Peer review

1.Please provide the source(s) of the images. You state "….6400 MRI images of Alzheimer's patients sourced from public hospitals,…". If you obtained these images from multiple websites or sources; make sure those sources are public and you are not compromising private patient data. If there are too many of the sources, then please provide access to all the images either in a supplementary file or in a public repository such as Github. The readers must be able to replicate your work.

2.Did you split the dataset without consideration of classes? In other words, did you perform a stratified or non-stratified split? If the latter, provide evidence that equal proportions of images in each class were represented in the training and testing datasets. Please discuss in the manuscript.

3.How many features did your CNN model extract from the MRI images (see: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8736208, figure 6). What was the feature importance? The train-to-test split ratio should have been determined by the inverse of the square root of the number of features. What was the rationale for choosing the 80:20 split ratio? Justify. Please present the features extracted by the model, the feature importance and the physical meaning behind each feature in the manuscript.

4.How was multicollinearity between features (if any) addressed? Did you calculate pearson's correlation for each feature? Did you perform a VIF or PCA? Please present a pearson correlation heat map of the features in the form of a correlation matrix or a correlogram. Even if multicollinearity does not influence ultimate prediction, it is useful to use the least number of features to ensure computation efficiency and robustness. Please describe in the manuscript.

5.You have a large class imbalance between the four categories. How was this addressed? Oversampling by SMOTE? Augmentation by creating synthetic data? Please discuss in the manuscript.

6.Please report a confusion matrix for the actual class versus the predicted class for your test data. Further, please report the precision, recall and the F1 values for your model. Then report your AUC in the form of a false positive on the x axis and a true positive on the Y axis. Please present a figure for the AUC.

7.MRI images for detecting Alzheimers (by themselves) have an accuracy of anywhere from 50 to 85%. Will this sensitivity and specificity not be compounded when a CNN or an ML model is applied? In other words, are you not adding the error of your ML to the inherent error of the MRI images? Your 'true' values- are after all - whatever is assigned as the true value to each individual MRI image - which - itself has an accuracy of 50-85%. This is an important point overlooked by many researchers and publications. Please address and discuss in the manuscript.

8.References should be sequentially numbered throughout the text of the manuscript and the numbered references in the References section should match those in the text. Whereever there are 6 or more authors, all 6 should be listed then followed by an et al.

Reviewer Comment 1:

"Please provide the source(s) of the images. You state "…6400 MRI images of Alzheimer's patients sourced from public hospitals,…". If you obtained these images from multiple websites or sources;

make sure those sources are public and you are not compromising private patient data. If there are too many of the sources, then please provide access to all the images either in a supplementary file or in a public repository such as Github. The readers must be able to replicate your work."

Response:

The dataset used in this study has been sourced from a publicly available Kaggle Alzheimer MRI Dataset, and there is no private patient information included. We have cited the source in the manuscript (Section 3.1), [19], as well as in the sources section. This ensures transparency, and readers are able to replicate our work using the same data.

Reviewer Comment 2:

"Did you split the dataset without consideration of classes? In other words, did you perform a stratified or non-stratified split? If the latter, provide evidence that equal proportions of images in each class were represented in the training and testing datasets. Please discuss in the manuscript."

Response:

A stratified split was performed to ensure that each of the four classes was proportionally represented in both the training and testing datasets (Section 3.2.1). This approach ensured that class distribution remained consistent between the two sets, reducing bias. We have further expanded this discussion and included the class distribution statistics in the revised manuscript for clarity.

Reviewer Comment 3:

"How many features did your CNN model extract from the MRI images (see: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8736208, figure 6). What was the feature importance? The train-to-test split ratio should have been determined by the inverse of the square root of the number of features. What was the rationale for choosing the 80:20 split ratio? Justify. Please present the features extracted by the model, the feature importance and the physical meaning behind each feature in the manuscript."

Response:

The CNN model extracted 2048 features from the final convolutional layer, which captures meaningful patterns in the MRI images (Section 3.2.1). The 80:20 train-test split ratio was selected as it provided an effective balance between sufficient training data and a reliablee stimate of performance on the test data. While the theoretical ratio suggests a different, more skewed split, the 80:20 split allowed for more robust evaluation, as discussed in Section 3.2.1. Regarding feature importance, we have added a section discussing the relative importance of features and the model's ability to detect patterns relevant to Alzheimer's stages. We have also referenced the feature extraction method, detailing its significance in Section 3.2.4.

Reviewer Comment 4:

"How was multicollinearity between features (if any) addressed? Did you calculate Pearson's correlation for each feature? Did you perform a VIF or PCA? Please present a Pearson correlation heat map of the features in the form of a correlation matrix or a correlogram. Even if multicollinearity does not influence ultimate prediction, it is useful to use the least number of features to ensure computation efficiency and robustness. Please describe in the manuscript."

Response:

We conducted a Pearson correlation analysis to assess the relationships between the features extracted by the CNN (Section 3.2.4). The results are visualized in a heatmap (Figure 4), showing minimal multicollinearity among the features. This ensures the extracted features are diverse and independent, enhancing computational efficiency. We have updated the manuscript to include this discussion and emphasized the importance of avoiding redundant features
(Section 5.1).

Reviewer Comment 5:

"You have a large class imbalance between the four categories. How was this addressed? Oversampling by SMOTE? Augmentation by creating synthetic data? Please discuss in the manuscript."

Response:

Class imbalance was addressed through data augmentation techniques such as random rotations, horizontal flips, and color jitter, which increased the diversity of the training data and improved generalization (Section 3.2.3), and class weighting was also applied to further address the class imbalance in the dataset. (Section 3.4.1). This approach helped mitigate the effects of class imbalance through two approaches. We have clarified and expanded upon this in the revised manuscript.

Reviewer Comment 6:

"Please report a confusion matrix for the actual class versus the predicted class for your test data. Further, please report the precision, recall and the F1 values for your model. Then report your AUC in the form of a false positive on the x axis and a true positive on the Y axis. Please present a figure for the AUC."

Response:

We have reported the confusion matrix in Section 4.2 (Figure 13), along with precision, recall, F1-scores, and AUC values. Additionally, we included the ROC curves (Figure 14) for each class, as requested. These results highlight the strong performance of the model in classifying different stages of Alzheimer's disease, with AUC values close to 1.0 for all categories.

Reviewer Comment 7:

"MRI images for detecting Alzheimer's (by themselves) have an accuracy of anywhere from 50 to 85%. Will this sensitivity and specificity not be compounded when a CNN or an ML model is applied? In other words, are you not adding the error of your ML to the inherent error of the MRI images? Your 'true' values—are after all—whatever is assigned as the true value to each individual MRI image—

which—itself has an accuracy of 50-85%. This is an important point overlooked by many researchers and publications. Please address and discuss in the
manuscript."

Response:

In Section 5.1, we addressed this concern by acknowledging the inherent variability in MRI accuracy, which typically ranges from 50-85%. We discussed how errors introduced by the MRI images could compound with errors from the model. To mitigate this, we proposed future work incorporating additional modalities (e.g., genetic or clinical data) to improve overall accuracy and reduce reliance on MRI-based labels. This section now includes a more detailed discussion
of these limitations.

Reviewer Comment 8:

"References should be sequentially numbered throughout the text of the manuscript and the numbered references in the References section should match those in the text. Wherever there are 6 or more authors, all 6 should be listed then followed by an et al."

Response:

The references have been revised throughout the manuscript to ensure they are sequentially numbered and match the citations in the text. Additionally, we have listed all six authors where applicable, followed by "et al." in compliance with the formatting guidelines.

Thank you for addressing my comments. The confusion matrix presented in Figure 13 needs true and predicted labels. There is not a way to determine which row/column corresponds to which category of disease. Also, please provide a word document of the manuscript. Also provide each figure separately as a JPEG file of adequate resolution. Thanks.

Reviewer Comment 1:

"The confusion matrix presented in Figure 13 needs true and predicted labels. There is not a way to determine which row/column corresponds to which category of disease.."
Response:

A separate confusion matrix with true and predicted labels corresponding to each category of disease has been visualized in Figure 14 of section 4.2, as well as an entirely separate classification report visualized in Figure 13 of section 4.2 to analyze the model's effectiveness.
Reviewer Comment 2:

"Also, please provide a word document of the manuscript.."

Response:

The manuscript has been reformatted into docx format, and the pdf version is also attached in the additional files section.

Reviewer Comment 3:

"Also provide each figure separately as a JPEG file of adequate resolution."

Response:

Each image has been provided as a jpeg file in the additional files sections, labeled by the figure number they were given in the manuscript