Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-161

**Response to Reviewer-A Comments**

The manuscript focuses on prediction of glioma grade. The authors extract features using CNNs and then use these features in machine learning models. The performance was evaluated using F1 scores, accuracy, recall, and precision. My comments are below:

**Comment-1**. Recall and sensitivity both focus on the positive class, it would be useful to also see specificity and negative predictive value. In addition, all of the measures proposed focus on discrimination and showing how calibrated the algorithm is would be useful.
**Reply-1:** Thank you for your insightful feedback. We acknowledge the importance of considering a comprehensive range of metrics, including specificity and negative predictive value (NPV), to provide a more holistic evaluation of the proposed method. In response to your suggestion, we have updated our results to include both specificity and NPV in the results table. We have now included the calibration plot for the proposed algorithm in the discussion section of the revised manuscript.
**Changes to the text:** We have added specificity and negative predictive value (NPV) to the results table (please see Page 15 Table 1).
Added calibration plot as Figure 9 in Page 20 and the corresponding description in Page 19 Line 331-333.

**Comment-2:** No measures of uncertainty about the model performance are provided making it hard for the reader to judge how different the methods actually are.
"These findings highlight the potential of CNN models to positively impact clinical practice by enhancing the accuracy of glioma grading." How do the authors envision this impacting clinical practice? Similarly, "These insights are valuable to medical professionals and researchers seeking optimal methods for glioma grade prediction. Our findings contribute to the advancement in the field of glioma grading, paving a way for improved diagnosis and treatment." This seems like a strong statement and listing the steps the authors think should be done in order for getting this into clinical practice would help.
**Reply-2:** Thank you for the suggestion. We agree with the reviewer. We have now computed the confidence interval (CI) for F1-scores for various methods discussed in the manuscript. Please see Page 17 Line 286-294.
We revised our statements in both abstract and conclusion sections of the manuscript and added a paragraph in the discussion section outlining the steps needed to transform the proposed method into standard medical practice.
**Changes in the text:** Computed CI and added the results to Section 3 of the manuscript. Please see Page 17 Line 286-294.

Updated the last sentence in the abstract Conclusions part. Please see Page 1 Line 24-25.
Updated the last sentence in the Conclusions section of the manuscript. Please see Page 21 Line 356-357.
Added a new paragraph to the discussion section of the manuscript. Please see Page 19-20 Line 334-340.

**Comment-3:** Given the small sample size and the high performance I suspect that differences in precision and recall are sensitive to a very small number of cases making the results less replicable. To evaluate that giving the 2x2 matrix of true status vs predicted status would be useful.
**Reply-3:** Thank you for the suggestion. We have now added the 2x2 matrix of true status vs predicted status as Figure 8 in Page 16 of the revised manuscript.
**Changes in the text:** Added a new figure showing the confusion matrix for the proposed method as Figure 8 in Page 16 and added corresponding description in Page 16-17 Line 274-285.

**Comment-4:** There is no information on how to select the cut-off for defining positive or negative predictions.
**Reply-4:** Sorry about that. We have now clarified the selection of cut-off for defining positive and negative predictions. We performed five-fold cross-validation and used F1-score as a metric to determine the optimal threshold: i.e., we computed F1-score at different thresholds and selected the one that has the highest F1-score.
**Changes in the text:** Added the information on cut-off selection for defining positive and negative predictions. Please see Page 8 Line 139-145. The optimal threshold value for the proposed method was given in the results section of the revised manuscript (see Page 16 Line 271-272).

## Response to Reviewer-B Comments

The authors present a method to classify glioma grade. The experiments compare various combinations of feature extraction methods and classification methods. The authors propose to extract CNN-based features. The challenge lies in the lack of data and diversity of the data collection process. The authors solve the challenge by preparing 2D slices from 3D MRI scans. This manuscript should be improved in the following aspects.

**Comment-1:** The existing methods and the proposed method should be clearly separated, especially around the autoencoder and U-Net parts. For example, making a separate section for the existing method would suffice.
**Reply-1:** Thank you for the suggestion. We have now added a separate section for the existing methods. Please see Section 2.3.2 (Page 10) for the existing methods and Section 2.3.3 (Page 13) for the proposed method in the revised manuscript.
**Changes in the text:** Added a separate Section 2.3.2 (see Page 10 Line 172) for existing methods and Section 2.3.3 for the proposed method (see Page 13 Line 222).

**Comment-2:** The results could be more readable by marking bold on the best performance or plotting the numbers.
**Reply-2:** Thank you for the suggestion. We have now marked the best performance in the results section in bold. Please see Tables 1-3 in the revised manuscript.
**Changes in the text:** Best performance is indicated in bold in Table 1 (Page 15), Table 2 (Page 18), Table 3 (Page 19).

This manuscript can be further improved by answering the following questions.

**Comment-3:** What are the characteristics of the 2D slices of 3D scans? I suppose that the neighboring slices would be very similar to each other leading to redundancy.
**Reply-3:** We agree with the reviewer. The neighboring slices in 3D scans are very similar to each other leading. We have now elaborated on the characteristics of 2D slices in Section 2.3.1 of the revised manuscript.
**Changes in the text:** Elaborated on the image extraction step. Please see Page 7-8 Line 129-137 in the revised manuscript.

**Comment-4**. Why do we need different classifiers (SVM, decision tree, ...) other than the one used for training the CNN?
**Reply-4:** Sorry for not being clear. We have now compared the performance of CNN (for feature extraction) + ML (for glioma grade classification) model with using only CNN (for both feature extraction and grade classification). The CNN + ML models performed better compared to the CNN alone. A detailed comparison was given in Table 3, which clearly suggests the need of further training ML models on learned features extracted from CNN methods.
**Changes in the text:** Added results of CNN alone network to determine the need of using ML models for grade classification. Please see Table 3 in Page 19 and its corresponding text in Page 19 Line 324-330.