

Dear reviewer,

Thank you for your constructive comments which have helped us to improve our manuscript. We have revised our manuscript according to your and the other reviewer's comments. Our point-by-point responses to your comments can be seen below.

Comments in black are the original comments from the reviewer. Our responses are in red and italics refers to the phrase revised in the text.

#### Comments and Suggestions for Authors

Using the Combination of Texture Analysis in B-mode Ultrasound, the authors compared different ML classifiers to identify Myofascial Trigger Points. Following are the comments that need to be addressed in the revised version of the paper.

1. Line 19: what is the meaning of the three methods calculated for each image? Does it mean three types of features are extracted? SEGL must be defined first before using it.

*Thank you for your comment. It was meant to mean the different category of features as described in the methods. We have revised the abstract to be more accessible and more general to better describe the methods and results at a higher level with details to be found in the main text if the reader is interested.*

2. Line 20: What is the meaning of "Then, the statistical features (e.g., entropy) were calculated over the B-mode-US images and those three methods."

*Thank you for pointing this out. The abstract has been completely rewritten as mentioned above to better describe the paper from a higher level. In terms of what that line meant... It was supposed to mean that the different features (the 7 in the paper) were calculated over the B-mode ultrasound images as well as the LBP and Gabor images and in SEGL method.*

3. The abstract needs re-writing to describe clearly what has been done.

*The abstract has been completely re-written.*

*(Line 12-24, Page 1) Myofascial Pain Syndrome is a chronic pain disorder characterized by myofascial trigger points (MTrPs). Quantitative ultrasound (US) techniques could be used to discriminate MTrPs from healthy muscle. In this study, 90 B-mode US images of upper trapezius muscles were collected from 63 participants (left and/or right side(s)). Four texture feature approaches (individually and a combination of them) were employed that focused on identifying spots and edges were used to explore the discrimination of the three groups: active MTrPs (n = 30), latent MTrPs (n = 30) from healthy muscle (n = 30). Machine learning (ML) and a one-way analysis of variance was used to investigate the discrimination ability of the different approaches. Statistically significant results were seen in almost all examined features for each texture feature approach but in contrast ML techniques struggled to produce robust discrimination. The ML techniques showed that 2 texture features (i.e. correlation and mean) within the combination of texture features were most important in classifying the 3 groups. This discrepancy between traditional statistical analysis and ML techniques prompts the need for further investigate of texture-based approaches in US for the discrimination of MTrPs.*

4. Line 47: what is the US? Always describe the full names before using the abbreviations.

*We have now defined ultrasound (US) in the text.*

(Line 40, Page 1) *“Ultrasound (US) is an attractive modality for this problem as it has been used to identify MTrPs.”*

5. Line 62: Earlier, the authors explained using Doppler and elastography ultrasound images to distinguish the affected muscle. Isn't it a gold standard for the classification?

Doppler and elastography are indeed the gold standard for classification but we meant to refer to the idea in B-mode ultrasound as most clinics or hospitals are equipped with a B-mode machine as opposed to doppler or elastography capabilities. This sentence has been revised.

(Line 57-58, Page 2) *“... there is currently no “gold standard” to detect the presence of MTrPs with B-mode US images.”*

6. Line 67: It is better if all these four categories of techniques may be compared, and the authors must highlight the pros and cons of these techniques. Summarizing in the form of tables will be very useful. Moreover, a comparison of performance on different datasets should be done in the literature review.

These four categories including the methods associated with each category were described in a review paper by Laleh Armi et al. From these four categories, we chose features / approaches that focuses on the spots, edges, pattern within the B-mode US images were consider due to the fact that myofascial trigger points are typically referred to as “knots” and thus we hypothesized that features that are designed for spots, edges or patterns would do well. We have revised the text to reflect this.

(Line 85-96, Page 2) *“Of these categories, we focused on features that may better describe spots, edges and patterns. This is because a variety of studies describe MTrPs as “knots” in the muscle. A wide variety of studies describe the MTrPs as a hyperechoic band, hypoechoic elliptical region, or simply a different echo-architecture than the surrounding muscle tissue in clin-ical examination (e.g., US screening) [17], [18].*

*For many clinicians and investigators, the finding of one or more MTrPs is required to assure the diagnosis of MPS. However, there remains a lack of optimal methods for characterizing these muscle structures and achieving an objective characterization of MTrPs has the potential to enhance their localization and diagnosis. This can facilitate the development of clinical measures [16]. One of the leading challenges in the classification of B-mode US images is that they may differ in scale, view or intensity. For these reason, various approaches attempt to address these challenges.”*

Reference:

Armi, L., & Fekri-Ershad, S. (2019). *Texture image analysis and texture classification methods - A review.* <https://doi.org/10.48550/arxiv.1904.06554>

7. Table 2: Why equations are shown in terms of a table. They can be explained in the text with the help of a figure. Several symbols in the equations are not defined.

We have revised the manuscript as you suggested. (Line 228-239, Page 6)

8. In Figure 2, what is the purpose of showing the Gabor feature for an MTrP subject and LBP for a healthy subject? It is better to show and identify the differences in the Gobor, LBP, and other feature maps.

Initially we were showing just examples of the different features for the reader, but following your suggestion we have revised the figure to show the different features on the active myofascial trigger point subject as this seems more informative to show how the MTrP may be characterized by those features. (Figure. 2) (Line 240, Page 7)

9. Figure 3: What are the RED color connections? How (40x8) is 280?

The red color represents the SEGL method, a combination of Statistical, Edge, gray-level-co-occurrence-matrices and LBP; the chart has been revised to better represent that (Figure. 3). This also has been added to the caption for the clarification. (Line 248-249, Page 7)

The 40 x 8 was a typo. We have revised this mistake.

10. Table 3: Explain and cite all the ML classifiers in this table. How are these classifiers selected from many ML classifiers in the literature?

We have added citations for the ML classifiers. These ML classifiers were selected due to their common use in many problems, their different strength and ease of implementation via the Scikit Learn library. This has been revised in the manuscript.

(Line 262-270) “The features calculated from each approach (Table 2) were used to train a variety of ML models to discriminate muscle with MTrPs (A-MTrPs and L-MTrPs) from healthy muscle. ML models were implemented in python using the Scikit Learn library. These ML models were: logistic regression (LR) [36], decision tree (DT) [37], random forest (RF) [38], k-nearest neighbors (kNN) [39], naive bayes (NB) [40], support vector machine (SVM) [41], and artificial neural network (NN) [42], [43]. These models were used because they are common in the literature [44], have different strengths and could easily be implemented. Each method used the libraries’ default parameters and other hyperparameters such as number of neighbors in kNN was tuned using grid search (Table 3).”

## References

- (36) Hosmer, D. W.; Lemeshow, S.; Sturdivant, R. X. Applied Logistic Regression, 3rd Edition. *Wiley Series in Probability and Statistics* **1989**, 528.
- (37) Quinlan, J. R. Induction of Decision Trees. *Machine Learning* **1986** 1:1 **1986**, 1 (1), 81–106. <https://doi.org/10.1007/BF00116251>.
- (38) Breiman, L. Random Forests. *Mach Learn* **2001**, 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- (39) Cover, T. M.; Hart, P. E. Nearest Neighbor Pattern Classification. *IEEE Trans Inf Theory* **1967**, 13 (1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>.
- (40) Lindley, D. V. Fiducial Distributions and Bayes’ Theorem. *Journal of the Royal Statistical Society: Series B (Methodological)* **1958**, 20 (1), 102–107. <https://doi.org/10.1111/J.2517-6161.1958.TB00278.X>.
- (41) Kecman, V. Support Vector Machines – An Introduction. **2005**, 1–47. [https://doi.org/10.1007/10984697\\_1](https://doi.org/10.1007/10984697_1).
- (42) McCulloch, W. S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull Math Biophys* **1943**, 5 (4), 115–133. <https://doi.org/10.1007/BF02478259/METRICS>.
- (43) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986** 323:6088 **1986**, 323 (6088), 533–536. <https://doi.org/10.1038/323533a0>.
- (44) Uddin, S.; Khan, A.; Hossain, M. E.; Moni, M. A. Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. *BMC Med Inform Decis Mak* **2019**, 19 (1), 1–16. <https://doi.org/10.1186/S12911-019-1004-8/FIGURES/12>.

11. Line 266: Explain (256x512x128) in the NN. Is 256 the size of the input layer? What are these 256 inputs?

Thanks for mentioning this, this was a transcribing error from the 2 manuscripts we are working on. The actual neural network was a single hidden layer network with 512 nodes and a dropout layer (50% of nodes are dropped) for this manuscript and have been revised accordingly.

The numbers you were asking about refer to the hidden layers, the inputs are explained in a new paragraph. These revisions should provide more clarity now.

(Line 274 - 284) *"The NN was a single hidden-layer network (512) with a dropout layer (50% of nodes dropped). All activation functions were Rectified Linear Units. The output layer was a 3-node output with an activation function of SoftMax. The NN was trained for 250 epochs with early stopping criterion of 7 epochs of no improvement in the validation loss. The learning rate was the default set by Keras and was decreased by a factor of 0.1 after 3 epochs of no improvement to a minimum of learning rate = 0.00001."*

*Input to all classifiers were the features from each approach as seen in Table 2. A leave-one-site-out approach was used due to the low number of images to better evaluate performance. The remaining examples were used for training (i.e., LR, DT, RF, k-NN, NB, SVM) with the exception of the NN approach where they were split into 75% training and 25% validation sets."*

12. Line 272: Describe the inputs to the ML classifiers more clearly.

We have added a new table to directly show the number of inputs features clearly and also have added a new paragraph in section 2.4. (Table. 2) (Line 250, Page 7)

(Line 280 – 284, Page 8): *"Input to all classifiers were the features from each approach as seen in Table 2. A leave-one-site-out approach was used due to the low number of images to better evaluate performance. The remaining examples were used for training (i.e., LR, DT, RF, k-NN, NB, SVM) with the exception of the NN approach where they were split into 75% training and 25% validation sets."*

13. Line 274: How the training and testing datasets are divided is unclear. It will be helpful if the authors make a table to show how many data points are used for training, validation, and testing.

We have revised the text to better reflect the division of the datasets with an example.

(Line 284 – 287, Page 8): *"For example, a training set would consist of 356 images (89 sites x 4 US images) and a test set consist of 4 images. In the case of the NN, the training and validation sets would consist of 268 and 88 images (67 sites x 4 US images; 22 sites x 4 US images) respectively."*

14. Line 278: The ROC curve and area under the ROC curve can also be good performance metrics.

Thank you for the suggestion. The ROC curves are a way of graphically displaying true-positives versus false-positives across a range of cut-offs. In this study, we have calculated a variety of metrics (true-positive, false-positive, etc.) that can provide us with similar information to evaluate our models. Thus, in this revision, we have chosen to not include an ROC curve into the manuscript.

## Reference

Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev. 2008 Aug;29 Suppl 1(Suppl 1):S83-7. PMID: 18852864; PMCID: PMC2556590.

15. Line 281: Two "And" are in the title. The ensemble has a meaning in ML, a combination of classifiers not reflected' in this section.

We have revised the text to better explain the ensemble approach. We indeed did use an ensemble approach where we make use of the different machine learning classifiers to make a prediction. It was done by selecting the best performing classifiers and using a majority vote of their predictions to obtain the final result.

(Line 291, Page 8): The title was revised to *"Ensemble Approaches, Feature Importance & Statistical Analysis"*

(Line 292-296, Page 8): *"The ML techniques were further investigated using an ensemble approach. The best performing trained classifier for each technique (e.g. kNN, SVM, etc.) was selected based on the mean performance across all 4 feature approaches (i.e., B-mode LBP, Gabor, SEGL), as shown with asterisks (\*) in Table. 3. These selected classifiers were then used to perform a majority vote for classification task."*

16. Line 188: What is the meaning of leave on feature out approach? Typically, different combinations of features are used without mentioning leaving one out.

We have revised the text to better explain what we mean. This approach was meant to observe how important a single feature was on the classification performance. This was why we investigated the effect of the classification result by using a single feature and by removing the particular feature (i.e. removal of entropy from the set of 7 features to see the change in the classification performance.). We have removed the phrase leave-one-feature-out from the text to mitigate any misunderstanding.

(Line 298-300, Page 8) *"In addition, to determine which features were more important toward the classification task we examined the classification performance of using the single statistical feature (e.g., entropy, mean, etc.) and removing a single feature (from the set of 7)."*

17. Section 3: This section is poorly written. Results are reported in Tables 4 and 5, but the explanation is unclear. A better and more detailed description is required. How is the majority voting used? The majority voting is not explained in the methodology section. Training, validation, and testing are not described in the results section. Results should be reported as mean and standard deviation in Table 4.

We have completely rewritten this section to provide better clarity. The results section now flows from the initial results with each classifier for each approach and then to the ensemble methods and the statistical results with the interpretations moved to the discussion (Line 373-404). The methodology has also been revised to ensure the content is clear to the reader The relevant revisions are shown below:

Methods: (Line 284-308, Pages 8 & 9)

*"For example, a training set would consist of 356 images (89 sites x 4 US images) and a test set consist of 4 images. In the case of the NN, the training and validation sets would consist of 268 and 88 images (67 sites x 4 US images; 22 sites x 4 US images) respectively.*

*The ML techniques were further investigated using an ensemble approach. The best performing trained classifier for each technique (e.g. kNN, SVM, etc.) was selected based on the mean performance across all 4 feature approaches (i.e., B-mode LBP, Gabor, SEGL), as shown with asterisks (\*) in Table. 3. These selected classifiers were then used to perform a majority vote for classification task. This was implemented via the function majorityvote in MATLAB.*

*In addition, to determine which features were more important toward the classification task we examined the classification performance of using the single statistical feature (e.g., entropy, mean, etc.) and removing a single feature (from the set of 7).*

*For the single statistical feature case, we took the features from all approaches (as seen in table 2) and used only the statistical feature of interest. In cases, where there were more than 7 features (i.e., Gabor and SEGL), the mean values were used (i.e. Gabor: 40 entropy features converted into a single mean entropy feature for the Gabor approach). This results in a vector of 4 values (i.e. entropy feature from the 4 approaches).*

*For the removal of a single feature case, the same procedure was used except the features that were not removed were used as inputs (i.e. vector of 24 values (6 statistical features x 4 approaches)."*

Results: (Line 312-330, Page 9)

*"Table. 4 shows the classification accuracy (%), F1-score, sensitivity, specificity, PPV, and NPV for the best parameter of each ML technique with the bolded values showing the best performance for each parameter: for each approach (B-mode, LBP, Gabor feature, and SEGL).*

*Figure. 5 shows the confusion matrixes of the ML techniques for: each approach (B-mode, LBP, Gabor feature, and SEGL); each approach with the majority voting; a single statistical feature; the removal of a single statistical feature. For each analysis, the ML classifier with the parameter that presented the best performance are shown in Table.5.*

**Table. 5** shows the classification accuracy (%), F1-score, sensitivity, specificity, PPV, and NPV for the ensemble approach and the effects of using a single statistical feature and the removal of a single statistical feature. Highest performance can be seen with the “correlation” feature (accuracy = 53.33%, F1-score = 0.4861) and the removal of variance (accuracy = 51.67%, F1-score = 0.518).

**Table. 6** shows results of the statistical analysis (mean and standard deviation) from all four approaches (B-mode, SEGL, Gabor, LBP) between all 3 groups: A-MTrPs, L-MTrPs, and healthy controls with bolded values showing statistical significance. Statistical differences ( $p < 0.05$ ) were seen for all almost all features for all approaches except in B-mode (i.e., entropy, contrast and energy) and Gabor (i.e., mean and correlation).”

18. What is the effect of parameters in ML classifiers, and how the parameters are optimized?

The parameters can be optimized to ideally improve performance of the classifier and usually optimized via some kind of tuning. In this case we tuned it by performing a simple grid search.

(Like 267-270, Page 8) “These models were used because they are common in the literature<sup>44</sup>, have different strengths and could easily be implemented. Each method used the libraries’ default parameters and other hyperparameters such as number of neighbours in kNN was tuned using grid search (Table 3).”

19. It is essential to compare the reported results with the published literature in this domain. I think there are many papers in this area. So, it is necessary to describe how the reported results are better than the published results and the novelty of the work. Two of the papers are mentioned below,

Ryan G.L. Koh, Banu Dilek, Gongkai Ye, Alper Selver, Dinesh Kumbhare, Myofascial Trigger Point Identification in B-Mode Ultrasound: Texture Analysis Versus a Convolutional Neural Network Approach, Ultrasound in Medicine & Biology, Volume 49, Issue 10, 2023, Pages 2273-2282

Behr M, Saiel S, Evans V, Kumbhare D. Machine Learning Diagnostic Modeling for Classifying Fibromyalgia Using B-mode Ultrasound Images. Ultrasonic Imaging. 2020;42(3):135-147. doi:10.1177/0161734620908789

Thanks very much for your comments. We have added 2 paragraphs in the discussion comparing the results seen here with Koh et al. and Kumbhare et al. The second paper you suggested focused more on fibromyalgia and instead we have compared it to the other study we have included here which focuses on MPS.

(Line 365-370, Page 13): These results are comparable to another study that compared texture features to a CNN approach[45]. Their F1-score ranged from 0.383 to 0.477 for their texture approaches (i.e. first-order statistical, LBP, and blob analysis) when classifying these three groups using a NN. This study shows better performance in the texture feature approach, which may be attributed to the combined ensemble approach and features that focus on the structural information (i.e. spots, and edges).

(Line 390-401, Page 13 & 14): The result of our traditional statistical analysis agrees with results seen in previous literature [8]. One study using LBP and blob analysis demonstrated statistically significant results between healthy individuals and patients with MPS ( $p < 0.001$ )[8]. Based on this they suggested that a combination of texture features (i.e., LBP, blob area and count) can be used to describe differences between the individuals MPS and healthy individuals using a principal component analysis. However, this study grouped individuals with both A-MTrPs and L-MTrPs into the group of individuals with MPS. Koh et al. demonstrated better performance in classifying MTrPs (i.e., A-MTrP and L-MTrP grouped) from healthy muscle compared to the three-group case (i.e., A-MTrP, L-MTrP and healthy muscle) [45]. These studies within the literature plus the results seen in this study suggest that MTrPs can be distinguished from healthy muscle but may not be sufficient for discrimination between the two types of MTrPs (i.e., A-MTrPs from L-MTrPs).

Reference

Kumbhare, D., Shaw, S., Ahmed, S., & Noseworthy, M. D. (2020). Quantitative ultrasound of trapezius muscle involvement in myofascial pain: comparison of clinical and healthy population using texture analysis. *Journal of Ultrasound*, 23(1), 23–30. <https://doi.org/10.1007/s40477-018-0330-5>

The content has been revised for the English language coherence. We also added the symbols that were not defined in the text. So, if some parts are highlighted, that is due to the revision of the content to improve clarity.