Response to reviewers

We truly appreciate all the constructive comments and suggestions from the reviewer and editor. We have modified the manuscript as per the reviewer's suggestion. We hope the manuscript will now be acceptable for publication. Please find below our point-to-point responses to the reviewers' comments.

Updated title: ***"An evaluation of AI models for osteoarthritis of the knee using deep learning algorithms for orthopaedic radiographs (DLAOR)."***

Reviewer #1: An ambitious work to validate machine learning approach for interpreting knee radiographs aiming detection of OA. But please keep in mind nowhere in the manuscript this statement appears, the reader is obliged to decipher him or herself what the work is about.

Response: We thank the reviewer for the appreciation and his comments. We acknowledge the lack of clarity in putting forth the use of ML in radiograph interpretation. The revised manuscript contains the description in abstract itself. It now reads:

"Deep learning, a form of artificial intelligence (AI), has shown promising results for interpreting radiographs. In order to develop this niche machine learning program of interpreting orthopaedic radiographs with accuracy a project named Deep Learning algorithm for Orthopaedic radiographs (DLAOR) was conceived. In the first phase, the diagnosis of Knee osteoarthritis as per the KL scale in medical images was done using the DLAOR."

Comments

- The abstract needs to be drastically shorten, please do not use numbered enumeration neither references within abstract, keep to essential focusing especially on the aim of the study (in the current form it is unclear what was the aim of the study) and on very briefly presenting results and conclusion Introduction OA is a degenerative disease of the whole joint not only of the cartilage. As to my knowledge references should be listed in the order of citation (why starting text with reference 42?).

  Response: Firstly, we appreciate your knowledge and understanding about this novel study to utilise the existing technology to benefit health care professionals. As per the suggestions, we have shortened the abstract, updated the objectives, and mentioned the possible drawbacks of the transfer learning approach. We apologise for the sloppiness in preparation of manuscript. The references have been now inserted as per the journals directive for authors.

- Please keep the information in the introduction to the point, what has driverless cars to do with this manuscript, this is supposed to be a scientific writing and not a blog or social media post.

  Response: We apologise for not being clear in the submission of introduction section. The changes have been made in the revised manuscript. I appreciate you highlighting the same, but the intention was to point out that image classification has been successfully applied in fields other than healthcare. Thank you for the suggestion, the necessary changes have been updated in the manuscript.

- Please clarify the paragraph "Deep Learning Algorithms for orthopaedic radiographs (DLAOR) had adopted transfer learning technique.

  Response: We have named this project DLOAR, so any reference to this acronym indicates the activities performed under this project. mentioned on page no. line no

  "*DLAOR stands for Deep Learning Algorithms for Orthopaedic Radiographs in the present study. DLAOR is an image classification system where images are fed into the model to classify the Knee Osteoarthritis (KOA) grade in medical images. It is built using a data ingestion pipeline, modelling engine and classification system".*

- The aim of this solution was to evaluate the feasibility and efficacy of the transfer learning algorithms in accurately assessing orthopaedic radiographs. Three key focus areas were determination of the disease pathology like osteoarthritis, identification of fractures and identifications of the implants". Is this study about OA AND fractures and implants as it is suggested by this paragraph or is solely about knee OA?

  Response: We have updated the objectives and made them precise to indicate and align with the outcomes of the proposed title. The idea is to identify the suitable model whose performance is good on the set of X-ray images (which are mostly standardised) of our hospital and thus bring automation for preliminary detection of KL grade by machine. Therefore, we have listed multiple objectives, however as guided in comments, we have shortened the objective and made it more targeted and precise. Page no. line no.

  "*This solution aimed to evaluate the feasibility and efficacy of the transfer learning algorithms in accurately assessing orthopaedic radiographs. Three key focus areas were determining the disease pathology like osteoarthritis, identifying fractures, and identifying the implants. The purpose of this study was to compare eight different transfer learning deep learning models for knee osteoarthritis grade detection from a*

radiograph and to identify the most appropriate model for detecting knee osteoarthritis grade."

- Please specify protocol for X Ray exposure, radiological incidence, were they taken by same radiologist/technician or not? X ray protocol can influence OA diagnostic and it is mandatory this remains consistent with training, testing and validation x Ray sets.

  Response: All the x-rays were standardised weight bearing X-rays taken on digital format using Seimens machine. The method is included in the manuscript:

  "Standardised Knee X-rays: For the AP view of the knee, the leg is extended and centred to the central ray. The leg is rotated slightly inward in order to place the knee and the lower leg in a true anterior-posterior position. The image receptor is centred on the central ray. For a lateral view, the patient is positioned on the affected side, with the knee flexed 20-30 degrees"

- Table 2 please give the formula for all parameters listed within the table.

  Response: Thank you for highlighting this, the necessary changes have been incorporated in the manuscript. it will improve the impact of the study and parameters used for evaluation. The update has been made as guided.

Table 2. Evaluation of parameters for KOA detection	
Accuracy	*This determines the accuracy of the standalone model inaccuracy to detect the presence of KOA and its classification in the input image Accuracy = (True Positive + True Negative)/ (True Positive + True Negative + False Positive + False Negative)*
Precision	*Precision = True Positive/ (True Positive + False Positive)*
Recall	*Recall = True Positive/ (True Positive + False Negative)*
Loss	*This study had leveraged cross-entropy multiple class loss. It is the measure of divergence of the predicted class from the actual class. It is calculated as a separate loss for each class label per observation and sums the result.*

- There is absolutely no comment regarding possible flaws of transfer learning model especially for this category of images. Negative transfer as well as overfitting are not only possible but highly probable in this kind of images and this could lead to the ridiculous interpreting of images. Validation set used here is not consistent enough to out rule this probability. AI based misdiagnosis is already there (we see it many times in lab data tests) and it really introduces way more hardship in resolving them. It is interesting to be high positive on AI based image interpretation but the consequences of many possible flaws need to be taken into account and, at least, discussed as such.

  Response: Firstly, appreciate your feedback to improve on regarding possible flaws of transfer learning models and their consequences. We thank the reviewer for this valid suggestion. A total of more than 2000 X-ray images have been reviewed for the study over the last two years. The idea is to test the accuracy and model results over the limited sample. As we move forward, we will train more images and train them in sophisticated systems in order to achieve higher accuracy and boost the results achieved in the current study. Our study with limited samples showed great promise and further enhanced our confidence in developing an assisted system for healthcare professionals. The Future prospects section describes the possible drawbacks of the model and key challenges associated with the data preparation and validation process. Also, it describes the approaches taken to overcome them. The necessary changes have been incorporated in the revised manuscript at page no. line no.

  "Any image classification project starts with a substantial number of images classified by an expert. This will provide inputs to the training of the model. There are multiple challenges that can be resulted from this miss classification of data during validation of the model by a machine, such as Intra-class variation, Image Scale variation, Viewpoint variation, i.e. Orientation of image, Occlusion – multiple objects in the

image, Illumination and Background clutter. While humans can overcome some of these possible challenges with naked eyes, it will be difficult for a machine to reduce the impacts of such factors while classifying the images. This results in an increased amount of loss and effect on the outcome of the prediction."

"*Another major drawback of utilising transfer learning models for image classification is negative transfer, where transfer learning ends up decreasing the accuracy or performance of the model. Thus there is a need to customise the models and their parameters to see the impact of tuning and epocs on the performance or accuracy of the model*"

We hope that the editor and the reviewers will find the revised manuscript acceptable as an important addition to the scientific literature. Once again thanking all the reviewers and the editors for their time and their detailed analysis and their help improving the quality of this study.