

Peer Review File

Learning Co-Plane Attention Across MRI Sequences for Diagnosing Twelve Types of Knee Abnormalities



Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

Review of NCOMMS-23-35552

De-COMA: Deep CO-plane and Multi-modal Attention Learning for Classifying Twelve Knee Abnormalities from Multi-sequence MRI

Summary:

In this study, a multi-center data set is constructed with images from 5 different clinical centers. This multi-center data set is then used to train a neural network to be able to detect 12 different knee abnormalities from multi-sequence MRI.

General comments:

This study provides an interesting new approach to a difficult and important problem in musculoskeletal imaging. The technical development is compelling and I like the probability matrix used, but I thought certain improvements of the study design and reporting of results could strengthen the manuscript.

Specific comments:

Abstract:

1. Lines 33-34: When people say 3D MRI, they usually mean 3D acquisitions, i.e., exciting a volume and phase encoding it, instead of exciting slices. I think when the authors say 3D MRI, they simply mean multiple slices. I would rephrase this.
2. Lines 33-34: The authors say "... different scanning modalities and planes...". I think instead of "modalities", the authors should say something like "contrast". By different imaging modalities, people usually mean MRI vs CT vs ultrasound etc. I would suggest changing this term throughout the manuscript.
3. Lines 44-45: I don't fully understand what the authors mean by "the plane-abnormality correlation discovery", can the authors rephrase?
4. Lines 48-49: The authors talk about how the network performed compared to the junior and senior radiologists but haven't previously mentioned in the abstract that such a comparison would be done. Could the authors add such a description for clarity?

Introduction:

5. Lines 63-64, left column: "Timely surgery is reacquired" – I assume the authors mean "required"?
6. Lines 54-66, right column: In this paragraph, the authors first describe the motivation being due to radiologists' lack of sensitivity and expertise, but end the paragraph talking about the need to reduce workload. To me, these are two separate motivations – do the authors think the network needs to be better than a radiologist, faster, or both?

1 Deep learning enables automatic knee abnormality classification

7. Lines 107-108, left column: I'm not sure what the authors mean by "traditional image analytical

methods with hand-crafted features” – can the authors rephrase?

2 Results:

8. Lines 130-133, right column: I would suggest moving interpretation of the results to the Discussion section.
9. Lines 154-177, left column: The class activation mapping is not mentioned in the Methods section, I would suggest adding a description there.
10. Lines 154-158, left column: Did the network correctly classify this individual to have MENI, ACL, CART and EFFU?
11. Lines 169-172, left column: Is there a way to see the regions that get activated for each of the abnormalities of this individual (MENI, ACL, CART, EFFU)?
12. Lines 190-194, left column: The analysis with radiologists is also not mentioned in the Method section, I think this methodology should be described there, with the results described here.
13. Lines 194-196, left column: Am I understanding correctly that the tests to compare with radiologists were done on the whole internal test data sets (220 subjects)? This could be made more clear in the Methods.
14. Lines 190-194, left column: How do you define “junior” and “senior” radiologist? How much experience did these radiologists have?
15. Lines 163-167, right column: You say that your model is competitive with the senior radiologist, but the ROC curves in the supplementary material (1.3.2) show the senior radiologist to be substantially better. I suggest the authors rephrase accordingly.
16. Lines 180-185, right column: I am unclear about where these likelihood numbers come from – can the authors explain a little better? Is this based on experimental or clinical data?
17. Lines 186-190, right column: Instead of calculating $|p - 0.5|$, wouldn't it simply be best to use the success rate the network has in detecting the various abnormalities in the different planes, similar to the numbers in Figure 6a?

3 Discussion

18. Lines 232-235, left column: I have the same comment as before regarding the senior radiologist, in the supplementary material, he seems to outperform the network so I would rephrase.
19. Lines 205-209, right column: The authors talk about T1 and T2 weighting, but you can have different amounts of T1 and T2 weighting – not all T1W and T2W sequences are exactly the same. Can the authors comment on what would happen if one used their network with PDW, T1W, and T2W sequences, but with somewhat different scan parameters (different TR, TE, flip angle, etc), so that the contrast wouldn't be quite the same?

4 Largest ever multi-center dataset with 12 knee abnormalities annotated from arthroscopy

20. Can the authors provide more information about the equipment used to obtain this data – scanner type, field strength, coils, etc?

5 Method:

21. Lines 268-270, right column: I don't fully understand what the authors mean by abnormalities happening in a small area around the meniscus. Are the authors essentially using the meniscus as a landmark for the whole knee joint?

22. Lines 276-278, right column: How large was the cropped area?
23. Lines 297-306, right column: In the example that the authors give, even though the rotated coronal patch would give high slice resolution, it only does this for a limited number of points (24 points along the A/P axis). Doesn't that mean that for most of the sampled points in the axial branch, you still don't have very high slice resolution?
24. Lines 346-349, left column: Can you please define C, D, H, and W here?

Tables and figures:

25. Figures 1-3: I commend the authors for creating these figures, which add a lot of clarity to the paper.

Supplementary material:

26. The authors describe data augmentation. How much was the data augmented (how many data sets did the authors have after augmentation)?

Reviewer #2 (Remarks to the Author):

Peer review comments:

- The quality of written English is low and there are too many awkward sentences/expressions with incorrect grammar. If the manuscript is written by a non-English native author(s), the manuscript should have undergone professional English screening/editing prior to submission.
- Many sentences in the Introduction of the paper are superfluous. Much of the first paragraph can be deleted. For example, young patients not wanting total knee arthroplasty is a matter of common sense. Also, 'accurate diagnosis of knee abnormalities' is always essential regardless of underlying pathology. It is not limited to the context of knee OA.
- 'Twelve knee abnormalities' – this is a vague terminology and it is unclear what these mean. What is the clinical value of making a deep learning model depict such abnormalities?
- The second paragraph of the page 2 of PDF file is also superfluous in contents. So many sentences are stating the obvious facts, e.g. 'MRI is needed due to the anatomical complexity and pathological diversity of the knee' and 'each MRI modality has a unique detection ability for specific types of knee abnormalities.' Also, there are a few incorrect pieces of information regarding basic principles of MRI. This raises a question of expertise level of the authors regarding MRI sequences. For example, PDW MRI does not focus on 'proton concentration', but it is a reflection of 'proton density' in a tissue.
- It is clear that the paper is written by non-radiologists, and the targeted readers are also not radiologists.
- In the third paragraph, the authors state "For radiologists, interpreting multiple MRI sequences with varying modalities and planes present a challenge." This is untrue. Any competent radiologists will not consider that to be a challenge, because MRI is always multiplanar and multisequence, and that's what it is. If anything, being multiplanar and multipsequence is helpful, i.e. opposite of a challenge. By the way, the authors made a mistake and stated "multimodality" in the aforementioned sentence. MRI is a single modality on its own.
- Making AI detect '12 abnormalities' in the generalized context is not clinically useful. I do not see any potential for this proposed algorithm to be applied in the clinical practice in the future.

- At present, typical AI algorithms only focus on one pathological feature at a time, because that's how we can train AI against ground truth.
- Different numbers of patients from 5 centers were included to make a cohort of total 1748 patients. Of note, the use of multiple MRI scanners at multiple sites introduces bias.
- Different numbers of patients are recruited from five centers, which is also a source of bias.
- How did authors account of bias in their dataset arising from differences in patient demography in each hospital?
- So, there were 773 training set, 110 for interval validation, and 220 for interval testing. How many patients were there for external validation?
- The way the authors selected 12 abnormalities is illogical. "Plica syndrome" is a clinical diagnosis which cannot be diagnosed based on MRI.
- They are assessing "meniscus tear" without differentiating medial meniscus and lateral meniscus, and they did not include meniscal root tears.
- The term "ligament injury" is too vague and one cannot train AI for such a vague entity. They should have been more specific and say either "ligamentous tear" or "ligamentous sprain", which appear differently on MRI.
- How can they arthroscopically diagnose bone contusion?
- Also, it does not make sense to diagnose patellar retinacular injury using arthroscopy.
- "Cysts" are too vague. Is it a subchondral cyst (=geode) or a cyst in the soft tissue like Baker's cyst, parameniscal cyst, or other ganglion cyst?

Reviewer #3 (Remarks to the Author):

Summary: The paper presents a deep learning-based framework for knee abnormality classification. The authors propose an attention-based mechanism to fuse information from multiple sequences and orientations. The work builds upon the MRNet framework (reference 23) by extending the evaluation to 5 centers, including 12 classes, and introducing a cross-modality, cross-plane attention-based approach to enhance model performance.

Overall: The paper's readability is satisfactory, but its novelty is limited. Several aspects of the paper lack clarity, and the experiments require more comprehensive details and analysis. As a result, the paper's strengths are not convincing. Placing the results section ahead of the explanation of experimental setups creates confusion, as readers lack essential context for understanding the outcomes.

Novelty: The proposed method appears to be an incremental extension and combinations of existing approaches -- attention-based methods have been explored in the field, and various methods of information fusion exist. The authors' proposal of an abnormality probability matrix for final output weighting is quite interesting, and its value is demonstrated in the supplementary material.

Experiments: Generally, the experiments are reasonable. The comparison with three baseline methods through AUC-ROC and ACC metrics is fine. However, the evaluation against radiologists could be much improved. Comparing with only one senior radiologist poses limitations, and inter-rater variability cannot

be assessed. Details about the junior radiologist group size and their expertise levels are lacking. While the AI-assisted radiologist performance comparison in supplementary materials is interesting, the provided information (is it classification label or class probabilities, etc?) isn't clear. The grouping of sites B and C, and D and E, without reporting individual site performance diminishes the evaluation's significance.

Structure:

- As mentioned, presenting the result before explaining the experimental set up was confusing to me.
- The paper lacks critical information:
 1. Details of the feature encoder are unclear.
 2. Fusion of anisotropic volumes is poorly explained. If 3D convolution extracts features, rotating patches might be redundant. Is the input 224x224x224 or 224x224x24? How are encoded features fused in the attention block?
 3. It's not clear how the consistency in training parameters across baseline methods was achieved, given varying architectures.
 4. Radiologist qualifications and expertise aren't specified.
 5. AI-assisted study details, including washout period and presented information, are missing.

Others:

1. The author mentions in Section 2.1 L129 and in Supplementary material that the single-task network have easier time classifying some classes, but this argument is not so strong. If this is the case, then the authors should investigate the model performance of the proposed approach for single class prediction tasks. This is crucial to evaluate the added value of multi-task approach.
2. The proposed approach struggles to generalize -- dropping from 0.812 AUC to 0.721/0.726 for external dataset. The author says this is "within an acceptable range". The authors must provide clinical justification for this comment. Also conduct additional experiments to compare the generalizability of different models (eg. relative drop in performance, etc.), to highlight the difference of your method compared to the baseline.
3. The comparison to clinical performance evaluation (wrt radiologists) should be performed with respect to external dataset.
4. Evaluating EigenCAM with radiologists would validate model correctness and interpretability. This should be studied to see if it can boost radiologists' study or reduction in time taken to read images.
5. Page 7 - line 307 not sure why only PDW was rotated. T1w and T2w can be too.

Typo:

- Section 5.5 L377

Recommendation: Please provide in depth information listed in Structure section, and perform additional experiments listed in Others 1-5 to strengthen the contribution of the paper.

Dear Editor, and Reviewers,

We sincerely appreciate your time and efforts in reviewing our manuscript. We have thoroughly examined your insightful comments and valuable suggestions, which have greatly contributed to improving the quality of the revised manuscript. Furthermore, we have improved the contents of the paper with more clarity and readability. Here we present the specific changes accordingly. Our response to your comments is marked in blue, and revisions made to the manuscript are highlighted in red.

Thank you once again for your invaluable contribution to this process.

Yours Sincerely,

Authors of the manuscript

Comments from reviewer #1 (Remarks to the Author):

Review of NCOMMS-23-35552

De-COMA: Deep CO-plane and Multi-modal Attention Learning for Classifying Twelve Knee Abnormalities from Multi-sequence MRI

Summary:

In this study, a multi-center data set is constructed with images from 5 different clinical centers. This multi-center data set is then used to train a neural network to be able to detect 12 different knee abnormalities from multi-sequence MRI.

General comments:

This study provides an interesting new approach to a difficult and important problem in musculoskeletal imaging. The technical development is compelling and I like the probability matrix used, but I thought certain improvements of the study design and reporting of results could strengthen the manuscript.

Specific comments:

Abstract:

1. Lines 33-34: When people say 3D MRI, they usually mean 3D acquisitions, i.e., exciting a volume and phase encoding it, instead of exciting slices. I think when the authors say 3D MRI, they simply mean multiple slices. I would rephrase this.
2. Lines 33-34: The authors say "... different scanning modalities and planes...". I think instead of "modalities", the authors should say something like "contrast". By different imaging modalities, people usually mean MRI vs CT vs ultrasound etc. I would suggest changing this term throughout the manuscript.
3. Lines 44-45: I don't fully understand what the authors mean by "the plane-abnormality correlation discovery", can the authors rephrase?
4. Lines 48-49: The authors talk about how the network performed compared to the junior and senior radiologists but haven't previously mentioned in the abstract that such a comparison would be done. Could the authors add such a description for clarity?

Introduction:

5. Lines 63-64, left column: "Timely surgery is reacquired" – I assume the authors mean "required"?
6. Lines 54-66, right column: In this paragraph, the authors first describe the motivation being due to radiologists' lack of sensitivity and expertise, but end the paragraph talking about the need to reduce workload. To me, these are two separate motivations – do the authors think the network needs to be better than a radiologist, faster, or both?

1 Deep learning enables automatic knee abnormality classification

7. Lines 107-108, left column: I'm not sure what the authors mean by "traditional image analytical methods with hand-crafted features" – can the authors rephrase?

2 Results:

8. Lines 130-133, right column: I would suggest moving interpretation of the results to the Discussion section.

9. Lines 154-177, left column: The class activation mapping is not mentioned in the Methods section, I would suggest adding a description there.

10. Lines 154-158, left column: Did the network correctly classify this individual to have MENI, ACL, CART and EFFU?

11. Lines 169-172, left column: Is there a way to see the regions that get activated for each of the abnormalities of this individual (MENI, ACL, CART, EFFU)?

12. Lines 190-194, left column: The analysis with radiologists is also not mentioned in the Method section, I think this methodology should be described there, with the results described here.

13. Lines 194-196, left column: Am I understanding correctly that the tests to compare with radiologists were done on the whole internal test data sets (220 subjects)? This could be made more clear in the Methods.

14. Lines 190-194, left column: How do you define "junior" and "senior" radiologist? How much experience did these radiologists have?

15. Lines 163-167, right column: You say that your model is competitive with the senior radiologist, but the ROC curves in the supplementary material (1.3.2) show the senior radiologist to be substantially better. I suggest the authors rephrase accordingly.

16. Lines 180-185, right column: I am unclear about where these likelihood numbers come from – can the authors explain a little better? Is this based on experimental or clinical data?

17. Lines 186-190, right column: Instead of calculating $|p - 0.5|$, wouldn't it simply be best to use the success rate the network has in detecting the various abnormalities in the different planes, similar to the numbers in Figure 6a?

3 Discussion

18. Lines 232-235, left column: I have the same comment as before regarding the senior radiologist, in the supplementary material, he seems to outperform the network so I would rephrase.

19. Lines 205-209, right column: The authors talk about T1 and T2 weighting, but you can have different amounts of T1 and T2 weighting – not all T1W and T2W sequences are exactly the same. Can the authors comment on what would happen if one used their network with PDW, T1W, and T2W sequences, but with somewhat different scan parameters (different TR, TE, flip angle, etc), so that the contrast wouldn't be quite the same?

4 Largest ever multi-center dataset with 12 knee abnormalities annotated from arthroscopy

20. Can the authors provide more information about the equipment used to obtain this data – scanner type, field strength, coils, etc?

5 Method:

21. Lines 268-270, right column: I don't fully understand what the authors mean by abnormalities happening in a small area around the meniscus. Are the authors essentially using the meniscus as a landmark for the whole knee joint?

22. Lines 276-278, right column: How large was the cropped area?

23. Lines 297-306, right column: In the example that the authors give, even though the rotated coronal patch would give high slice resolution, it only does this for a limited number of points (24 points along the A/P axis). Doesn't that mean that for most of the sampled points in the axial branch, you still don't have very high slice resolution?

24. Lines 346-349, left column: Can you please define C, D, H, and W here?

Tables and figures:

25. Figures 1-3: I commend the authors for creating these figures, which add a lot of clarity to the paper.

Supplementary material:

26. The authors describe data augmentation. How much was the data augmented (how many data sets did the authors have after augmentation)?

Our response:

Comment 1 and 2.

Lines 33-34: When people say 3D MRI, they usually mean 3D acquisitions, i.e., exciting a volume and phase encoding it, instead of exciting slices. I think when the authors say 3D MRI, they simply mean multiple slices. I would rephrase this.

Lines 33-34: The authors say "... different scanning modalities and planes...". I think instead of "modalities", the authors should say something like "contrast". By different imaging modalities, people usually mean MRI vs CT vs ultrasound etc. I would suggest changing this term throughout the manuscript.

Response: Thanks for your reminding. We have rephrase all the terms using in this paper. We agree that "modality" should not be used to address different MRI sequences and apologize for the confusion. The multi-sequence MRI is now being described as with different scanning plane or contrast. This step will enhance the overall clarity and presentation of the content.

Comment 3. Lines 44-45: I don't fully understand what the authors mean by "the plane-abnormality correlation discovery", can the authors rephrase?

Response: Thank you for your valuable feedback. "The plane-abnormality correlation discovery" refers to the correlation of MRI scanning planes and the abnormality classification result we found in the model output. We have made the changes in the abstract as follows: "Furthermore, we found

consistency with clinical knowledge when analyzing our model's results, which improves the interpretability of our model's results."

Comment 4. Lines 48-49: The authors talk about how the network performed compared to the junior and senior radiologists but haven't previously mentioned in the abstract that such a comparison would be done. Could the authors add such a description for clarity?

Response: Thanks for your advice. We have added the description of radiologists' evaluation in the abstract and enhance the overall clarity.

Comment 5. Lines 63-64, left column: "Timely surgery is reacquired" – I assume the authors mean "required"?

Response: Thanks for your comments. We have corrected the typos and will ensure that manuscripts are professionally screened and edited by native English speakers before resubmission.

Comment 6. Lines 54-66, right column: In this paragraph, the authors first describe the motivation being due to radiologists' lack of sensitivity and expertise, but end the paragraph talking about the need to reduce workload. To me, these are two separate motivations – do the authors think the network needs to be better than a radiologist, faster, or both?

Response: Thank you for your valuable feedback. In this study we aim to improve the accuracy of the radiologists. Our result has shown that the assistance of deep learning model can significantly raise diagnosis performance of the radiologists, especially for junior radiologists. We have modified the sentence as follows: "Therefore, automatic classification methods have a huge clinical demand for accurate diagnosis of knee abnormalities."

Comment 7. Lines 107-108, left column: I'm not sure what the authors mean by "traditional image analytical methods with hand-crafted features" – can the authors rephrase?

Response: Thanks for your comments. Here we mean the traditional image analysis methods that use hand-crafted features, to compare with the deep learning methods. We have rephrased the sentence as "Due to its complexity in spatial characteristics, traditional image analysis methods that use hand-crafted features are limited in designing effective morphological features."

Comment 8. Lines 130-133, right column: I would suggest moving interpretation of the results to the Discussion section.

Response: Thank you for your valuable suggestion. We have move the interpretation to the Discussion section and conduct more comprehensive analysis there to enhance the overall clarity.

Comment 9. Lines 154-177, left column: The class activation mapping is not mentioned in the Methods section, I would suggest adding a description there.

Response: Thanks for your valuable suggestion. We have move the class activation map analysis to the Discussion section. Considering that the CAM (Class Activation Map) using in this study is a straightforward implementation of existing technology and to maintain reading consistency, we just provide the description right before the analysis and also mention it in Sec. 3.6.

Comment 10. Lines 154-158, left column: Did the network correctly classify this individual to have MENI, ACL, CART and EFFU?

Response: Thanks for your reminding. Our model successfully classified the MENI, ACL, CART, and EFFU, but misclassified MCL and PLICA. Specially, with a false positive prediction on MCL. We have added the description as: "...to illustrate the activated regions in three planes on a subject with MENI, ACL, CART, and EFFU. CoPAS successfully classified the MENI, ACL, CART, and EFFU, but misclassified MCL and PLICA... it is worth noting that this leads to a false positive prediction on MCL"

Comment 11. Lines 169-172, left column: Is there a way to see the regions that get activated for each of the abnormalities of this individual (MENI, ACL, CART, EFFU)?

Response: Thanks for your suggestion. Since the EigenCAM [21] we used in this study is implemented in the way with no class discrimination, we add GradCAM++ [22] to illustrate the activated regions in three planes on a subject with MENI, ACL, CART. We illustrate and analyse the result in Sec2. Discussion and Fig.4.

Comment 12. Lines 190-194, left column: The analysis with radiologists is also not mentioned in the Method section, I think this methodology should be described there, with the results described here.

Response: Thank you for your valuable feedback. The analysis with radiologists is mentioned in the Sec. 2.1 Study design before result section to provide a concept for the readers in advance. We have made the revision on the paper structure to enhance the overall clarity, grammar, and presentation.

Comment 13. Lines 194-196, left column: Am I understanding correctly that the tests to compare with radiologists were done on the whole internal test data sets (220 subjects)? This could be made more clear in the Methods.

Response: Thank you for your valuable feedback. The comparison with radiologists are performed both on the internal test dataset (220 subjects) and selected external dataset (220 subjects with similar abnormality proportion to the internal set). The results are shown in the main manuscript Fig. 2 and the supplementary material Section 3.2. The experiment is mentioned in the Sec. 2.1 Study design in advance to improve the readability, and described with more details in supplementary material. We have revised the paper accordingly to enhance the overall clarity, grammar, and presentation of the content.

Comment 14. Lines 190-194, left column: How do you define "junior" and "senior" radiologist? How much experience did these radiologists have?

Response: Thanks for your reminding. The radiologists are defined as "junior" and "senior" according to the time of their clinical experience. The junior radiologists have less than 10 year experience and

the senior are with 10 and 30 years. We've added the [Sec. 3.7 Observer study](#) for more details and improved the paper clarity.

Comment 15 and 18. Lines 163-167, right column: You say that your model is competitive with the senior radiologist, but the ROC curves in the supplementary material (1.3.2) show the senior radiologist to be substantially better. I suggest the authors rephrase accordingly.

Lines 232-235, left column: I have the same comment as before regarding the senior radiologist, in the supplementary material, he seems to outperform the network so I would rephrase.

Response: Thank you for your feedback. Based on our analysis, we have observed that our model outperforms the senior radiologist in 5 out of the 12 abnormalities with the internal dataset, which demonstrates its effectiveness. Indeed, it is undeniable that the model's performance is slightly lower than that of senior radiologist 2 in the external dataset. We attribute this discrepancy to potential data shifting across different centers, as our model was trained solely on the internal dataset, whereas the senior radiologist may be less affected by this factor due to his extensive 30+ years of experience. Furthermore, we have observed significant performance improvement by the junior radiologist when assisted by the model, comparable to that of the senior radiologists. This observation further reinforces our belief that the model exhibits highly competitive performance compared to the senior radiologists.

Comment 16. Lines 180-185, right column: I am unclear about where these likelihood numbers come from – can the authors explain a little better? Is this based on experimental or clinical data?

Response: Thanks for your comments. These likelihood numbers were determined based on the prior clinical knowledge. Here we give some examples to illustrate clearly. It is recommended that meniscal tears be evaluated first on sagittal MR images and then on coronal images [1]. Because sagittal images can detect 97% of medial meniscal tears and 96% of lateral meniscal tears. Assessment of the anterior and posterior cruciate ligaments is typically performed on sagittal T2 images [2]. The medial and lateral collateral ligaments are best visualized on coronal images [3]. The above explanations have been added to the [Sec. 1.4](#) to improve the clarity.

Comment 17. Lines 186-190, right column: Instead of calculating $|p - 0.5|$, wouldn't it simply be best to use the success rate the network has in detecting the various abnormalities in the different planes, similar to the numbers in Figure 6a?

Response: Thanks for your valuable suggestion. Here we want to quantify the confidence of the modal prediction by calculating $|p - 0.5|$ because it is intuitively more similar to the human decision. The success rate is limited in a case that model yields false-positive results in all planes, but the certainty can still reflect the model decision preference. Nevertheless, the success rate is also a reasonable measurement, so we revised our paper by adding the prediction result into consideration and demonstrated in [Fig. 3](#), similar conclusion is drawn in [Sec. 1.4](#).

Comment 19 and 20. Lines 205-209, right column: The authors talk about T1 and T2 weighting, but you can have different amounts of T1 and T2 weighting – not all T1W and T2W sequences are exactly the same. Can the authors comment on what would happen if one used their network with PDW, T1W, and T2W sequences, but with somewhat different scan parameters (different TR, TE, flip angle, etc), so that the contrast wouldn't be quite the same?

Can the authors provide more information about the equipment used to obtain this data – scanner type, field strength, coils, etc?

Response: Thanks for your suggestions. We have added [Tab. 3](#) into the supplementary file to describe the MRI protocols used in this study, including repetition time (TR), echo time (TE), field of view (FOV), etc. To reduce the influence of un-unified image quality result from various machines and scanning parameters, we performed image pre-processing including resampling isotropic voxels, intensity standardization and knee area alignment. In fact, our cross-center study addressed this issue: while migrating to other centers may lead to a decrease in performance, our model still outperforms other methods and provides significant improvement for junior radiologists.

Comment 21. Lines 268-270, right column: I don't fully understand what the authors mean by abnormalities happening in a small area around the meniscus. Are the authors essentially using the meniscus as a landmark for the whole knee joint?

Response: Thanks for your comment. Sorry for the confusion brought by “a small area around the meniscus”, here we mean that locating meniscus can help the model to find the centre of the knee, and subsequently alleviate the misalignment between MRI scans. We use the meniscus as the landmark of knee and our ablation study demonstrates its effectiveness. For clarity, we modified the sentence in [Sec. 3.3](#): “Firstly, in order to establish a reference point for the knee, we segment the meniscus by U-Net [47] as an upstream task.”

Comment 22. Lines 276-278, right column: How large was the cropped area?

Response: Thanks for your reminding. The cropped area is 448×448 , with the slice number unchanged, and then interpolated into 224×224 . We have made changes to the corresponding content in supplementary as “The images were cropped in an area of 448×448 according to the ROI segmented from the meniscus (a rough segmentation performed by U-Net as an upstream task) before being interpolated into the shape of 224×224 .”

Comment 23. Lines 297-306, right column: In the example that the authors give, even though the rotated coronal patch would give high slice resolution, it only does this for a limited number of points (24 points along the A/P axis). Doesn't that mean that for most of the sampled points in the axial branch, you still don't have very high slice resolution?

Response: Thanks for your reminding. Your understanding is correct, since our images have large space between slices, the original information will only be stored in these slices. If we perform the interpolation, the resolution is higher in the slice plane. After the rotation, it will lead to higher information density along the corresponding axis (as shown in [Fig. 5](#)). And that is the reason we assign three branches (sagittal, coronal, and axial) in our model, to explore the dense information in different planes. We have modified the [Method](#) section accordingly to improve the clarity.

Comment 24. Lines 346-349, left column: Can you please define C, D, H, and W here?

Response: Thanks for your reminding. The C, D, H, and W represent the channel, depth, height, and width of the feature. We have added the sentence “That is, the feature map will be pooled and reshaped from $R^{C \times D \times H \times W}$ to $R^{D \times C}$, where C, D, H, and W represent the channel, depth, height, and width of the feature.” to improve clarity.

Comment 26. The authors describe data augmentation. How much was the data augmented (how many data sets did the authors have after augmentation)?

Response: Thanks for your reminding. The overall sample number in the dataset is 4 times before augmentation, so the subject number in the internal training set was 3092. We have added the sentence in supplementary “The overall sample number in the dataset is 4 times before augmentation.”.

Comments from reviewer #2 (Remarks to the Author):

Peer review comments:

- The quality of written English is low and there are too many awkward sentences/expressions with incorrect grammar. If the manuscript is written by a non-English native author(s), the manuscript should have undergone professional English screening/editing prior to submission.
- Many sentences in the Introduction of the paper are superfluous. Much of the first paragraph can be deleted. For example, young patients not wanting total knee arthroplasty is a matter of common sense. Also, 'accurate diagnosis of knee abnormalities' is always essential regardless of underlying pathology. It is not limited to the context of knee OA.
- 'Twelve knee abnormalities' – this is a vague terminology and it is unclear what these mean. What is the clinical value of making a deep learning model depict such abnormalities?
- The second paragraph of the page 2 of PDF file is also superfluous in contents. So many sentences are stating the obvious facts, e.g. 'MRI is needed due to the anatomical complexity and pathological diversity of the knee' and 'each MRI modality has a unique detection ability for specific types of knee abnormalities.' Also, there are a few incorrect pieces of information regarding basic principles of MRI. This raises a question of expertise level of the authors regarding MRI sequences. For example, PDW MRI does not focus on 'proton concentration', but it is a reflection of 'proton density' in a tissue.
- It is clear that the paper is written by non-radiologists, and the targeted readers are also not radiologists.
- In the third paragraph, the authors state "For radiologists, interpreting multiple MRI sequences with varying modalities and planes present a challenge." This is untrue. Any competent radiologists will not consider that to be a challenge, because MRI is always multiplanar and multisequence, and that's what it is. If anything, being multiplanar and multipsequence is helpful, i.e. opposite of a challenge. By the way, the authors made a mistake and stated "multimodality" in the aforementioned sentence. MRI is a single modality on its own.
- Making AI detect '12 abnormalities' in the generalized context is not clinically useful. I do not see any potential for this proposed algorithm to be applied in the clinical practice in the future.
- At present, typical AI algorithms only focuses on one pathological feature at a time, because that's how we can train AI against ground truth.
- Different numbers of patients from 5 centers were included to make a cohort of total 1748 patients. Of note, the use of multiple MRI scanners at multiple sites introduces bias.
- Different numbers of patients are recruited from five centers, which is also a source of bias.
- How did authors account of bias in their dataset arising from differences in patient demography in each hospital?
- So, there were 773 training set, 110 for interval validation, and 220 for interval testing. How many patients were there for external validation?
- The way the authors selected 12 abnormalities is illogical. "Plica syndrome" is a clinical diagnosis which cannot be diagnosed based on MRI.
- They are assessing "meniscus tear" without differentiating medial meniscus and lateral meniscus, and they did not include meniscal root tears.

- The term "ligament injury" is too vague and one cannot train AI for such a vague entity. They should have been more specific and say either "ligamentous tear" or "ligamentous sprain", which appear differently on MRI.
- How can they arthroscopically diagnose bone contusion?
- Also, it does not make sense to diagnose patellar retinacular injury using arthroscopy.
- "Cysts" are too vague. Is it a subchondral cyst (=geode) or a cyst in the soft tissue like Baker's cyst, parameniscal cyst, or other ganglion cyst?

Our response:

Comment 1. The quality of written English is low and there are too many awkward sentences/expressions with incorrect grammar. If the manuscript is written by a non-English native author(s), the manuscript should have undergone professional English screening/editing prior to submission.

Response: Thanks for your comments. We apologize for unpleasant reading experience. We have revised the paper thoroughly and ensure that manuscripts will be professionally screened and by native English speakers before resubmission. This step will enhance the overall clarity, grammar, and presentation of the content.

Comment 2. Many sentences in the Introduction of the paper are superfluous. Much of the first paragraph can be deleted. For example, young patients not wanting total knee arthroplasty is a matter of common sense. Also, 'accurate diagnosis of knee abnormalities' is always essential regardless of underlying pathology. It is not limited to the context of knee OA.

Response: Thanks for your advice. We have modified the content of the introduction, made a concise and clear statement to the background related to knee abnormalities. The new version of the introduction clearly states that our focus is on knee abnormalities caused by aging and injury, not limited to the context of osteoarthritis.

Comment 3. 'Twelve knee abnormalities' – this is a vague terminology and it is unclear what these mean. What is the clinical value of making a deep learning model depict such abnormalities?

Response: Thanks for your reminding. We specify 12 types of knee abnormalities in the new edition, including meniscal tears, cartilage damage, anterior cruciate ligament tear, posterior cruciate ligament injury, medial collateral ligament injury, lateral collateral ligament injury, patellar retinaculum injury, infrapatellar fat pad injury, synovial plica, cysts (such as popliteal cyst, ganglion cyst, meniscus cyst, and synovial cyst), bone contusion, and joint effusion. Different types of knee abnormalities often occur together but are best treated with quite different methods: severe abnormalities (e.g., anterior cruciate ligament tear [4] and meniscal tear [5]) require surgical treatment; and mild abnormalities (e.g., bone contusion, infrapatellar fat pad injury and effusion) usually require only conservative treatment and follow-up observation. The clinical value of the deep learning model for multiple knee joint abnormalities is that it can automatically and accurately diagnose knee abnormalities, which is crucial for formulating personalized treatment plans and improving the overall quality of life of patients. The modification is as follow: "Various knee abnormalities ..., including meniscal tears, cartilage damage, anterior cruciate ligament tear, posterior cruciate ligament injury, medial collateral ligament injury, lateral collateral ligament injury, patellar retinaculum injury, infrapatellar fat pad injury, synovial plica,

cysts (such as popliteal cyst, ganglion cyst, meniscus cyst, and synovial cyst), bone contusion, and joint effusion [6][7].”

Comment 4. The second paragraph of the page 2 of PDF file is also superfluous in contents. So many sentences are stating the obvious facts, e.g. ‘MRI is needed due to the anatomical complexity and pathological diversity of the knee’ and ‘each MRI modality has a unique detection ability for specific types of knee abnormalities.’ Also, there are a few incorrect pieces of information regarding basic principles of MRI. This raises a question of expertise level of the authors regarding MRI sequences. For example, PDW MRI does not focus on ‘proton concentration’, but it is a reflection of ‘proton density’ in a tissue.

Response: Thanks for your advice. We have deleted redundant content to accurately and clearly introduce the role of proton density weighted (PDW) MRI in the examination of knee joint abnormalities. We apologize for the incorrect expression of ‘proton concentration’, and corrected with ‘proton density’ for ‘PDW MRI’. We have revised the sentence as “In addition to conventional T1- and T2-weighted (T1W and T2W) imaging sequences, knee MRI requires proton density-weighted (PDW) imaging sequence because of its high signal-to-noise ratio and spatial resolution to accurately detect abnormalities [8][9].”.

Comment 5. It is clear that the paper is written by non-radiologists, and the targeted readers are also not radiologists.

Response: Thank you for your valuable feedback. We do this work with a team of experts in the field of engineering technology and radiologists. Our research focuses on engineering technology innovation. We have asked collaborating radiologists to carefully revise the radiology content to attract the interest of the target readers (including experts in the field of engineering technology and radiologists) and better meet the expectations of reviewers and readers. We believe that this kind of interdisciplinary research has an important role in promoting the development of the fields of engineering technology and radiology.

Comment 6. In the third paragraph, the authors state “For radiologists, interpreting multiple MRI sequences with varying modalities and planes present a challenge.” This is untrue. Any competent radiologists will not consider that to be a challenge, because MRI is always multiplanar and multisequence, and that’s what it is. If anything, being multiplanar and multipsequence is helpful, i.e. opposite of a challenge. By the way, the authors made a mistake and stated “multimodality” in the aforementioned sentence. MRI is a single modality on its own.

Response: Thank you for your suggestions. We acknowledge that the statement “For radiologists, interpreting multiple MRI sequences with varying modalities and planes presents a challenge.” is imprecise and apologize for the confusion. The word “challenge” has a vague meaning, and its usage here does not convey specific limitations. As mentioned in the comment, multi-sequence and multi-plane are characteristics of MRI, and the limitation of radiologists interpreting MRI is that this work is very time-consuming, tedious, and relies on the radiologist’s experience. We have corrected the previous misstatement as “However, knee MRI interpretation is very time-consuming and labor-intensive. Some subtle lesions of the knee joint can easily be overlooked by radiologists with insufficient work experience.”. In addition, we agree that “modality” should not being used to distinguish different MRI sequences and have carefully revised the terms in this paper.

Comment 7. Making AI detect '12 abnormalities' in the generalized context is not clinically useful. I do not see any potential for this proposed algorithm to be applied in the clinical practice in the future.

Response: Thank you for your comments. We selected 12 types of knee abnormalities according to expert consensus on MRI knee joint whole-organ assessment [10]. Different types of knee abnormalities often occur together but are treated with quite different methods. The automatic, accurate, and rapid diagnosis of knee abnormalities based on deep learning model is crucial for reducing radiologist's workload, developing appropriate treatment plans and improving prognosis for patients.

Comment 8. At present, typical AI algorithms only focuses on one pathological feature at a time, because that's how we can train AI against ground truth.

Response: Thank you for your valuable feedback. As mentioned in the comment, usually AI algorithms are built to detect a single disease, however, such algorithms cannot yet meet the needs of clinical practice well. Because imaging examination can provide information about a variety of diseases. As we have the ground truth for each abnormality, and previous works have proved that AI has the potential in detecting multiple diseases. For example, it is feasible to use algorithms to detect more than 100 lesions on a single chest radiograph [11]. We believe that building this algorithm to detect multiple diseases can more comprehensively assess the patient's physical health status and make more personalized treatment plans.

Comment 9. Different numbers of patients from 5 centers were included to make a cohort of total 1748 patients. Of note, the use of multiple MRI scanners at multiple sites introduces bias.

Response: Thank you for your advice. Collecting patients who underwent different MRI scanners at multiple sites does introduce bias. In fact, knee MRI scanning protocols vary widely between hospitals. We have added the details of MRI protocols in this study in the **supplementary Tab. 4**. However, our aim is to construct a model that is generalizable and applicable to different institutions. The point is that we collected data from multiple centers for external validation, which can well test the generalization of the model.

Comment 10. Different numbers of patients are recruited from five centers, which is also a source of bias.

Response: Thank you for your suggestions. The training dataset is collected from center A, which is a high-level orthopedic hospital, and the number of patients with knee joint abnormalities treated is much higher than that of other centers. For other centers, we also collected patients with the same inclusion and exclusion criteria and inclusion time. Most importantly, the data we collected reflect the distribution of patients with knee abnormalities in real world practice, which allow us to report more reliable result when evaluating the model across centers.

Comment 11. How did authors account of bias in their dataset arising from differences in patient demography in each hospital?

Response: Thank you for your reminding. We performed statistical analysis on patient demographic characteristics. The Kruskal–Wallis test was used to compare the age of patients in different centers, and the chi-square test was used to compare sex. Data were analyzed using SPSS software (version 23.0,

IBM, Armonk, NY). The results showed statistically significant differences in age and gender among patients at different centers (all $P < 0.001$). The result is reported in [supplementary Tab. 2](#). The dataset was divided into multiple subgroups based on gender and age, and we analysed the performance and reported the results in the [supplementary Sec. 3.3](#).

Comment 12. So, there were 773 training set, 110 for interval validation, and 220 for interval testing. How many patients were there for external validation?

Response: Thank you for your comments. For external validation, we included 500, 37, 64, and 44 patients in Center B, C, D, and E, respectively. We combined patients from centers B and C into external test set I and patients from centers D and E into external test set II to obtain two data sets with sufficient balance of positive samples. Details were shown in [supplementary Tab. 3](#).

Comment 13. The way the authors selected 12 abnormalities is illogical. "Plica syndrome" is a clinical diagnosis which cannot be diagnosed based on MRI.

Response: Thank you for your feedback. These 12 types of knee abnormalities usually arise from aging or injury, including meniscal tears, cartilage damage, anterior cruciate ligament tear, posterior cruciate ligament injury, medial collateral ligament injury, lateral collateral ligament injury, patellar retinaculum injury, infrapatellar fat pad injury, synovial plica, cysts (such as popliteal cyst, ganglion cyst, meniscus cyst, and synovial cyst), bone contusion, and joint effusion.[6][7] These knee abnormalities are relatively common and significantly affect patients' quality of life and cause a substantial social cost [12][13]. We apologize for the incorrect term "plica syndrome", which is diagnosed by clinician and not appropriate to use here. The target abnormality in our study is actually "synovial plica" that diagnosed via MRI [14]. We have revised the paper and carefully checked the terms to ensure clarity.

Comment 14. They are assessing "meniscus tear" without differentiating medial meniscus and lateral meniscus, and they did not include meniscal root tears.

Response: Thank you for your reminding. The meniscal tears we evaluated included tears that occurred in the medial meniscus, lateral meniscus, and meniscal root. One of the limitations of our study was that the meniscal tears were not more specifically localized. In fact, it is easy for radiologists to determine the location of a meniscal tear if they know it exists.

Comment 15. The term "ligament injury" is too vague and one cannot train AI for such a vague entity. They should have been more specific and say either "ligamentous tear" or "ligamentous sprain", which appear differently on MRI.

Response: Thank you for your comments. We agree that the term "ligament injury" involve a "ligamentous tear" or a "ligamentous sprain". However, we focus on anterior cruciate ligament tears because surgical treatment is often required [4]. There is no clear evidence to support the need for surgical treatment for "sprains" and "tears" of the posterior cruciate ligament, medial collateral ligament, and lateral collateral ligament [15-17]. Therefore, the latter three ligaments "sprain" and "tear" are uniformly labelled as "injury". We have added the detail descriptions of the abnormalities in this study in [Tab. 1](#).

Comment 16. How can they arthroscopically diagnose bone contusion?

Response: Thank you for your comment. Sorry for the inaccurate description of “we use arthroscopy as gold standard”. We only use arthroscopy as the gold standard when labelling meniscal tears, ligament injuries, and cartilage damage. In determining the label of bone contusion, consensus on MRI interpretation by two senior radiologists (with more than 10 years of experience in knee MRI interpretation) was used as the gold standard. We described each label in detail in the supplementary material of the new version of the manuscript.

Comment 17. Also, it does not make sense to diagnose patellar retinacular injury using arthroscopy.

Response: Thank you for your comment. In fact, the patellar retinaculum cannot be visualized using arthroscopy. The label of patellar retinacular injury was determined by consensus on MRI interpretation by two senior radiologists with more than 10 years of experience in knee MRI interpretation.

Comment 18. "Cysts" are too vague. Is it a subchondral cyst (=geode) or a cyst in the soft tissue like Baker's cyst, parameniscal cyst, or other ganglion cyst?

Response: Thank you for your valuable feedback. We clarify that in our research, the term "Cysts" encompasses cysts in soft tissues, including popliteal cysts, synovial cysts, meniscal cysts, and ganglion cysts. Subchondral cyst is usually caused by cartilage damage [18], and we included cartilage damage as one type of knee abnormalities in our research. We have revised the paper and added **Tab. 1** for describing the abnormalities in this study.

Comments from reviewer #3 (Remarks to the Author):

Summary: The paper presents a deep learning-based framework for knee abnormality classification. The authors propose an attention-based mechanism to fuse information from multiple sequences and orientations. The work builds upon the MRNet framework (reference 23) by extending the evaluation to 5 centers, including 12 classes, and introducing a cross-modality, cross-plane attention-based approach to enhance model performance.

Overall: The paper's readability is satisfactory, but its novelty is limited. Several aspects of the paper lack clarity, and the experiments require more comprehensive details and analysis. As a result, the paper's strengths are not convincing. Placing the results section ahead of the explanation of experimental setups creates confusion, as readers lack essential context for understanding the outcomes.

Novelty: The proposed method appears to be an incremental extension and combinations of existing approaches -- attention-based methods have been explored in the field, and various methods of information fusion exist. The authors' proposal of an abnormality probability matrix for final output weighting is quite interesting, and its value is demonstrated in the supplementary material.

Experiments: Generally, the experiments are reasonable. The comparison with three baseline methods through AUC-ROC and ACC metrics is fine. However, the evaluation against radiologists could be much improved. Comparing with only one senior radiologist poses limitations, and inter-rater variability cannot be assessed. Details about the junior radiologist group size and their expertise levels are lacking. While the AI-assisted radiologist performance comparison in supplementary materials is interesting, the provided information (is it classification label or class probabilities, etc?) isn't clear. The grouping of sites B and C, and D and E, without reporting individual site performance diminishes the evaluation's significance.

Structure:

- As mentioned, presenting the result before explaining the experimental set up was confusing to me.
- The paper lacks critical information:
 1. Details of the feature encoder are unclear.
 2. Fusion of anisotropic volumes is poorly explained. If 3D convolution extracts features, rotating patches might be redundant. Is the input 224x224x224 or 224x224x24? How are encoded features fused in the attention block?
 3. It's not clear how the consistency in training parameters across baseline methods was achieved, given varying architectures.
 4. Radiologist qualifications and expertise aren't specified.
 5. AI-assisted study details, including washout period and presented information, are missing.

Others:

1. The author mentions in Section 2.1 L129 and in Supplementary material that the single-task network have easier time classifying some classes, but this argument is not so strong. If this is the case, then the authors should investigate the model performance of the proposed approach for single class prediction tasks. This is crucial to evaluate the added value of multi-task approach.
2. The proposed approach struggles to generalize -- dropping from 0.812 AUC to 0.721/0.726 for external dataset. The author says this is "within an acceptable range". The authors must provide clinical justification for this comment. Also conduct additional experiments to compare the generalizability of different models (eg. relative drop in performance, etc.), to highlight the difference of your method compared to the baseline.
3. The comparison to clinical performance evaluation (wrt radiologists) should be performed with respect to external dataset.
4. Evaluating EigenCAM with radiologists would validate model correctness and interpretability. This should be studied to see if it can boost radiologists' study or reduction in time taken to read images.
5. Page 7 - line 307 not sure why only PDW was rotated. T1w and T2w can be too.

Typo:

- Section 5.5 L377

Recommendation: Please provide in depth information listed in Structure section, and perform additional experiments listed in Others 1-5 to strengthen the contribution of the paper.

Our response:

Comment 1. As mentioned, presenting the result before explaining the experimental set up was confusing to me.

Response: Thank you for your valuable feedback. We have added [Sec. 2.1 Study design](#) before result section to provide a concept of our work for the readers in advance. We have made the revision on the paper structure to enhance the overall clarity, grammar, and presentation.

Comment 2. Details of the feature encoder are unclear.

Response: Thanks for your comment. We use ResNet3D with 18 layers [20] as the feature encoder. We have modify the [Fig. 6](#) and [Method](#) section to improve the clarity.

Comment 3. Fusion of anisotropic volumes is poorly explained. If 3D convolution extracts features, rotating patches might be redundant. Is the input 224x224x224 or 224x224x24? How are encoded features fused in the attention block?

Response: Thank you for your valuable feedback. The rotation is to align the orientation of the volumes from different planes so that they can fit into the same encoder. The weight of encoder is shared from the main volume (e.g., the original coronal volume in coronal branch) so that each branch focuses on different plane, even with 3D convolution (because the original coronal volume only has high resolution

in coronal plane). The input of each branch is three volumes of 224x224x224, and output of the encoder is reshaped into 224x512 (moving the 'depth' dimension to the front), which can be regarded as 512-dimension embeddings for 224 slices. We are sorry for the confusion and have modified the **Method** section to improve the clarity.

Comment 4. It's not clear how the consistency in training parameters across baseline methods was achieved, given varying architectures.

Response: Thanks for your comment. We set the same hyper parameters (e.g., batch size, learning rate) across different methods. We keep the parameters within the models unchanged as their original papers. Consider the variant of different architectures, we performed some minor modifications to fit our tasks. For example, the MRNet [21] classified 3 abnormalities with 3 subnetworks, we can expand it to 12 abnormalities by duplicating to 12 subnetworks, and the parameters inside the subnetworks are unchanged. We have add the explanation of adaptation detail in the section **2.2 Adapting existing methods for comparison** in supplementary file to improve the clarity.

Comment 5. Radiologist qualifications and expertise aren't specified.

Response: Thanks for your reminding. In this study, the radiologists are defined as "junior" and "senior" according to the time of their clinical experience. The junior radiologists have less than 10 year experience, and the senior are with 10 and 30 years. We have added the **Sec. 4.6 Observer study** to improve the paper clarity.

Comment 6. AI-assisted study details, including washout period and presented information, are missing.

Response: Thanks for your reminding. In AI-assisted experiments, the radiologists fist perform diagnosis alone. Then we present the binary classification result from our model, re-order the data, and ask them to make diagnosis again. We have modified the **supplementary Sec. 3.2** accordingly.

Comment 7. The author mentions in Section 2.1 L129 and in Supplementary material that the single-task network have easier time classifying some classes, but this argument is not so strong. If this is the case, then the authors should investigate the model performance of the proposed approach for single class prediction tasks. This is crucial to evaluate the added value of multi-task approach.

Response: Thanks for your suggestions. We've evaluated the performance of our model under single task setting and added to the **supplementary Tab. 10**. Our study found the single-task network did show some advantage in certain tasks. However, the performance gain is trivial comparing to the large computational cost to ensemble 12 models, not to mention the overall performance is lower than multi-task model. We've revised the **supplementary Sec. 4 Ablation study**" accordingly to make a more comprehensive evaluation.

Comment 8. The proposed approach struggles to generalize -- dropping from 0.812 AUC to 0.721/0.726 for external dataset. The author says this is "within an acceptable range". The authors must provide clinical justification for this comment. Also conduct additional experiments to compare the

generalizability of different models (eg. relative drop in performance, etc.), to highlight the difference of your method compared to the baseline.

Response: Thank you for your comment. From our perspective, the decline in performance can be attributed to variations in image quality (such as different scanners and parameters) and the availability of different sequences (where only center A has 5 sequences). We have presented a comparison with different models in [supplementary Sec. 3.3](#) and [supplementary Tab. 6](#), which demonstrates that all models experience a performance drop, but our model still ranked in the top performance. However, the main objective of this study is to explore how deep learning can assist radiologists, and the results of our experiments on the external dataset have confirmed its effectiveness in improving the performance of junior radiologists. Sorry for the confusion of “within an acceptable range”, we have modified the [supplementary Sec 3.3](#) to improve the clarity.

Comment 9. The comparison to clinical performance evaluation (wrt radiologists) should be performed with respect to external dataset.

Response: Thanks for your suggestions. The comparison between our model and the radiologists on the external dataset is presented in [supplementary Fig. 1](#). In this figure, we illustrate the ROC curve of our model alongside the (FPR, TPR) point of the radiologists, which can easily tell our model have surpassed the junior radiologists. In addition, the figure also shows the performance gain of radiologists with the assistance of our model. We’ve re-drawn the figure to improve the readability and clarity.

Comment 10. Evaluating EigenCAM with radiologists would validate model correctness and interpretability. This should be studied to see if it can boost radiologists' study or reduction in time taken to read images.

Response: Thanks for your suggestion. This study aims to investigate how attention learning can enhance the accuracy of diagnosis in multi-sequence MRI. The comprehensive analysis of existing data has provided valuable insights, revealing a significant improvement in the performance of junior radiologists. While EigenCAM can aid in understanding the model, it has limitations when assisting radiologists. The low resolution of the map due to upsampling may highlight a large relative area, regardless of the presence of abnormalities. For instance, indicating the location of the meniscus to radiologists is redundant as they are already aware of it, and excessive guidance may even lead to misinterpretation. Nevertheless, the importance of further research on reducing radiologists' time is undeniable. We have addressed the limitations and proposed potential avenues for future investigations in the modified [Sec. 2 Discussion](#).

Comment 11. Page 7 - line 307 not sure why only PDW was rotated. T1w and T2w can be too.

Response: Thanks for your comment. Normally, we consider the PDW images to carry more informative details about abnormalities during the diagnostic process. On the other hand, T1W and T2W images provide a clear view of anatomical structures. To leverage the advantages of each image type, we rotate and fuse the PDW patches to complement each other in different planes. T1W and T2W images are served as the contrast reference in our model, enhancing the overall diagnostic capability. We’ve modified [Sec. 4.4](#) to explain the reason and improve clarity.

Comment 12. Typo: Section 5.5 L377

Response: Thanks for your reminding. We have corrected the typos and will ensure that manuscripts are professionally screened and edited before resubmission.

References

- [1] De Smet, Arthur A. "How I diagnose meniscal tears on knee MRI." *American Journal of Roentgenology* 199.3 (2012): 481-499.
- [2] Barberie, John E., et al. "Oblique sagittal view of the anterior cruciate ligament: comparison of coronal vs. axial planes as localizing sequences." *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 14.3 (2001): 203-206.
- [3] Khanna, A. Jay, et al. "Magnetic resonance imaging of the knee: Current techniques and spectrum of disease." *JBJS* 83.2 (2001): S128-141.
- [4] Musahl, Volker, and Jon Karlsson. "Anterior cruciate ligament tear." *New England Journal of Medicine* 380.24 (2019): 2341-2348.
- [5] Doral, Mahmut Nedim, et al. "Modern treatment of meniscal tears." *EFORT open reviews* 3.5 (2018): 260-268.
- [6] Roos, Ewa M., and L. Stefan Lohmander. "The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis." *Health and quality of life outcomes* 1.1 (2003): 1-8.
- [7] Kim-Wang, Sophia Y., et al. "Distribution of Bone Contusion Patterns in Acute Noncontact Anterior Cruciate Ligament-Torn Knees." *The American journal of sports medicine* 49.2 (2021): 404-409.
- [8] Tyler, Philippa, Abhijit Datir, and Asif Saifuddin. "Magnetic resonance imaging of anatomical variations in the knee: Part 2: Miscellaneous." *Skeletal radiology* 39 (2010): 1175-1186.
- [9] Chien, Alexander, et al. "Magnetic resonance imaging of the knee." *Polish journal of radiology* 85.1 (2020): 509-531.
- [10] Peterfy, C. G., et al. "MRI protocols for whole-organ assessment of the knee in osteoarthritis." *Osteoarthritis and cartilage* 14 (2006): 95-111.
- [11] Seah, Jarrel CY, et al. "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study." *The Lancet Digital Health* 3.8 (2021): e496-e506.
- [12] Adams, Bryan G., Megan N. Houston, and Kenneth L. Cameron. "The epidemiology of meniscus injury." *Sports medicine and arthroscopy review* 29.3 (2021): e24-e33.
- [13] Connelly, James W., et al. "Patient acceptable symptom state at 1 and 3 years after total knee arthroplasty: thresholds for the knee injury and osteoarthritis outcome score (KOOS)." *JBJS* 101.11 (2019): 995-1003.
- [14] García-Valtuille, Roberto, et al. "Anatomy and MR imaging appearances of synovial plicae of the knee." *Radiographics* 22.4 (2002): 775-784.
- [15] Bedi, Asheesh, Volker Musahl, and James B. Cowan. "Management of posterior cruciate ligament injuries: an evidence-based review." *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* 24.5 (2016): 277-289.

- [16] Miyamoto, Ryan G., Joseph A. Bosco, and Orrin H. Sherman. "Treatment of medial collateral ligament injuries." JAAOS-Journal of the American Academy of Orthopaedic Surgeons 17.3 (2009): 152-161.
- [17] Grawe, Brian, et al. "Lateral collateral ligament injury about the knee: anatomy, evaluation, and management." JAAOS-Journal of the American Academy of Orthopaedic Surgeons 26.6 (2018): e120-e127.
- [18] Gao, Liang, Magali Cucchiaroni, and Henning Madry. "Cyst formation in the subchondral bone following cartilage repair." Clinical and Translational Medicine 10.8 (2020): e248.
- [19] Tran, Du, et al. "A closer look at spatiotemporal convolutions for action recognition." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.
- [20] Bien, Nicholas, et al. "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet." PLoS medicine 15.11 (2018): e1002699.
- [21] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class activation map using principal components. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–7. IEEE, 2020.
- [22] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV), pages 839–847. IEEE, 2018.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Review of NCOMMS-23-35552A-Z

CoPAS: learning Co-Plane Attention across MRI Sequences for diagnosing twelve types of knee abnormalities: A multi-center retrospective study

Comments:

I thank the authors for answering my comments in their revised manuscript. I still have a few outstanding comments:

In your response to my comment #7, you say:

Thanks for your comments. Here we mean the traditional image analysis methods that use hand-crafted features, to compare with the deep learning methods. We have rephrased the sentence as “Due to its complexity in spatial characteristics, traditional image analysis methods that use handcrafted features are limited in designing effective morphological features.”

I’m sorry, I still don’t really understand what this sentence is trying to tell me. I don’t understand what is meant by “spatial characteristics”, “handcrafted features”, or “designing effective morphological features”. Can you please rephrase or just cut this sentence?

Regarding comment 26:

Thank you for clarifying. However, I suggest the authors phrase this differently to avoid confusion, to say something like “the data set was augmented by a factor of 4 to a total of 3092”, or something along those lines.

Reviewer #3 (Remarks to the Author):

Thank you for addressing my comments. The clarity of the manuscript has greatly improved. I have only minor comments:

1. Comment 7: Table 10: Items (vi) and (vii) show similar performances (AUC-ROC separated by 0.004). Is this difference statistically significant?
2. Comment 8: Supplementary Section 3.3 and Table 6 appear to refer to a subgroup study, and I do not see the comparison of model performances. Please add this experimental result.
3. Typo: Page 10, Line 461: "Sagittl"

Reviewer #3 (Remarks on code availability):

The code for the encoder, the proposed model, data loading, training, and some aspects of evaluation is available.

There are no tests, so it is hard to tell if this will work or not. Additionally, there is no code to reproduce the experimental results. In this regard, reproducibility may be limited.

Reviewer #4 (Remarks to the Author):

This work proposes a multi-purpose DL framework for knee abnormalities diagnosis. It utilizes attention structure to combine features extracted from MR images of different contrasts and orientations, and obtain satisfying diagnostic performances compared to that from radiologists. It seems the manuscript has been through a substantial revision following comments of the last round of review. I only have a few very minor comments left.

1. Page 5, left column, line 220-221, “a decline of...”, supposedly you meant the opposite.
2. Page 5, right column, line 236, what do you mean by “multi-center gap”?
3. Maybe with what you have accomplished, you can explore, in the future, training a model that describes what features it has observed in which region of the image and thus it “thinks” what injuries are presented.

Dear Editor, and Reviewers,

We sincerely appreciate your time and efforts in reviewing our manuscript. We have thoroughly examined your insightful comments and valuable suggestions and have improved the contents of the paper with more clarity and readability. Here we present the specific changes accordingly. Our response to your comments is marked in blue, and revisions made to the manuscript are highlighted in red.

Thank you once again for your invaluable contribution to this process.

Yours Sincerely,

Authors of the manuscript

Reviewer #1 (Remarks to the Author):

Review of NCOMMS-23-35552A-Z

CoPAS: learning Co-Plane Attention across MRI Sequences for diagnosing twelve types of knee abnormalities: A multi-center retrospective study

Comments:

I thank the authors for answering my comments in their revised manuscript. I still have a few outstanding comments:

In your response to my comment #7, you say:

Thanks for your comments. Here we mean the traditional image analysis methods that use hand-crafted features, to compare with the deep learning methods. We have rephrased the sentence as

“Due to its complexity in spatial characteristics, traditional image analysis methods that use handcrafted features are limited in designing effective morphological features.”

I’m sorry, I still don’t really understand what this sentence is trying to tell me. I don’t understand what is meant by “spatial characteristics”, “handcrafted features”, or “designing effective morphological features”. Can you please rephrase or just cut this sentence?

Regarding comment 26:

Thank you for clarifying. However, I suggest the authors phrase this differently to avoid confusion, to say something like “the data set was augmented by a factor of 4 to a total of 3092”, or something along those lines.

Our response:

Comment 1:

I’m sorry, I still don’t really understand what this sentence is trying to tell me. I don’t understand what is meant by “spatial characteristics”, “handcrafted features”, or “designing effective morphological features”. Can you please rephrase or just cut this sentence?

Response: Thank you for the suggestion. Here we mean the image features are manually designed in traditional image processing methods, such as the scale-invariant feature transform (SIFT) [1] algorithm. However, these methods are less effective when applied to images with unique spatial characteristics, such as 3D MR images. In contrast, deep learning can capture complex spatial information and outperform traditional methods in detecting morphological changes. To increase the clarity, we have revised the sentence as “Traditional image analysis methods that rely on hand-crafted features are less effective in handling complex 3D images.”

Comment 2:

Thank you for clarifying. However, I suggest the authors phrase this differently to avoid confusion, to say something like “the data set was augmented by a factor of 4 to a total of 3092”, or something along those lines.

Response: Thanks for your suggestion. We have revised the sentence in supplementary as “The dataset was augmented by a factor of 4, resulting in a total of 3092 samples.”.

Reviewer #3 (Remarks to the Author):

Thank you for addressing my comments. The clarity of the manuscript has greatly improved. I have only minor comments:

1. Comment 7: Table 10: Items (vi) and (vii) show similar performances (AUC-ROC separated by 0.004). Is this difference statistically significant?
2. Comment 8: Supplementary Section 3.3 and Table 6 appear to refer to a subgroup study, and I do not see the comparison of model performances. Please add this experimental result.
3. Typo: Page 10, Line 461: "Sagittl"

Reviewer #3 (Remarks on code availability):

The code for the encoder, the proposed model, data loading, training, and some aspects of evaluation is available.

There are no tests, so it is hard to tell if this will work or not. Additionally, there is no code to reproduce the experimental results. In this regard, reproducibility may be limited.

Our response:

Comment 1:

Comment 7: Table 10: Items (vi) and (vii) show similar performances (AUC-ROC separated by 0.004). Is this difference statistically significant?

Response: Thanks for your valuable feedback. Regarding Supplementary Table 10, items (vi) and (vii), statistical analysis reveals significant differences ($P < 0.05$, Delong's test) among 10 out of the 12 classes, with the exception being EFFU and PR. Despite a marginal difference of only 0.004 in average AUC-ROC values, the multi-task model (vi) consistently outperforms the single-task model (vii) across a majority of the tasks. We've included this sentence in the ablation study section to enhance clarity.

Comment 2:

Supplementary Section 3.3 and Table 6 appear to refer to a subgroup study, and I do not see the comparison of model performances. Please add this experimental result.

Response: Thanks for your valuable feedback. Sorry for the confusion. Supplementary Table 6 presents the subject-level performance, which is calculated using the testing dataset without dividing it into subgroups. However, in Supplementary Section 3.3, we conducted a subgroup analysis by dividing the participants based on gender and age, considering the patients' characteristics outlined in Supplementary Table 2. The results of these subgroup analyses are presented in Supplementary Figure 2, where the bar chart illustrates the model performance using the AUC-ROC metric.

Comment 3:

Typo: Page 10, Line 461: "Sagittl"

Response: Thanks for your reminding. We have corrected the spelling and will check our paper carefully.

Comment 4:

There are no tests, so it is hard to tell if this will work or not. Additionally, there is no code to reproduce the experimental results. In this regard, reproducibility may be limited.

Response: Thank you for the reminder. We have reorganized our code and made the necessary updates in the repository. Specially, we have added 50 test cases for evaluation. The experiments can be reproduced by passing the arguments and modifying the configuration files in our code. We have also updated the README file, which now provides clearer instructions.

Reviewer #4 (Remarks to the Author):

This work proposes a multi-purpose DL framework for knee abnormalities diagnosis. It utilizes attention structure to combine features extracted from MR images of different contrasts and orientations, and obtain satisfying diagnostic performances compared to that from radiologists. It seems the manuscript has been through a substantial revision following comments of the last round of review. I only have a few very minor comments left.

1. Page 5, left column, line 220-221, “a decline of...”, supposedly you meant the opposite.
2. Page 5, right column, line 236, what do you mean by “multi-center gap”?
3. Maybe with what you have accomplished, you can explore, in the future, training a model that describes what features it has observed in which region of the image and thus it “thinks” what injuries are presented.

Our response:

Comment 1:

Page 5, left column, line 220-221, “a decline of...”, supposedly you meant the opposite.

Response: Thank you for the reminder. We apologize for the incorrect description. Here, we meant to convey that the performance has declined from 0.812 to 0.712 and 0.726 when transitioning from the internal dataset to the two external datasets. We have revised the sentence as follows: “a decline of the average AUC-ROCs from 0.812 to 0.721 and 0.726 is observed when transitioning from the internal dataset to the two external datasets.”

Comment 2:

Page 5, right column, line 236, what do you mean by “multi-center gap”?

Response: Thank you for your valuable feedback. Here, we mean to express that the data collected from different centers exhibits variations in several factors, including scanning parameters, image quality, and cohort distribution. These differences contribute to the existence of a noticeable “domain gap” between the centers, and thus may affect the model performance. To improve clarity, we have revised the sentence as: “We found that, despite the model performance decline due to the data variations across multiple centers, CoPAS still outperforms certain senior radiologists.”

Comment 3:

Maybe with what you have accomplished, you can explore, in the future, training a model that describes what features it has observed in which region of the image and thus it “thinks” what injuries are presented.

Response: Thank you for your suggestion. We strongly agree this is a promising direction for future investigation. You can find a similar expression in our discussion section, “leveraging the CAM technique with a pre-trained model to generate a heatmap indicating the areas of potential abnormalities, we can significantly expedite the diagnosis process.”. Therefore, we will certainly consider in present a more trustworthy, explainable model for future research.

References

- [1] Lowe, David G. "Object recognition from local scale-invariant features." Proceedings of the seventh IEEE international conference on computer vision. Vol. 2. IEEE, 1999. Barberie, John E., et al. "Oblique sagittal view of the anterior cruciate ligament: comparison of coronal vs. axial planes as localizing sequences." Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 14.3 (2001): 203-206.

REVIEWERS' COMMENTS

Reviewer #3 (Remarks to the Author):

The authors addressed all of my concerns.

Reviewer #3 (Remarks on code availability):

The authors have added a test dataset to allow users to run the training and evaluation code. As such, while there is no trained model, it has minimal amount needed for a researcher to understand the code and run on their own dataset.