Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-38

Reviewer A

Comment 1: Some sentences are too long or hard to understand and should be rephrased.

Reply 1: We rephrased and divided long sentences to improve readability throughout the paper.

Change in text:

Line 550 (original): In the setting of implementation of the machine learning risk stratifier as an additional feature involved in decision making, E_i makes accurate decisions using highly specific features.

Revised: In the setting of implementation of the machine learning risk stratifier as an additional feature involved in decision-making, E_i makes accurate decisions by highlighting specific features of the nevus.

Comment 2: In some sections of the manuscript related to the experiments, authors used the term DL, but the experiments are related both to DL and ML methods.

Reply 2: To the best of our knowledge, we did not notice any instances where this was the case. All references to N_i were with regards to DL networks and references to E_i involved EL networks.

Comment 3: The symbols on line 405 and 447 have not been defined.

Reply 3: For Line 405, we reorganized the placement of N_i to directly follow the word "networks" with the intention of indicating N_i refers to the deep learning networks trained. Additionally, for Line 447, the symbols LR and WD are defined on Line 411 to represent learning rate and weight decay, respectively. As such, we do not believe there is an indication to alter Line 447.

Change in text:

Line 405: Five total networks (N_i: N_A, N_D, N_M, N_R, N_S)

Comment 4: The adoption of the one-hot encoding should be detailed in the case of lesion location.

Reply 4: We understand this can provide clarity for researchers attempting to recreate the project and have included a table to systematically break down how each lesion location was numerically categorized. All subsequent tables references were adjusted accordingly to accommodate the new table.

Change in text: The addition of Table 2: One hot encoding breakdown of a) sex and b) lesion location.

Line 433: One-hot encoding was implemented on biological sex and location of lesions to numerically categorize the classes included for training (Table 2).

Comment 5: Figures 4 and 5 are not flow charts.

Reply 5: We agree with this sentiment and decided to reference the figure as "pipeline".
Change in text: Figure 4 and 5 captions are changed to replace "Flowchart" with "Pipeline"

Comment 6: Figure 8 is blurred and hard to read.
Reply 6: We captured a high-definition picture from the Python output and utilized this image.
Change in text: Figure 8a and 8b (now Figure 9a and 9b) have been replaced with high-definition pictures.

Comment 7: No comparisons with the state of the art have been carried out.
Reply 7: In our Discussion section, we highlighted several studies utilizing state-of-the-art network architectures for classification of skin lesions in the HAM10000 dataset.
Change in text: Following Line 622 of the original manuscript, we included the following paragraph: This study served as a baseline in demonstrating the improvement of classification performance of melanoma with the inclusion of a machine learning risk score as an additional feature in deep ensemble training. We utilized historically-stable classification networks to test our hypothesis. EfficientNet (56) is a novel network architecture that has demonstrated impressive capabilities in classification challenges. Ali et al. demonstrated its diagnostic accuracy with EfficientNet base networks and modified derivatives yielding AUCs ranging from 0.96-0.98 (57). Jeyakumar et al. tested five modern deep learning architectures for multi-class classification of the HAM10000 dataset, exemplifying the classification performance of the GoogleNet architecture (58). The novel architecture yielded an AUC of 0.98, granting near-perfect predictive capabilities. Furthermore, lightweight networks, such as the DeepSkinNet by Abhiram et al. have demonstrated exceptional multi-class classification performance (59). Their novel network yielded a testing classification accuracy of 0.9734 while having significantly fewer parameters than an industry-standard AlexNet. We are motivated to further test the capabilities of integrating demographic data into modern DL architectures in the hopes of augmenting performance through ensemble learning.

Reviewer B

In this manuscript, a binary classification approach for skin lesions using dermatoscopic images is presented. The authors propose the use of a demographic machine learning risk stratified to inform the decisions of convolutional neural networks for the classification of melanoma. The paper is relatively well written and technically sound, though the related work section provides a deficient analysis of related approaches in the literature. More related papers such as [1,2,3] should be discussed.

[1] Hierarchy-aware contrastive learning with late fusion for skin lesion classification; Computer Methods and Programs in Biomedicine, 2022.
[2] Melanoma detection using adversarial training and deep transfer learning; Physics in Medicine & Biology, 2020.
[3] MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge; IEEE Transactions on Medical Imaging, 2021.

My general comments are summarized as follows:

Comment 1: More related papers such as [1,2,3] should be discussed.
Reply 1: We agree that more recent papers should be discussed to show the progression of deep learning performance for skin lesion classification. The purpose of this study is to establish a baseline of improvement in performance with standard models. In future papers, we will implement more recent network architectures.
Change in text: Following Line 159, we included the summarized results of the two studies.

Comment 2: More recent DL network architectures such as EfficientNet and HybridNets should be included.
Reply 2: In our Discussion section, we highlighted several studies utilizing state-of-the-art network architectures for classification of skin lesions in the HAM10000 dataset.
Change in text: Following Line 622 of the original manuscript, we included the following paragraph: This study served as a baseline in demonstrating the improvement of classification performance of melanoma with the inclusion of a machine learning risk score as an additional feature in deep ensemble training. We utilized historically-stable classification networks to test our hypothesis. EfficientNet (56) is a novel network architecture that has demonstrated impressive capabilities in classification challenges. Ali et al. demonstrated its diagnostic accuracy with EfficientNet base networks and modified derivatives yielding AUCs ranging from 0.96-0.98 (57). Jeyakumar et al. tested five modern deep learning architectures for multi-class classification of the HAM10000 dataset, exemplifying the classification performance of the GoogleNet architecture (58). The novel architecture yielded an AUC of 0.98, granting near-perfect predictive capabilities. Furthermore, lightweight networks, such as the DeepSkinNet by Abhiram et al. have demonstrated exceptional multi-class classification performance (59). Their novel network yielded a testing classification accuracy of 0.9734 while having significantly fewer parameters than an industry-standard AlexNet. We are motivated to further test the capabilities of integrating demographic data into modern DL architectures in the hopes of augmenting performance through ensemble learning.

Comment 3: The quality of the figures should be improved as some labels in

the futures are quite small and virtually unreadable.
Reply 3: We agree that Figure 8 (now Figure 9) was difficult to read due to the labels being too small to be read. To the best of our knowledge, all other figures were readable, but were modified to increase readability without affecting the aesthetics of the paper.
Change in text: Figure 8 was recaptured to yield zoomed in versions of the figures. All other figures were magnified to fill the most space without compromising organization of the manuscript.

Comment 4: Eq. (3) seems incomplete/incorrect. It needs to be corrected.
Reply 4: In the Word document submitted, the equation was not formatted correctly. We altered the equation to be an image and will submit this updated form to demonstrate the complete t-statistic equation.
Change in text: Equation 3 is an image with the correct equation to compute the t-statistic.

Comment 5: The runtime analysis needs to be discussed.
Reply 5: Runtime is an essential component in determining the practicality of network performances. We included runtime analysis by averaging the runtime of each network's three-fold cross-validation.
Change in text: At the beginning of Results>Evaluation of performances between deep and ensemble networks, we included a paragraph and new figure (Figure 6) to represent the differences in runtime between N_i and E_i.

Comment 6: It is not clear how the choice of the hyperparameter k of the k-fold CV would affect the overall performance of the different models.
Reply 6: The description of how *k* affects overall performance was not explicitly mentioned. We revamped this description to include the benefits and limitations of varying *k* and the associated impact on classification tasks.
Change in text: Methods and Materials>ML Algorithms>*k*-Nearest Neighbors paragraph rewritten to:
The *k*-Nearest Neighbors (KNN) is a supervised algorithm that excels at data classification and discriminant analysis when there is little prior knowledge of the database of interest. When plotted, categorical training data can cluster into discernable groups. Testing data, with no assigned output, can be classified based on the number of nearest neighbors to the datapoint. *k* is the hyperparameter that defines the number of nearest neighbors to classify a point of interest. With *k*=1, the unknown datapoint is grouped into the category with the nearest neighbor. Higher *k* allows for more neighboring datapoints to be included in the classification task. The unknown datapoint is classified by the category with the greatest amount of nearest neighbors. However, overly-elevated *k* can result in the over-representation of a category with few samples. KNN is an efficient algorithm for small-scale data but has a high sensitivity to outliers and large categories (35).