
Reviewers' Comments:

======= Review 1 =======

> *** Strong aspects: Comments to the author: what are the strong aspects of the paper

There exist many approaches that attempt to classify the existence of rotator cuff ruptures. However, many of these previous studies utilize MRI to make accurate classification, which can be expensive. Devising an approach that uses accessible wearable sensors can be a complementary or alternative approach that has procedurally and financially advantages. The submitted work seems to be an interesting step towards that end.

> *** Weak aspects: Comments to the author: what are the weak aspects of the paper?

The authors first show how the conventional logistic regression algorithm fails on their data. Based on their observation, the authors propose an approach that 1) trains a list of univariate logistic regression models using the dataset that leads to maximum precision, and then 2) ORs the classification results from these models to make the final classification results. The authors compare the results of the proposed approach against the results of other ensemble classification algorithms. Despite the better results compared to those of the existing algorithms, the overall study can be better justified by addressing the following points.

A. In section 2, the authors explained that clinicians diagnose rotator cuff injuries in various ways, one of which is to observe how the shoulder blades move. However, the authors chose to deploy the sensors on patients' shoulders, elbows, wrist, and sternum. I believe that the authors could have extracted the features of greater quality if the authors chose to deploy the sensors that can better describe the clinicians' thinking process. If there existed any practical challenges that prevent the authors from optimally deploying the sensors, it would be great to explain in the paper.

B. In section 2.3, the authors stated that they chose to extract the minimum and maximum shoulder angles for different arm configurations without any justification. I wonder if the authors can provide any academic papers that suggest a relationship between the features that the authors chose to use and the ways that clinicians diagnose the disease in question? I believe this is especially important as the authors make a great emphasis on the results being interpretable and convincing by clinicians. If the selected features do not convey any clinical implications, it is difficult to believe that clinicians will give credit to the classified results.

C. Due partially to the comments I made above, I question the value of section 3. It may be the poor choice of features that resulted in the poor classification results of logistic regression models. My question applies to section 3.3 and section 3.4 as well.

D. In section 3.2, the authors explained that 'a visual inspection of the exercises reveals that a possible explanation is that half of the patients have their hands on their lap at the start of the exercise, which the control group members do not.' I wonder if the authors failed to control the data collection procedure, where the study participants in two different groups assumed seemingly different postures. Albeit the control failure, I wonder if the authors could have post-process the data so that such uncontrolled movements at the very beginning of each data point can be systematically removed. Or it could have been even better if the authors could have

devised the features of higher quality that can systematically capture only the informative portion of the data point.

E. In section 3, I wonder why the author did not try to apply feature selection algorithm, which is a widely adopted approach, before learning classification models. The systematically selected subset of features will give the authors the most relevant features and rule out uninformative features that can negatively impact the classification results. The learned models may provide the authors with better classification results if the authors choose to use only the selected subset of features.

F. In section 4, I do not feel comfortable with one-sided logistic regression. It appears that the authors indirectly 'hand-pick' the decision boundary rather than letting the data identify the decision boundary, which is often the rationale of data-driven machine learning approaches.

G. In section 4.1, the authors hand-picked the hyperparameter f. However, I wonder if there is a reason that the authors did not identify this using nested validation?

H. In section 4.2, the authors explained the rationale of their proposed approach using a real-world scenario: '3 physicians diagnose the same patient on 3 completely different pieces of evidence. If one of the physicians diagnoses a patient as positive, this should overrule a negative diagnosis from the physicians that did not observe that bit of information." I do not believe this appropriately describes the actual medical practice. In reality, each clinician will consider all the evidence that he/she can collect and try to make the most informed decision. These clinicians will then gather together to discuss and make a final decision unless they have radically and exclusively different expertise.

I. In section 4.2, the authors compared the results of their proposed approach against the results of other algorithms. I wonder if the authors properly tuned the hyperparameters of these algorithms. Otherwise, it is difficult to say all the algorithms were compared fairly, and the proposed approach outperforms other algorithms on this particular data set.

> *** Recommended changes: Recommended changes. Please indicate any changes that should be made to the paper if accepted.

One potential change and maybe the easiest that the authors can make is to make a fair comparison against other ensemble classification algorithms. In other words, the authors are recommended to better tune the hyperparameters of the algorithms in comparison and report their best performances.

> *** Relevance and timeliness: Rate the importance and timeliness of the topic addressed in the paper within its area of research.

Little (2)

> *** Technical content and scientific rigour: Rate the technical content of the paper (e.g.: completeness of the

analysis or simulation study, thoroughness of the treatise, accuracy of the models, etc.), its soundness and scientific rigour.

Marginal work and simple contribution. Some flaws. (2)

> *** Novelty and originality: Rate the novelty and originality of the ideas or results presented in the paper.

Minor variations on a well investigated subject. (2)

> *** Quality of presentation: Rate the paper organization, the clearness of text and figures, the completeness and accuracy of references.

Readable, but revision is needed in some parts. (3)

======= Review 2 =======

> *** Strong aspects: Comments to the author: what are the strong aspects of the paper

Develop a prediction model for shoulder injury diagnosis that is easier to explain and hence more conducive for patients or doctors to accept when deployed.

Design a simple model where the univariate features can be easily extracted.

> *** Weak aspects: Comments to the author: what are the weak aspects of the paper?

The features used are relatively simple and hence may be the reason why they create false indicators as what the authors have described in Section 3.2.

Typically a good prediction model uses a combination of multiple features rather than purely on a particular feature e.g. max_AF or min_AF. In addition, the procedure taken to train one-sided logistic regression and the choice of threshold seems to hand crafted. Such procedure may result in a model that is too sensitive to the dataset used and not generalizable.

Dataset is small

If DL model can be used to extract distinctive features that may not create such false indicators.

Can the interquartile range of movement a better feature than just max, min, or range?

> *** Recommended changes: Recommended changes. Please indicate any changes that should be made to the paper if accepted.

if the authors can explore additional features that consider interquartile range of movements etc and see if that allows them not to have to use one sided classification to do this prediction and compares the 2 approaches to see which one is more robust, it will strengthen the paper.

> *** Relevance and timeliness: Rate the importance and timeliness of the topic addressed in the paper within its area of research.

Little (2)

> *** Technical content and scientific rigour: Rate the technical content of the paper (e.g.: completeness of the

analysis or simulation study, thoroughness of the treatise, accuracy of the models, etc.), its soundness and scientific rigour.

Marginal work and simple contribution. Some flaws. (2)

> *** Novelty and originality: Rate the novelty and originality of the ideas or results presented in the paper.

Minor variations on a well investigated subject. (2)

> *** Quality of presentation: Rate the paper organization, the clearness of text and figures, the completeness and accuracy of references.

Well written. (4)