[BIOINFO 2023]

COVID-19 severity risk prediction with machine learning approaches

Se-Hyun Chang, Sanghyun Lee, Sang Cheol Kim, Ju-Hee Lee, Hye-Yeong Jo*
Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, Korea National Institute of Health, Korea Disease Control and Prevention Agency
**Corresponding author: jhy1227@korea.kr*

After the coronavirus disease 2019 (COVID-19) Pandemic, one of the major challenges in treating patients with COVID-19 is to predict the severity of disease since more intensive surveillance and appropriate therapy can be considered in patients at high risk of COVID-19 progression in clinic. This study aimed to develop an machine learning (ML) model to predict the COVID-19 severity and identify clinical features associated to COVID-19 severity.

Here, we collected laboratory testing and clinical data for 620 participants, including 459 COVID-19 patients and 161 healthy controls. Out of a toal of 220 factors in clinical data, 34 features in demographics, comorbidities, symptoms on admission, and vital signs categories were considered. In addition, 37 variables were used in laboratory testing, including r-GTP, hematocrit, white blood cell (WBC), and red blood cell (RBC). After feature importance analysis was carried out with the logistic regression based on the forward stepwise selection, the machine learning classifiers with XGboost, LightGBM, RandomForest and Ensemble model with the top-ranked features were applied to predict COVID-19 severity. Subsequently, we evaluated the Odds Ratios (OR) to investigate whether the features of the best model were statistically significant for severity clsssification and the Shapley additive explanations (SHAP) values to explain how each feature contributes to the COVID-19 severity prediction models. We expect that our results can give insights into early screening for diagnosing COVID-19 severity, which, in turn, assist in further retrospective research for newly emerged infectious diseases.