.
Changes in the text: Page 11 / Paragraph 4 / Line 23.

c. Please also show in the result table all evaluation metrics including accuracy, precision and recall for the clinical model.
Reply 9: Thank you. Those metrics have been added as table 4.
Changes in the text: Table 4 added.

d. Could you also include results on the training and validation sets for the ML models perhaps in supplementary materials? This can help us examine the degree of overfitting, especially given that the AUC of most CNN models are below 0.7.
Reply d: Thank you. We have provided the results of training for best HNN (GoogleNet) and best CNN (AlexNet).
CNNs did not achieve clear convergence and there was overfitting as you predicted, we therefore added this to the text as well.
Changes in the text: Page 11 / Paragraph 3 / Line 16.

3. Discussion:
a. This section should be better structured. In the second and third paragraphs of your

discussion you were saying, slightly repeatedly, how non-invasive imaging features may help prediction, and why predicting mortality in CAD is important. However, these sound more like an introduction instead of discussion. In your discussion, you should highlight your important results and how you interpret them, but these were only briefly touched upon in the end. How your results compare to past literature was also not properly discussed.
Reply a: Thank you. The ‘Discussion’ section has been re-written.
Changes in the text: Page 12 / Paragraph 1-6 / Line 1. Page 13 / Paragraph 1 / Line 1.

b. Line 343: the statement of “this is the first application of using AI to link image pixels to prognosis” is simply not true. There have been many studies using imaging data and AI for disease prognosis.
Reply b: This is true for stress perfusion CMR, we added “stress perfusion CMR” to this statement. Changes in the text: Page 12 / Paragraph 5 / Line 21.

c. Line 357-362 “Although there was no statistical difference … better outcome prediction”: This has the risk of over-interpreting unsignificant results.
Reply c: Thank you. We have removed this sentence from the ‘Discussion’ section. Changes in the text: N/A.

d. Line 364-366 “current novel AI models…in the near future”: this comment is rather irrelevant and incorrect. Reinforcement learning is not unsupervised learning. Both are very different domains of ML from what is used in this research (supervised learning), and it is more meaningful to discuss your results instead of saying that AI as a whole might be useful. Besides, all prediction models are only useful if they can generalize well, so your point is unclear.
Reply d: Thank you. This issue has been addressed during re-writing the ‘Discussion’ section.
Changes in the text: Page 12 / Paragraph 6 / Line 26; Page 13 / Paragraph 1 / Line 1.

4. Conclusions:
a. Your results showed that the CNN models performed very poorly, and there was no significant difference between the performance of the clinical model and the HNN model. However, the second part of your conclusions were drawn highly favorably towards the ML models without strong evidence support. Moreover, your statement of “Further refinement of images … influence clinical decision-making in CAD” was not what you did in your study, so they should not be your conclusions.
Reply a: Thank you. The ‘Conclusion’ section has been re-written to address those issues.
Changes in the text: Page 13 / Paragraph 4 / Line 12.

5. Writing:
a. Caption of Figure 1: Should it be “NLP” instead of “LP” to stand for natural language processing? Where is this abbreviation used in the figure? Also, adding more text to

explain the process will be more helpful.
Reply a: Thank you. Sorry for typo, it is ‘NLP’. This term actually does not appear in the figure, so it has been removed. Further details on data extraction have been elaborated in the text.
Changes in the text: Page 7 / Paragraph 2 / line 20.

b. Caption of Table 1: Where is the abbreviation of HB (heart block) used?
Reply 17: Thank you. This is another error where the term has not been used, so we removed it.
Changes in the text: N/A.