Expanding from unilateral to bilateral: a robust deep learning-based approach for predicting radiographic osteoarthritis progression

Reviewer 1, Reviewer 2

**Main revision**

1. The body of the text (Introduction, Method, Results and Discussion) of a full-length manuscript is limited to 4,000 words maximum.
2. The abstract should be structured by objective, method, results, conclusion, with a size limit of 250 words.
3. Check the references to see that they comply with the journal reference style: e.g., six authors are to be listed prior to " et al.
4. There should be 2 files submitted: one file is a clean final copy, the second file has all changes marked.
5. Attached rebuttal: : list reviewer comments one by one, with each 'reviewer comment' followed by 'author response', and by 'author action' In 'author action' you describe exactly what change has been made, if any, referring to page/paragraph/line of the revised manuscript. Repeat this for each reviewer comment made by all reviewers.

Recommend:   
Also, you can now include interactive data visualizations in your publication and let your readers interact and engage more closely with your research.

Please follow the instructions here: <https://www.elsevier.com/authors/author-services/data-visualization> to find out about available data visualization options and how to include them with your article.

**Advantage:**

**R1**: develop and validate a novel bilateral knee neural network (BikNet). I believe the work done in this article to be novel and that it provides valuable contributions to the scientific literature. I

**R2**: The premise of the study is interesting and potentially impactful

**R2**: While the paper is generally well-written, it might benefit from a review of a native English speaker.

**Limitation:**

**R1**: Comment 1, Pages 5-6, Lines 101-102

"The participants were initially divided into a development set (from B, C, or D) and a testing set 1 (from 102 A or E) based on the enrolled site."  
Why divide the subjects based on enrollment site?

Demographic differences between sites could influence your analysis.

Answer:不确定我们是怎么做的 这里统计上应该要按比抽取, 比如每个site抽20%作为test set.

**R1**: Comment 2, Page 6, Lines 104-105

"To further evaluate the models' robustness and mimic clinical scenarios, testing set 2 (n=2,653) was created by obtaining 4-year follow-up radiographs."  
Why was the 48 month follow-up data not used for training as well? It seems that your model could perform even better with a larger training set at more time points.  
I would have liked to see a more typical randomization and splitting of the data into the training, validation, and test sets using the data from all sites and both time points.

Answer: 这个他不合理但是我们要做。首先我们就按他的做，合并这个问题和第一个问题（Comment 1）去做数据集的分割，然后在更多时间点上训练。然后我们解释当然有48m的数据训练会更好，但是大部分的人并不会来同一个地方复诊，或者是这种数据很难拿，所以我们从实际data可获得的角度认为这种设定不合理。

**R1**: Comment 3

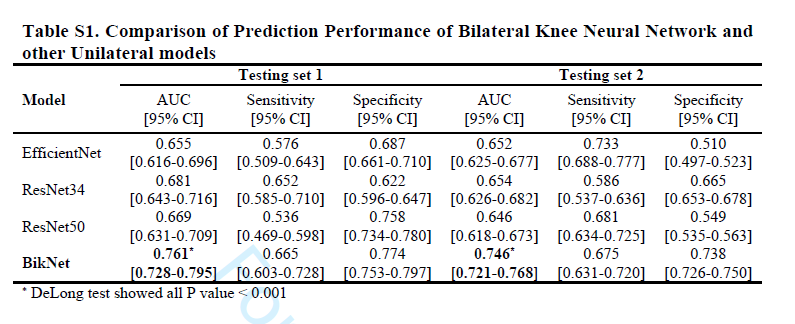
The comparison between the unilateral convolutional networks and the new bilateral cross-attention network seems a little biased. Why not train bilateral versions of the DenseNet and ResNext models? That would make for a much more fair comparison. Otherwise how would you know that that model architecture is the reason for the better performance, which seems to be the claim, and not just the presence of more information from the bilateral images?  
  
Answer: 我们要train bilateral的其他模型，给他结果就行。先说他的意见很中肯，然后说我们做了这个实验，但是篇幅没有放入文章，我们会在revised的版本加入。

**R1**: Comment 4  
Figure 4-D seems to be missing a legend. I would suggest just using a filled and unfilled shape for Figure 3-C to make it easier to tell what is being shown in 4-D.  
Answer: Figure 4-D确实没图例，这个标一下。然后3-C用和4-D一样的方法去作图，从一个unfill的形状到一个filled的

**R1**: Comment 5, Page 12, Lines 239-240  
"Figure 5C illustrates examples of prediction errors caused by poor image quality and obscured bony structures."  
For Figure 5-C, I don't know that here is any justification to claim that the errors are due to poor image quality or obstruction compared to Figure 5-B. Why couldn't it be the case that the model just isn't robust enough to accurately handle these images?  
Answer: 问到点了，从我们的角度解释就是我们先发现这个例子错了，然后我们再找成因发现CAM的效果不好说明模型没关注到有用的信息，但是我这个回答不行。你可能可以从医生的角度说这个人也看不出来？

**R1**: Comment 6, Page 13, Lines 270-275  
"To improve the model's performance, we designed our model to simulate a clinical diagnosis process, where a doctor first identifies anatomical landmarks, assesses joint space narrowing, and measures the knee alignment to predict the potential risk of OA progression. Our method utilized multitask learning to design auxiliary tasks that explicitly predict the OA diagnosis and landmarks to mimic the abovementioned progress."  
None of this is analyzed in the above evaluations are analyzed in this paper. I don't think it's appropriate to discuss the value of these steps in improving the model when there is no comparison to a model without these steps.

Answer: 需要比较model without these steps，即我们去掉multi-task的ablation



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | T1 AUC | T1 Sensitive | T1 Specificity | T2 AUC | T2 Sensitive | T2 Specificity |
| ConvNeXt, unilateral | 0.627 | 0.589 | 0.653 | 0.618 | 0.601 | 0.630 |
| -> bilateral | 0.610 | 0.577 | 0.642 | 0.596 | 0.587 | 0.614 |
| + cross attention | 0.725 | 0.629 | 0.726 | 0.706 | 0.599 | 0.713 |
| + multi task (The same as below | 0.761 | 0.665 | 0.774 | 0.746 | 0.675 | 0.738 |
| BikNet (ConvNext) | 0.761 | 0.665 | 0.774 | 0.746 | 0.675 | 0.738 |
| BikNet (ResNet34) | 0.738 | 0.649 | 0.691 | 0.714 | 0.633 | 0.720 |

Cuz ResNet34 shows better performance than ConvNeXt in unilateral form, we should also justify why we use ConvNeXt.

a unilateral model to directly quantify the added benefit of using the contralateral knee?

**R2**: Comment 1,

The definitions of the testing sets are not clear.  As I understand it, Testing Set #1 uses the BL images to predict the 48 month status.   Does Testing set #2 use the 48mo to predict 96?  If this is correct, then there is substantial overlap in the OAI participants between the training set and Testing Set #2, while there is no overlap with Testing Set #1.  Can the authors clarify this?  The second issue is lack of a truly valid comparison to a unilateral model.

Answer: 合并R1的第一二个问题一起详细解释Test Set

**R2**: Comment 2,

As the authors point out comparison to different studies is problematic.  Why don't the authors use the same data and a unilateral model to directly quantify the added benefit of using the contralateral knee?

Answer: 合并R2的第三个问题，然后说我们选哟一个高精度的model，所以我们也探索了不同backbone的效果。

**R2**: Comment 3, Line 118: Since the OAI radiographs have a wide range of pixels spacing, 700 x 700 pixels will cover vastly different anatomical areas.  Were the radiographs also rescaled to a constant spacing prior to cropping?  If so, what was it?  Also, can I assume each 700 x 700 ROI covers a single knee? i.e. 2 ROIs per radiograph.

Answer: 这个你比较清楚了，他的意思就是原始OAI很大，700x700应该只覆盖了很小的一部分。

**R2**: Comment 4, Line 123: The authors need to provide more details about the quality control step.  What was done and by whom?  Is this reflected in Figure 1?

Answer: 比这个我不怎么理解。是指每个人的工作的部分嘛。他好像关心的是数据怎么处理。

**R2**: Comment 5, Line 243: Was the process actually fully automated?  Was the cropping step 100% successful with no reader corrections necessary? Could the QC step be considered manual intervention?

Answer: 我们的数据处理好像是固定的。你可以说我们follow了谁谁谁，他是这么做的。因为是固定的没有case by case，所以没有intervention

**R2**: Comment 6,   
Line 307: It's with also mentioning that the OAI used a specialized positioning protocol and hardware frame.  These are not generally used in the clinic.  
  
Answer: 你可能需要解释一下这个

**R2**: Comment 7,

Figure 1. This is very confusing and I believe incorrect.  We know from the right hand side that 3,576 subjects had a 48mo follow-up.  Presumably these same subjects should be included for the left-hand side since a 48mo follow-up is necessary to train and test the models.  Yet, there are 3,585 left after all selections,more than 3,576.  For "Not OA or underwent TKR" do the authors mean that the TKR was before 48mo?  I don't understand "Not OA" since a large fraction of the OAI subjects (Incidence Cohort) did not have OA at BL by definition.  And of course there are the (> 100) subjects in the Control Cohort with no risks for OA.  It seems to me that the first selection criterion on the left-hand side should be drop everyone who did not have a 48mo radiograph.  In any case, this this figure needs clarification.

Answer: 需要加一个annotation和一些details。 我没仔细看。

**R2**: Comment 8,  
  
Figure 3. It might be helpful to also label the "Auxiliary View" as the contralateral knee.

Answer: OK