APM ANNALS OF PALLIATIVE MEDICINE
AN OPEN ACCESS JOURNAL FOR HIGH-QUALITY RESEARCH IN PALLIATIVE MEDICINE

IMPACT FACTOR 1.681

Peer Review File

Article information: http://dx.doi.org/10.21037/apm-20-2625

Reviewer Comments

The paper presents a model to differentiate COVID-19 from common pneumonia. The proposed model uses a human-audited/deep learning-based segmentation techniques and four classic classifiers with handcrafted features such as L1 regularization, Lasso, Ridge and Ztest. It was experimentally tested in a dataset with CT-scan images from 103 patients. The paper is interesting and somehow novel, however I raise the following issues that should be addressed before publication:

Comment 1: The dataset is very short and balanced, which is not a real-world scenario. Maybe the authors could grab more CT-scan images and even use data augmentation techniques to increase the dataset size.

Reply 1: Thanks the reviewer for the valuable comment. The experiment was carried out at the beginning of March last year, which included all confirmed COVID-19 cases from two designated hospitals in Nanjing. As Nanjing was not the epi-center during the pandemic, and the situation in China was soonly controlled by the end of April, 2020. Therefore, we incorporated the deep learning-based segmentation scheme and radiomics features to perform the classification tasks in the current study. It is worth noting that data augmentation techniques were indeed used while building the deep learning model. Conventional data augmentation techniques, such as rotation, zooming, cropping and flipping was used as a part of data-processing to improve the segmentation accuracy of the deep learning model.

Comment 2:

- Why these four classic ML algorithms were chosen? E.g., why not Random Forest or other algorithms? Moreover, why not use Deep learning-based classifiers such as CNN?
- The size of the dataset may not be big enough for a Deep learning-based technique, but that shows a weakness of the study.

Reply 2: Thanks the reviewer for the valuable comments. Since there are four feature selection methods used in the study, the number of classifiers will cause the modeling experiment to increase geometrically. For this reason, we have selected four widely used and representative machine learning methods (Image-Based Cardiac Diagnosis With Machine Learning: A Review), including logistic regression (LR), support vector machine (SVM), multilayer perceptron (MLP), and eXtreme Gradient Boosting (XGboost). Among all methods, XGBoost represents a tree model, which has proven remarkable in many practical applications (Mekov, E., Miravitlles, M., & Petkov, R. (2020). Artificial intelligence and machine learning in respiratory medicine. Expert review of respiratory medicine, 14(6), 559-564.). Also, MLP represents a neural network model, which could be seen as a substitute of CNN. It is worth nothing that CNN mainly performs convolution operations on images directly to extract features, which is not suitable for omics features. Therefore, MLP is chosen as the representative of neural network for application in the study.

Exactly as the reviewer suggested here, the size of the dataset limits the use of deep CNN to perform the classification task, which may lead to over- or under- fitting of the model.

Comment 3: The authors have used only the AUC metric. This metric has an optimistic point of view, since it can mask the real performance of the model. The authors should also use metrics such as F-score or AUPRC.

Reply 3: Thanks the review for the valuable comments. We have added other performance measurements, including sensitivity, specificity, accuracy, F1-score, and area under the precision-recall curve (AUPRC) in the Results section.

Changes in the text:

Sixteen models were established in this study. For each model, the evaluation metrics presented here were AU-ROC, AU-PRC, sensitivity (SEN), specificity (SPEC), F1-score and accuracy (ACC). Table 4 summarized the varying performance for each classifier across different feature selection methods. Among all modes, Lasso regression yielded higher AU-ROC values for all used classifiers. Specifically, MLP classifier obtained the highest AU-ROC of 0.989 (95%CI: 0.962 - 1.000). The results indicated that LASSO combined with MLP classifiers was the best-performing model with an highest accuracy of 96.3%, sensitivity of 95.7%, specificity of 98.4%, and AU-PRC of 0.942 (Figure 3).

Table 4 Summary of the efficacy pf classifiers and feature selection methods with average predictive performance taken over the 5-fold validation.

Feature Selection	Classifiers	AU-ROC	AU-PRC	SEN	SPEC	F1-Score	ACC
L1 Regularization	LR	0.928	0.820	0.935	0.885	0.896	0.897
	MLP	0.941	0.804	0.935	0.869	0.887	0.888
	SVM	0.956	0.788	0.848	0.902	0.866	0.869
	XGboost	0.940	0.772	0.957	0.820	0.869	0.869
LASSO	LR	0.982	0.921	0.957	0.967	0.952	0.953
	MLP	0.989	0.942	0.957	0.984	0.962	0.963
	SVM	0.985	0.904	0.891	0.984	0.932	0.935
	XGboost	0.957	0.772	0.957	0.820	0.869	0.869
RIDGE	LR	0.957	0.836	0.935	0.902	0.905	0.907
	MLP	0.974	0.880	0.848	0.984	0.912	0.916
	SVM	0.966	0.817	0.870	0.918	0.885	0.888
	XGboost	0.963	0.808	0.913	0.885	0.886	0.888
ZTEST	LR	0.939	0.793	0.913	0.869	0.877	0.879
	MLP	0.948	0.769	0.870	0.869	0.857	0.860
	SVM	0.908	0.752	0.891	0.836	0.849	0.850
	XGboost	0.929	0.784	0.870	0.885	0.867	0.869

Abbreviations: LR, logistic regression; MLP, multi-layer perceptron; SVM, support vector machine; Xgboost, eXtreme gradient boosting. AU-ROC, area under the receiver characteristic curve; AU-PRC, area under the precision-recall curve; ACC, accuracy.

Comment 4: The English should be revised (maybe ask for a professional help).
- E.g.: "which composed of..." should be "which is composed of..."

Reply 4: Thanks the review for the comments. We have asked for a professional help on the https://editing.amegroups.cn/. The journal name is the Annals of Palliative Medicine and the order ID is AESE20210216. The manuscript shall return soon.

Comment 5: Although this is a hot topic nowadays, there is no related work section. The authors should present a related work.

- The paper is missing future work directions. Tip: The authors can use textural descriptors (LBP, RLBP,..) together with the radiomics features in a fusion schema.

Reply 5: Thanks the review for the comments. We have included several future work directions in the revised manuscript. We attempted to expand the dataset size using data from external institutions. Also, we would like to validate the possibility to implement the proposed model in the prediction of diffuse pulmonary diseases, such as pulmonary alveolar proteinosis and interstitial pneumonia. Moreover, novel edge-texture features such as local binary pattern (LBP) and robust LBP (RLBP) could be further investigated with the radiomics features to improve the discriminative capacity for disease recognition.

Change in the text:

Future work will extend this approach to a bigger dataset to further refine this techonology for diffuse pulmonary diseases, such as pulmonary alveolar proteinosis and interstitial pneumonia. Also, fusion of radiomics features and local binary pattern (LBP)-based edge-texture features may have a potential to handle the classification task with limited dataset in medical imaging.