Combining Habitat Analysis and Autoencoder Neural Network for Feature Extraction to Predict COVID-19 Infection in CT Images

Thinh Huynh[1,2], Sheeba J. Sujit[2], Maliazurina Saad[2], Kristy K. Brock[2], Carol C. Wu[3], Caroline Chung[4], Jia Wu[2]

Department of Chemistry, The College of Wooster[1], Wooster, OH
Department of Imaging Physics[2], Department of Thoracic Imaging[3], Department of Radiation Oncology[4], The University of Texas MD Anderson Cancer Center, Houston, TX

Background

For only over two years since its emergence, COVID-19 pandemic has caused an immerse burden on the global healthcare system and claimed over 4 million lives. Early detection and diagnosis are the keys to improve treatment outcome, control disease spreading, and alleviate logistical burden for healthcare facilities. Previous studies on incorporation of artificial intelligence-based methodologies have shown promising results for COVID-19 screening on medical images. However, one drawback of these studies is to not incorporate radiomics despite its powerful diagnostic and prognostic power in disease screening. In this study, we propose a combined approach integrating deep learning and radiomics for COVID-19 detection in CT scans obtained from different patient cohorts.

Objective

We utilized habitat analysis and an autoencoder neural network as feature extraction tools for machine learning classification to predict COVID-19 infection in CT scans.

Method

We analyzed the CT scans of 240 cancer patients diagnosed with COVID-19 at MD Anderson and 227 patients retrieving from the RSNA International COVID-19 Open Radiology Database (RICORD). RICORD database comprises both COVID-19 positive (n = 110) and negative (n = 117) patients.

Figure 1 shows the overall study design. Lung masks for COVID-19 negative and positive scans were auto-segmented using the pre-trained convolutional neural networks (CNN) U-net R231 and U-net R231CovidWeb, respectively. We performed habitat analysis to identify subregions within the lung. This partitioning method comprises of 2 clustering steps. First, at the individual level, the lung region was oversegmented into superpixels from 2 precalculated features: CT number and entropy of CT number. Then, at the population level, consensus clustering was performed on generated superpixels, whose similar image phenotypes indicate same habitats, to partition individual lung scans into different habitat regions. Here, the clustering number k = 3 was used, indicating that 3 different habitats were identified for each scan. Next, the spatial cooccurrence statistics among different habitats were generated as multiregional spatial interaction (MSI) matrix, which can be used as potential features for classification.

Method (Cont.)

For deep feature extraction, we utilized a 3D autoencoder built based on an U-net CNN architecture. The CT scans overlaid with corresponding lung masks were used as inputs. The inputs were split into 3 subsets for training (70%), validation (10%), and testing (20%). Mean squared error (MSE) loss function was used for training the autoencoder. After training and output quality check, we extracted the deep features from the bottleneck layer.

Multiple methods were utilized for feature selection and dimensionality reduction including pearson's correlation coefficient (PCC) and principal component analysis (PCA). The finalized features were used to train multiple machine learning (ML) classifiers with K Fold cross validation (K = 10). The outputs were COVID-19 negative and positive. We assessed the performance of each model using different evaluation metrics such as sensitivity, specificity, and confusion matrix.

Result

From habitat analysis, we identified 3 subregions with distinct image phenotypes (**Fig. 2**) and extracted 25 features including 3 habitat volume and 22 MSI features. Performing pair-wise PCC of 25 habitat features, we selected out 6 features for machine learning classification to avoid multicollinearity (**Fig. 3**). Fitting these features into different ML classifers yielded promising results with many evaluation metrics such as accuracy, recall, and AUC_ROC over 0.7 (Table 1). Further analysis indicated that quadratic discriminant

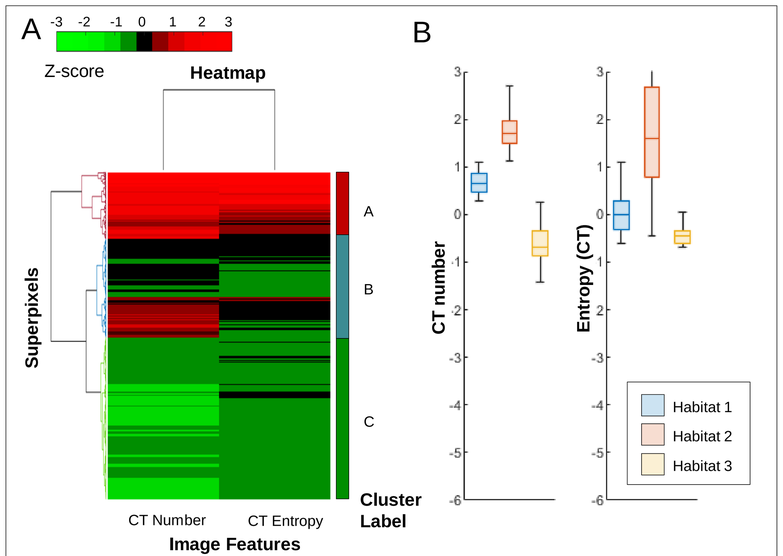

Figure 2. Different habitats with distinct image phenotype generated from habitat analysis. (**A**) Heatmap of superpixels by image features. (**B**) Boxplot of image features by habitats.

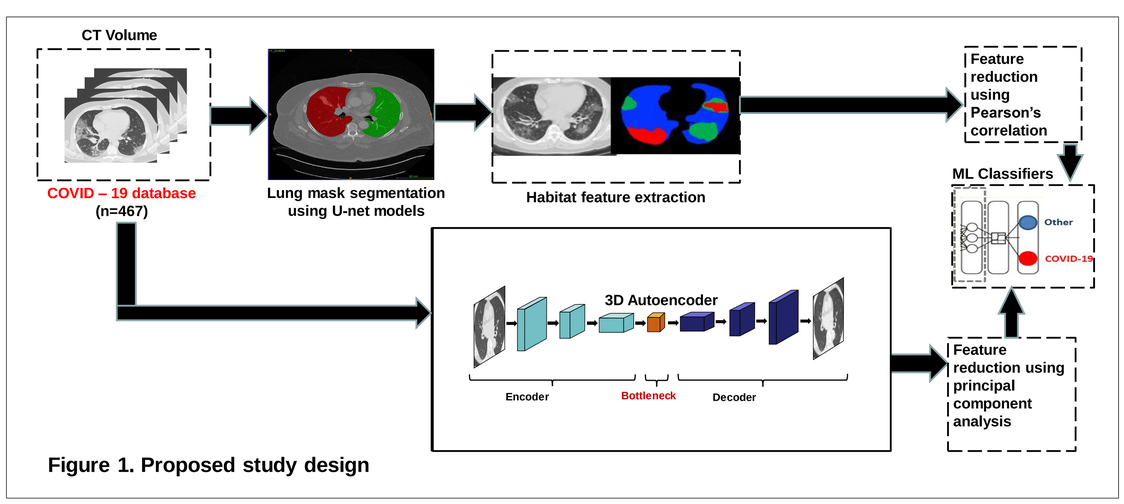

Figure 1. Proposed study design

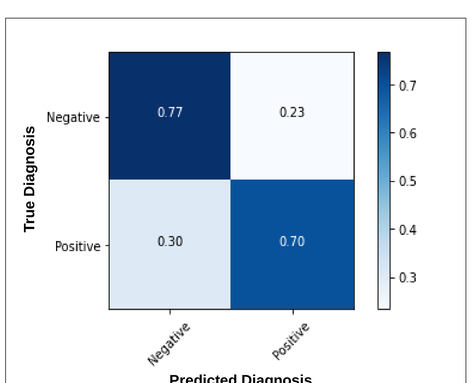

Figure 4. Confusion matrix of the performance of quadratic discriminant analysis with K Fold cross validation.

Results (Cont.)

analysis was the best classifier for COVID-19 infection using habitat-based features (**Fig. 4**)

Table 1. Summary Table of Different Classifiers' Performace

Model	Accuracy	Precision	Recall	F1 Score	AUC_ROC
Logistic Regression	0.719611	0.592803	0.515379	0.680781	0.728914
Support Vector Machine	0.770953	0.385476	0.500000	0.671262	0.716162
Random Forest	0.760685	0.73308	0.608081	0.745369	0.805088
Quadratic Discriminant Analysis	0.712026	0.728685	0.730859	0.712769	0.798232

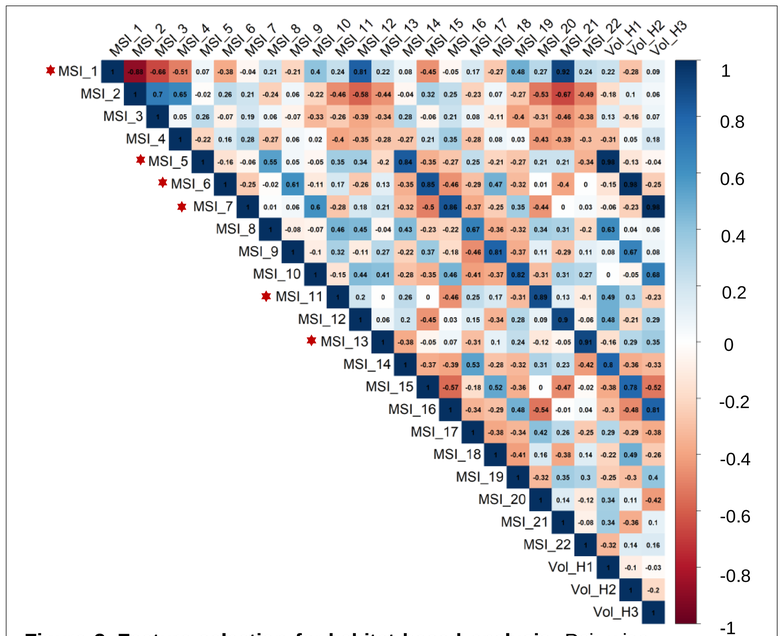

Figure 3. Feature selection for habitat-based analysis. Pair-wise Pearson's correlation coefficient matrix of all habitat-based features (red stars indicate feature used for classification).

Training the autoencoder with batch size of 6 and 150 epochs yielded good results with low train and validation loss (Fig. 5). After training, we extracted 1024 deep features from the bottleneck layer of the model.

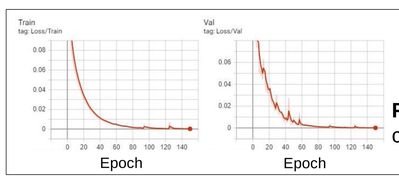

Figure 5. Train and validation loss curves of 3D autoencoder

Conclusions

These results indicate the diagnostic power of habitat analysis as a feature extraction tool for COVID-19 diagnosis in CT scans. A strength of our study is diverse patient cohorts However, a weakness of this study is the imbalanced data between COVID-19 positive and negative patients, which possibly resulted in low specificity and recall values during classification. Future works include dimensionality reduction of extracted deep features, performing classification on those features, and performing classification on both habitat-based and deep features to test incorporation of both types of features would yield better prediction than each type alone.

References

Tsai, E., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A.,Kalpathy-Cramer, J., Shen, J.,Hafez, M.A.F., John, S., Rajiah, P., Pogatchnik, B.P., Mongan, J.T., Altinmakas, E., Ranschaert, E., Kitamura, F.C., Topff, L., Moy, L., Kanne, J.P., & Wu, C. (2021). Data from Medical Imaging Data Resource Center (MIDRC) - RSNA International COVID Radiology Database (RICORD) Release 1c - Chest x-ray, Covid+ (MIDRC-RICORD-1c). The Cancer Imaging Archive. DOI: https://doi.org/10.7937/91ah-v663.

Wu, J., Gensheimer, M. F., Zhang, N., Guo, M., Liang, R., Zhang, C., Fischbein, N., Pollom, E. L., Beadle, B., Le, Q.-T., & Li, R. (2019). Tumor subregion evolution-based imaging features to assess early response and predict prognosis in oropharyngeal cancer. *Journal of Nuclear Medicine*, *61*(3), 327–336. https://doi.org/10.2967/jnumed.119.230037