Peer-Review

Wong, Gavin. 2025. “A Comparative Study of Convolutional Neural Network and Vision Transformer Models as Classifiers of East Asian Traditional Clothing.” *Journal of High School Science* 9 (3): 345–69. https://doi.org/10.64336/001c.143597.

This is a well designed and well executed manuscript. Congratulations. However, it has the potential to do better. Please see my comments below

1.To what extent is the classification accuracy dependent on the background (images of pagodas, cherry blossoms, or buildings or city styles distinctive of Korean or Japanese or Chinese culture…..). Please discuss. Is there a way you could use the grabCut algorithm (see references below) to purge your backgrounds and then run your ViT again? This will tell you to what extent your algorithm is factoring in backgrounds toward accurate classification.

2.I would like you to do the following. Gather Saudi Arabian, Indian, Bangladeshi, Indonesian traditional clothes images (from existing public datasets) and feed them into your algorithm as ONLY TESTING (NOT TRAINING) images. Your algorithm will obviously classify the images incorrectly as either Korean, Chinese or Japanese. However, by evaluating what features of clothing get classified into what category, you can obtain more information on the mechanics of your algorithm. For example, if your AI algorithm classifies all images with a central sash around the torso, or a “T” shape as Japanese; classifies all images with tight tops and high waistline as Korean and/or all images with ‘loose flowing’ motifs (without a clear distinction of top and dress) as Chinese, you can then infer that these features are being assigned greater importance in classification (or not).
The objective of this exercise is to improve the ‘explainability’ of your (and - in general - AI algorithms) which has been a bane toward their stand-alone applicability. This will go a long way to distinguish your manuscript from the published literature.

3.Also inform how many images were male and female in your algorithm and whether a stratified split (along country/culture) lines was performed.

4.What is the robustness of your algorithm? Was any k-crossvalidation performed? If not, why not? Please discuss in the manuscript.

5.Present whether CNN needed lesser epochs to get to its maximum accuracy ot 73.17%. Did your ViT algorithm need more epochs to get to 73.17% ? Present plateau accuracy of your algorithm versus CNN at the same number of epochs. If CNN were allowed to get to the same # of epochs as ViT, would its accuracy then be comparable to ViT? Discuss in the manuscript.

6.Present a confusion matrix for both the CNN and the ViT algorithms.
https://doi.org/10.3390/info15040196
https://dx.doi.org/10.21608/idj.2024.254693.1106

Comment 1: “To what extent is the classification accuracy dependent on the background (images of pagodas, cherry blossoms, or buildings or city styles distinctive of Korean or Japanese or Chinese culture…..). Please discuss. Is there a way you could use the grabCut algorithm (see references below) to purge your backgrounds and then run your ViT again? This will tell you to what extent your algorithm is factoring in backgrounds toward accurate classification.”

Revisions: To address this, I purged the backgrounds of my images using the Python library rembg as it was the fastest, yet still effective method. The method is detailed on pages 12 and 13 underneath the Model Explainability section. The results of this test are under Model Explainability Analyses on page 15. The results are discussed under the Model Robustness and Explainability section on page 17.

Comment 2: "I would like you to do the following. Gather Saudi Arabian, Indian, Bangladeshi, Indonesian traditional clothes images (from existing public datasets) and feed them into your algorithm as ONLY TESTING (NOT TRAINING) images. Your algorithm will obviously classify the images incorrectly as either Korean, Chinese or Japanese. However, by evaluating what features of clothing get classified into what category, you can obtain more information on the mechanics of your algorithm. For example, if your AI algorithm classifies all images with a central sash around the torso, or a "T" shape as Japanese; classifies all images with tight tops and high waistline as Korean and/or all images with 'loose flowing' motifs (without a clear distinction of top and dress) as Chinese, you can then infer that these features are being assigned greater importance in classification (or not).
The objective of this exercise is to improve the 'explainability' of your (and - in general - AI algorithms) which has been a bane toward their stand-alone applicability. This will go a long way to distinguish your manuscript from the published literature."
Revisions: Since no there were no public datasets of Saudi Arabian, Indian, Bengali, and Indonesian traditional clothing, I opted to create a dataset of the aforementioned cultures using the same Google Chrome extension used for my East-Asian dataset. The process was detailed on page 13 underneath the Model Explainability section. The results of the test are under the Model Explainability Analyses on page 16. The results are discussed under the Model Robustness and Explainability section on page 17. Although the comment asked to try and obtain more information on the mechanics of the algorithm, the results showed a very strong bias towards Chinese clothing that could not be easily explained upon a visual inspection of clothing characteristics. The discussion section expands upon this idea on page 17. Potential areas for future works in this area were discussed on page 19.

Comment 3: "Also inform how many images were male and female in your algorithm and whether a stratified split (along country/culture) lines was performed."
Revisions: On page 9, the disparity between the number of male and female clothing images was acknowledged in my Methodology. Furthermore, the table on page 9 now showcases the number of male images, allowing the reader to figure out the number of female images with simple calculations. On page 11, I discussed how a stratified split across cultural lines was used and that gender was not used as a basis for stratification. Finally, on page 18, I discuss the limitations of the gender disparity in my dataset.

Comment 4: "What is the robustness of your algorithm? Was any k-crossvalidation performed? If not, why not? Please discuss in the manuscript."
Revisions: The robustness of my algorithm was discussed in the Model Robustness and Explainability section on page 17. Since model robustness went hand in hand with the two additional analyses conducted to address comment 1 and 2, the sections were combined. Moreover, the robustness of my algorithms was discussed in the future works section on pages 18 and 19. As for k-fold cross-validation, I acknowledged the reasoning behind not using the technique on page 11 and the potential to use it for future works on pages 18 and 19. Finally, I addressed the limitations of both models in real world applications due to their low robustness to unseen data on pages 18 and 19.

Comment 5: "Present whether CNN needed lesser epochs to get to its maximum accuracy ot 73.17%. Did your ViT algorithm need more epochs to get to 73.17% ? Present plateau accuracy of your

algorithm versus CNN at the same number of epochs. If CNN were allowed to get to the same # of epochs as ViT, would its accuracy then be comparable to ViT? Discuss in the manuscript."
Revisions: On page 14, further analysis and discussion of plateau accuracy and algorithm convergence was added. Moreover, a table was added to showcase the peak accuracy of each model and the number of epochs it took to reach it.

Comment 6: "Present a confusion matrix for both the CNN and the ViT algorithms."
Revisions: On page 15, the confusion matrices for both models were added along with a discussion that analyzes the data within them. I also discussed once more how the ViT showed superiority as displayed in the figure of the confusion matrices.

Thank you for addressing my comments. Accepted. We are copyediting your manuscript and have made changes to content as well as language. Can you please review the preliminary copyedit (see attached) and write a conclusion section? You can write it into a separate document and send me only the conclusion section. Please send me the document when done, preferably ASAP.