**Supplementary Material**

Clinical Decision Support System for Severity Early Triaging of COVID-19 Severity

|  |  |
| --- | --- |
| **Table of Contents** | **Page (s)** |
| **Supplementary Table 1:** Baseline characteristics in severe and non-severe cohorts | 2 |
| **Supplementary Table 2:** Discriminative performance in the internal validation | 3 |
| **Supplementary Table 3:** Comparison in Net Reclassification Improvements among prediction models | 3 |
| **Supplementary Table 4:** Performance of prediction models on omicron-variant cases in the external validation | 4 |
| **Supplementary Figure 1.** Six feature engineering methods based on two ML algorithms and four feature importance measures to generate candidate features | 5 |
| **Supplementary Figure 2.** Distribution of each feature in the final model across centers | 6-7 |
| **Supplementary Figure 3.** Empirical distributions of AUROC scores in internal and external validations | 8 |
| **Supplementary Figure 4.** Empirical distributions of AUROC scores for ML algorithms across different class | 9 |
| **Supplementary Figure 5.** Reclassification Plots | 10 |
| **Supplementary Figure 6.** Average impact of selected features on severity prediction | 11 |
| **Supplementary Figure 7.** Characterization of patient characteristics for each variant based on clinical and laboratory characteristics | 12 |
| **Supplementary Text:** Study setting and design | 12 |
| **Supplementary Text:** Model Development and Evaluation with Feature-set Selection | 13 |
| **Supplementary Text:** Model Performance Visualization and Feature Interpretation | 14 |
| **Supplementary Text:** Model Transportability to New Variant Dominant Period | 14 |
| **Supplementary Text:** Statistical analysis | 14 |
| **Supplementary Text:** Comprehensive list of abbreviations and their definitions used throughout the study | 15 |

**Supplementary Table 1: Baseline characteristics in severe and non-severe cohorts**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Severe Cohort** | **Non-severe Cohort** | **ASMD** |
| **(N: 839)** | **(N: 5,106)** |
| **Patient Characteristics** | | | |
| Age (years) | 70 (60-75) | 55 (40-70) | 0·733 |
| Male Sex (%) | 499 (59·5%) | 2,388 (46·8%) | 0·255 |
| Diagnosed Date | -2 (-5-0) | 0 (-1-0) | 0·591 |
| **Comorbidities (%)** | | | |
| Hypertension | 462 (55·1%) | 1,652 (32·4%) | 0·462 |
| Diabetes Mellitus | 345 (41·1%) | 904 (17·7%) | 0·524 |
| CVD | 106 (12·6%) | 402 (7·9%) | 0·158 |
| Cancer | 75 (8·9%) | 402 (7·9%) | 0·038 |
| Others | 401 (47·8%) | 1,841 (36·1%) | 0·239 |
| **Clinical Symptoms (%)** | | | |
| Fever | 364 (43·4%) | 2,002 (39·2%) | 0·085 |
| Cough | 317 (37·8%) | 2,306 (45·2%) | 0·15 |
| Sputum | 181 (21·6%) | 1,321 (25·9%) | 0·101 |
| Dyspnea | 487 (58·0%) | 831 (16·3%) | 0·902 |
| Myalgia | 92 (11·0%) | 1,296 (25·4%) | 0·381 |
| Sorethroat | 59 (7·0%) | 1,083 (21·2%) | 0·42 |
| Loss of Sensor | 31 (3·7%) | 299 (5·9%) | 0·102 |
| GI symptom | 53 (6·3%) | 419 (8·2%) | 0·073 |
| **Vital Sign** | | | |
| BT (℃) | 36·7 (36·4-37·4) | 36·6 (36·3-37·2) | 0·187 |
| SBP (mmHg) | 130 (116-145) | 129 (116-140) | 0·056 |
| DBP (mmHg) | 73 (65-83) | 80 (70-88) | 0·447 |
| PR (%) | 85 (73-98) | 84 (75-95) | 0·116 |
| RR (%) | 22 (20-25) | 20 (18-20) | 0·911 |
| SpO2 (%) | 96 (93-98) | 97 (96-98) | 0·705 |
| **Blood biochemistry** | | | |
| WBC (103/µL) | 7·2 (5·0-10·2) | 5·1 (4·0-6·5) | 0·693 |
| ANC (103/µL) | 5·9 (3·8-8·8) | 3·3 (2·3-4·8) | 0·971 |
| ALC (103/µL) | 0·8 (0·5-1·1) | 1·3 (0·9-1·7) | 0·497 |
| PLT (103/µL) | 178 (131-237) | 203 (162-249) | 0·295 |
| CRP (mg/dL) | 9·1 (3·7-15·0) | 0·8 (0·2-3·5) | 0·248 |
| LDH (U/L) | 496 (364-723) | 295 (213-407) | 0·726 |

Data are median (IQR) or n (%). ASMD=absolute standardized mean difference. CVD=cardio-vascular disease. GI= gastrointestinal. BT=body temperature. SBP=systolic blood pressure. DBP=diastolic blood pressure. PR=pulse rate. RR=respiratory rate. SpO2=saturation of peripheral oxygen. WBC= white blood cell. ANC=absolute neutrophil count. ALC=absolute lymphocyte count. PLT= platelet count. CRP=c-reactive protein. LDH=lactate dehydrogenase.

**Supplementary Table 2:** **Discriminative performance in the internal validation**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predictive Measures** | | | | | | | |  |
| **Model** | **AUROC** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **LRP** | **LRN** | **DOR** | **Cut-Off** |
| DNN | 0·891 (0·889-0·892) | 0·852 (0·844-0·860) | 0·808 (0·796-0·819) | 46·58 (45·24-47·93) | 96·59 (96·45-96·73) | 4·57 (4·30-4·84) | 0·18 (0·17-0·19) | 25·96 (24·22-27·71) | 0·171 |
| MLR | 0·887 (0·886-0·888) | 0·847 (0·845-0·849) | 0·792 (0·789-0·795) | 44·40 (43·99-44·81) | 96·40 (96·35-96·45) | 4·17 (4·08-4·26) | 0·19 (0·19-0·20) | 21·93 (21·35-22·52) | 0·164 |
| RF | 0·894 (0·893-0·896) | 0·848 (0·831-0·866) | 0·807 (0·791-0·823) | 46·47 (44·98-47·96) | 96·51 (96·17-96·85) | 4·55 (4·27-4·82) | 0·19 (0·17-0·21) | 25·40 (23·75-27·05) | 0·193 |
| XGB | 0·878 (0·875-0·880) | 0·833 (0·813-0·853) | 0·784 (0·772-0·796) | 43·34 (42·36-44·33) | 96·08 (95·68-96·47) | 4·01 (3·82-4·20) | 0·21 (0·19-0·23) | 19·54 (18·05-21·02) | 0·105 |
| GBM | 0·879 (0·877-0·882) | 0·837 (0·820-0·853) | 0·791 (0·775-0·808) | 44·66 (42·94-46·37) | 96·23 (95·94-96·51) | 4·27 (3·95-4·60) | 0·20 (0·19-0·22) | 21·74 (20·75-22·74) | 0·010 |
| SVM | 0·834 (0·830-0·837) | 0·782 (0·771-0·794) | 0·833 (0·826-0·841) | 48·15 (47·20-49·09) | 95·20 (95·00-95·40) | 4·88 (4·69-5·08) | 0·26 (0·25-0·27) | 19·19 (17·97-20·42) | 0·105 |

All results, 95% CIs in parentheses, were computed from five iterations of stratified five-fold cross validations. Youden’s Index was used to determine optimal cut-off point. DNN=deep neural network. MLR=multivariable logistic regression. RF=random forest. XGB=eXtreme gradient boosting. GBM=gradient boosting machine. SVM=support vector machine. AUROC=area under receiver operating characteristic curve. PPV=positive predictive value. NPV=negative predictive value. LRP=likelihood ratio positive. LRN=likelihood ratio negative. DOR=diagnostic odds ratio.

**Supplementary Table 3. Comparison in Net Reclassification Improvements among prediction models**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **NRI** | **NRI+ (Severe)** | **NRI- (Non-Severe)** |
| **DNN vs MLR** | 0·0053 | -0·0458 | 0·0511 |
| **DNN vs RF** | 0·0137 | -0·0153 | 0·0290 |
| **DNN vs XGB** | 0·0614 | 0·0000 | 0·0614 |
| **DNN vs GBM** | 0·0394 | 0·0076 | 0·0317 |
| **DNN vs SVM** | 0·0290 | 0·0076 | 0·0214 |

NRI=net reclassification improvements. DNN=deep neural network. MLR=multivariable logistic regression. RF=random forest. XGB=eXtreme gradient boosting. GBM=gradient boosting machine. SVM=support vector machine.

**Supplementary Table 4. Performance of prediction models on omicron-variant cases in the external validation**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predictive Measures** | | | | | | | | |  | |
| **Model** | **AUROC** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **LRP** | **LRN** | **DOR** | **Cut-Off** | |
| Ensemble | 0·903 (0·897-0·910) | 0·897 (0·881-0·913) | 0·831 (0·814-0·849) | 54·56 (52·28-56·84) | 97·65 (97·29-98·01) | 6·70 (6·08-7·33) | 0·12 (0·10-0·14) | 56·22 (49·83-62·61) | 0·197 | |
| All-variants | 0·834 (0·825-0·844) | 0·876 (0·856-0·896) | 0·720 (0·701-0·740) | 41·22 (39·06-43·38) | 96·76 (96·30-97·22) | 4·07 (3·34-4·80) | 0·16 (0·14-0·19) | 29·00 (25·89-32·11) | 0·036 | |
| Omicron-only | 0·813 (0·800-0·826) | 0·912 (0·895-0·928) | 0·697 (0·683-0·712) | 39·14 (37·44-40·85) | 97·67 (97·30-98·04) | 3·37 (3·02-3·73) | 0·12 (0·10-0·14) | 33·79 (30·40-37·19) | 0·391 | |

Final model is Deep Neural Network (DNN) model with 11 important variables. All results, 95% CIs in parentheses, were computed from bootstrapping external validation cohorts 100 times with replacements. Youden’s Index was used to determine optimal cut-off point. AUROC=area under receiver operating characteristic curve. PPV=positive predictive value. NPV=negative predictive value. LRP=likelihood ratio positive. LRN=likelihood ratio negative. DOR=diagnostic odds ratio.

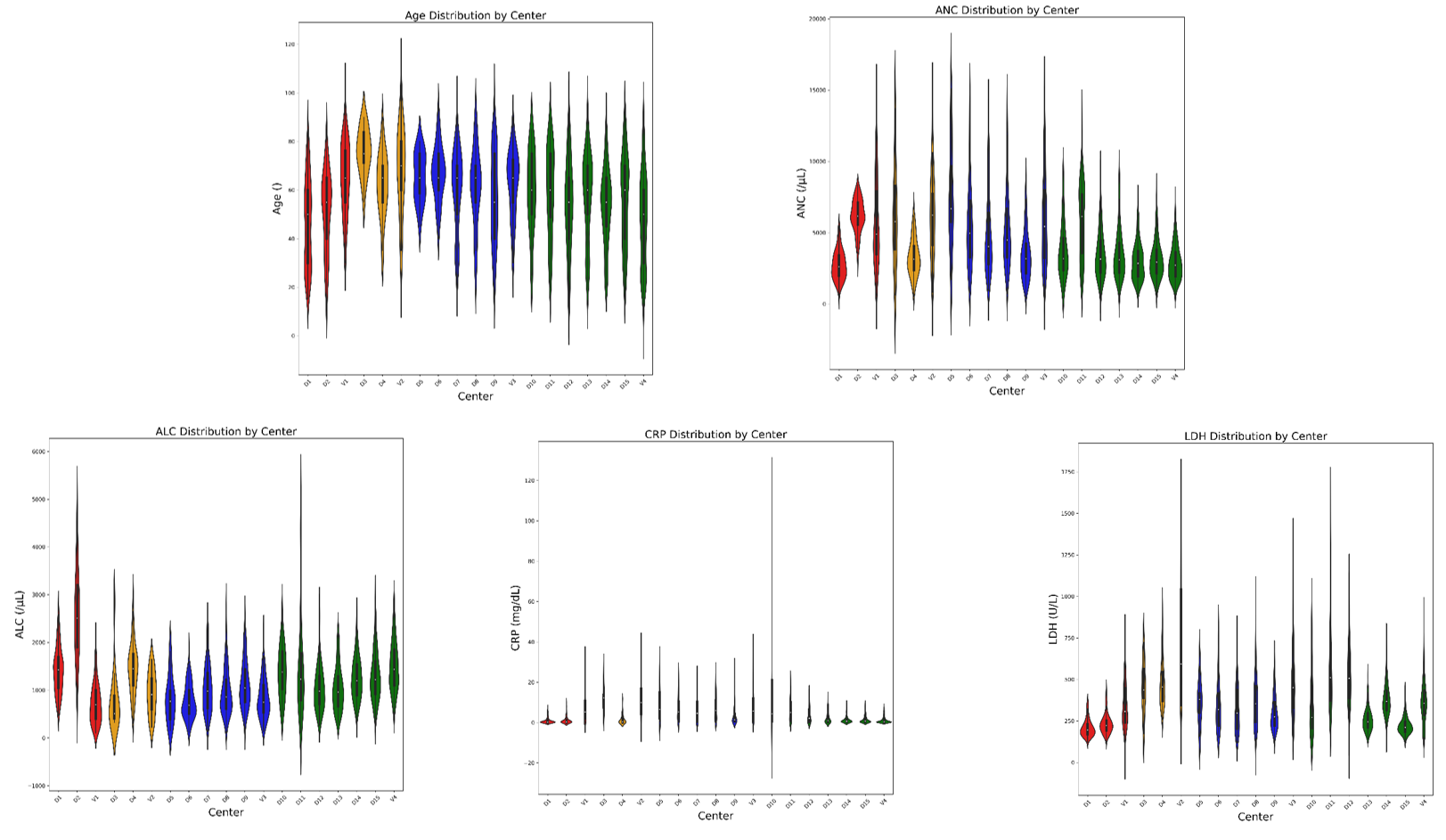
**Supplementary Figure 1. Six feature engineering methods based on two ML algorithms and four feature importance measures to generate candidate features**

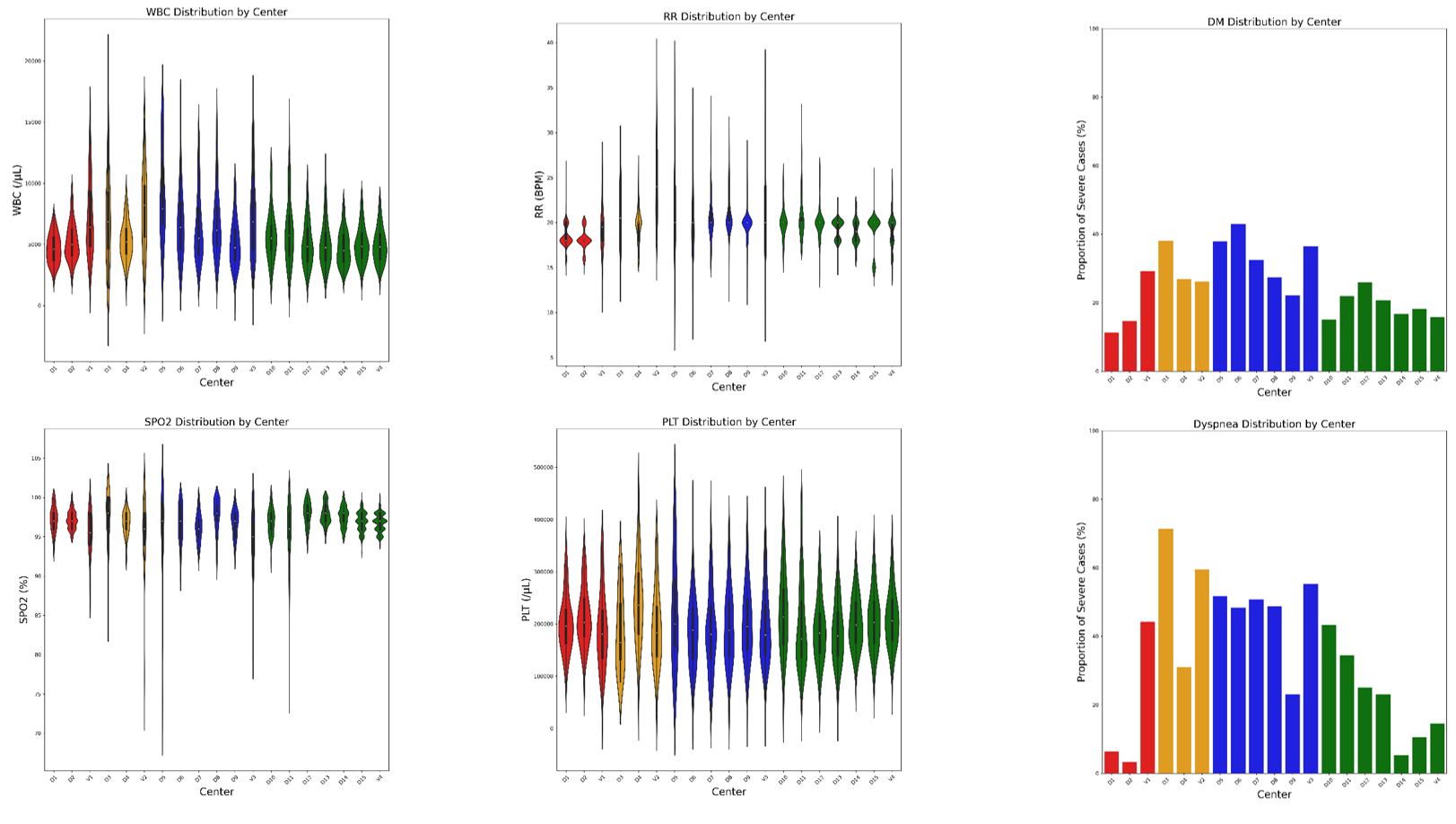
텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

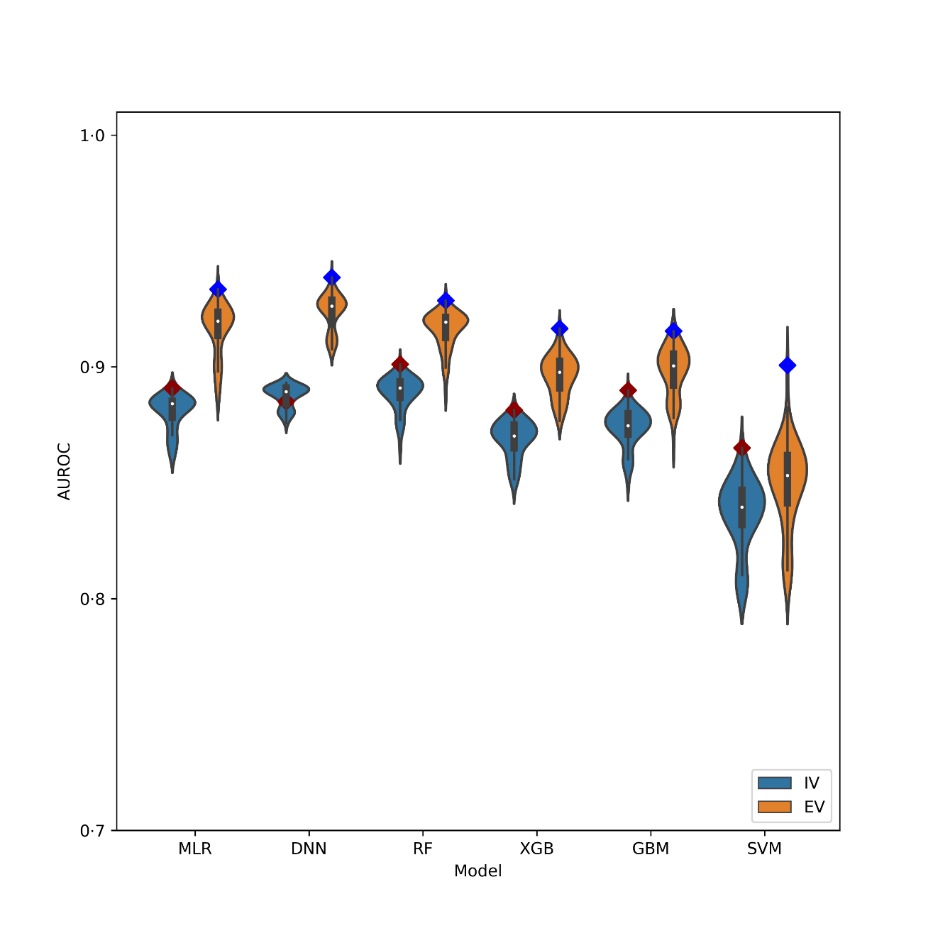
MDI=mean decrease in impurity. CVD=cardio-vascular disease. GI= gastrointestinal. BT=body temperature. SBP=systolic blood pressure. DBP=diastolic blood pressure. PR=pulse rate. RR=respiratory rate. SpO2=saturation of peripheral oxygen. WBC= white blood cell. ANC=absolute neutrophil count. ALC=absolute lymphocyte count. PLT= platelet count. CRP=c-reactive protein. LDH=lactate dehydrogenase.

**Supplementary Figure 2. Distribution of each feature in the final model across centers**



Z-score less than or equal to 3 were included to remove outliers. ANC=absolute neutrophil count. ALC=absolute lymphocyte count. CRP=c-reactive protein. LDH=lactate dehydrogenase. WBC=white blood cell. SPO2=saturation of peripheral oxygen. RR=respiratory rate. PLT=platelet count.

**Supplementary Figure 3. Empirical distributions of AUROC scores in internal and external validations**



Empirical distributions of area under receiver operating characteristic curve (AUROC) scores were estimated for each prediction model with all 60 feature subsets based on the resamples produced during the internal (IV) and external validations (EV). MLR=multivariable logistic regression. DNN=deep neural network. RF=random forests. XGB=eXtreme gradient boosting. GBM=gradient boosting machine. SVM=support vector machine.

**Supplementary Figure 4. Empirical distributions of AUROC scores for ML algorithms across different class**

**(A) internal validation results. (B) external validation results.** Empirical distributions of area under receiver operating characteristic curve (AUROC) scores were estimated for three ML algorithms across class weights from 1 to 10 all 60 feature subsets during the internal (IV) and external validations (EV). MLR=multivariable logistic regression. DNN=deep neural network. RF=random forest

**A**

**텍스트, 스크린샷, 스케치이(가) 표시된 사진

자동 생성된 설명**

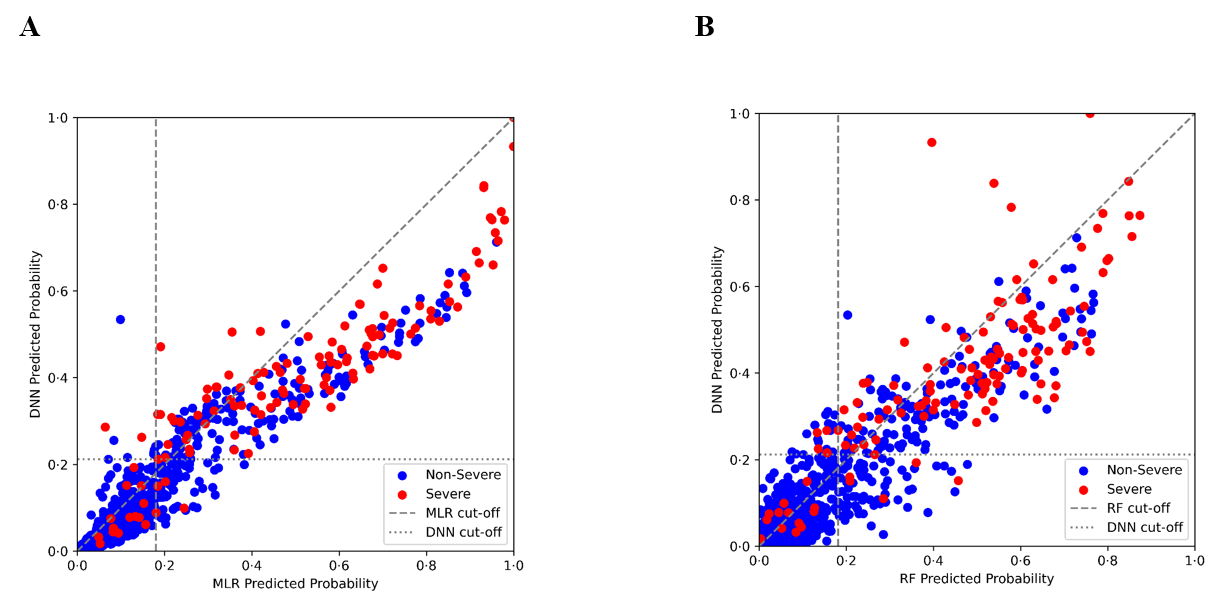
**B**

**스케치, 스크린샷, 그림이(가) 표시된 사진

자동 생성된 설명**

**Supplementary Figure 5. Reclassification Plots**

**(A) comparing the predictive performances between DNN and MLR models. (B) comparing the predictive performances between DNN and RF models.** DNN=deep nerual network. MLR=multivariable logistic regression. RF=random forests.



DNN=deep neural network. MLR=multivariable logistic regression. RF=random forests. \*Horizontal and vertical dotted lines represent Youden’s index-based cut-off point for each respective algorithm.

**Supplementary Figure 6. Average impact of selected features on severity prediction**

텍스트, 스크린샷, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

Used random forest-based model with 11 selected features to construct the summary plot above. Used TreeExplainer to compute Shapley Additive exPlanations (SHAP) values. CRP=c-reactive protein. LDH=lactate dehydrogenase. ALC=absolute lymphocyte count. RR=respiratory rate. ANC=absolute neutrophil count. WBC=white blood cell. PLT=platelet counts. SPO2=saturation of peripheral oxygen. DM=diabetes mellitus. \*Specific SHAP values are shown on the right of each feature names in the y-label.

**Supplementary Figure 7. Characterization of patient subgroups for each variant**

\*Darkred color indicates high concentration and lightgreen color indicates low concentration of each corresponding feature.

**도표, 라인이(가) 표시된 사진

자동 생성된 설명**

**Supplementary Text**

**Study setting and design**

***Source of Data***

The data sets were collected from 19 main referral hospitals across the nation: Ajou University Hospital, Chonnam National University Hospital, Chungang University Hospital, Chungbuk National University Hospital, Chungnam National University Hospital, Chungnam National University Sejong Hospital, Daegu Catholic University Medical Center, Gachon University Gil Medical Center, Jeonbuk National University Hospital, Kangwon National University Hospital, Keimyung University Dongsan Medical Center, Kyungpook National University Hospital, Pusan National University Hospital, Seoul Metropolitan Government Seoul National University Boramae Medical Center, Seoul National University Bundang Hospital, Samsung Medical Center, Seoul National University Hospital, Soonchunhyang University Hospital, and Uijeongbu Eulji Medical Center.

***Data Collection***

The clinical data included demographics (age and sex), comorbidities (hypertension, diabetes, cardiovascular disease, cancer, other), symptoms (fever, cough, sputum, myalgia, dyspnea, sore throat, loss of sensor, gastrointestinal symptom), and vital signs (body temperature, systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, saturation pulse oxygen). The laboratory data included white blood cell counts, absolute lymphocyte counts, absolute neutrophil counts, platelet counts, C-reactive protein, and lactate dehydrogenase. Lastly, real-time polymerase chain reaction results were labeled either positive or negative using a cycle threshold value of 40 as the cut-off.

**Model Development and Evaluation with Feature-subset Selection**

***Six ML-based algorithms***

We chose 6 representative ones from major streams of ML algorithms. The reason was two-fold. First, we wanted to identify robust features against the choice of ML algorithms. Second, the chance to find an optimal ML algorithm with very good performance would be increased by comprehensive search. The chosen ML algorithms included (1) deep neural network (DNN) among neural net-based algorithms; (2) random forest (RF) among bagging-based tree algorithms; (3) eXtreme gradient boosting (XGB), gradient boosting machine (GBM) among boosting-based tree algorithms; (4) support vector machine (SVM) among margin-based algorithms; and (5) multivariable logistic regression (MLR) among linear model-based algorithms.

***DNN Model Structure***

DNN model’s hyperparameters were empirically tuned and we achieved the best performance with the following parameters: a batch size of 32, an epoch of 50, an early stopper with patience level of 5, an Adam optimizer with learning rate of 0.01, a ReduceLROnPlateau learning rate scheduler—reduces when validation loss stops improving—with patience level of 1 and factor of 0.8, an input layer with the size of 128, two hidden layers with the size of 32 and 16, an output layer with sigmoid activation layer. Between hidden layers, the model was constructed with ReLU activation layer followed by a dropout layer with a 0.3 ratio and a batch normalization layer.

***Random States for Cross-Validation and Bootstrapping***

Different random states were used to partition the development cohort iteratively, and to obtain 100 bootstrap samples from the (external) validation cohort. In the meanwhile, random state parameters of all ML algorithms were fixed to a specific value (i.e., 0) that the randomness introduced in the algorithm was consistent across different trials.

***95% Confidence Intervals of Performance Metrics***

First, we estimated the values of each performance metric based on the five iterations of stratified five-fold cross-validation and the 100 bootstrap samples for internal and external validations, respectively. For each performance metric, we used these estimated values to generate the empirical distribution by assuming the normality with sample mean and standard deviation. Finally, 95% confidence intervals were estimated from the empirical distribution.

**Model Performance Visualization and Feature Interpretation**

***Calibration Plots***

The external cohort were segmented into 10 equal-sized bins according to the predicted probability for patients. For each bin, we calculated the observed risk as the ratio of the severe patients, and the predicted risk as the mean of predicted probabilities for all patients in the bin. The scatter plot of predicted and observed risks was constructed with 95% confidence bands for the predicted risk and the reference dashed line of identity function.

***DDRTree Plots***

We categorized the patients into subgroups using clinical characteristics. To graphically visualize subgroups, we condensed 11 dimensions into a two-dimensional tree structure. 4,787 patients formed the backbone of our DDRTree after removing the outliers from each continuous variable. These outliers were identified and eliminated by excluding cases beyond the 2% upper and lower thresholds following logarithmic transformation to each variable. During the tree generation, the fine-tuning of sigma (0.03) and gamma (0.007) parameters helped reveal the most distinctive structure. Lastly, the application of alpha shading (0.3) and jitter (0.04) illuminated the concentration of patients nested within particular zones of the tree.

***SHAP Summary Plots***

SHAP summary plot was utilized to interpret the contributions of individual features in the prediction model when classifying severe and non-severe cases. We modified the KernelExplainer method appropriately to compute SHAP values for features in DNN-based models while we used the TreeExplainer method with no modification for models built by tree-based ML algorithms.

***Decision Curve Analysis Plots***

Using each candidate threshold value of the predicted probability, we divided the patients into low- and high-risk groups. Then we assessed the potential net benefit of this risk grouping when applied to clinical decisions on intervention (i.e., intervention only for patients with high risk). The potential net benefit for this model-based intervention strategy was displayed across threshold values, along with two references, such as ‘always treat (i.e., intervention for all)’ and ‘never treat (i.e., intervention for none).’

**Model Transportability to New Variant Dominant Period**

According to the variant-dominant periods, we constructed variant-dominant subcohorts. First, the Omicron-dominant cohort consisted of the patients who were diagnosed with COVID-19 during the Omicron-dominant period: 337 patients from the development cohort (DC) and 112 patients from the validation cohort (VC). Then, we constructed the Alpha- and Delta-dominant DCs with 2,622 and 1,406 patients, respectively.

The final DNN-based model was called as the ‘all-variants’ model. We updated the final model with the Omicron-dominant DC and developed the ‘Omicron-only’ model’ as a result. We also built the Alpha-Omicron and Delta-Omicron models using the Alpha & Omicron-dominant DCs and the Delta & Omicron-dominant DCs, respectively. Then, we developed the ‘ensemble’ model by averaging prediction probabilities estimated in the Omicron-only, Alpha-Omicron, and Delta-Omicron models.

The ensemble model was externally validated and compared to the all-variants and Omicron-only models, using the Omicron-dominant VC.

**Statistical analysis**

All statistical analyses were performed using Python (Python Software Foundation, version 3·9) with the following libraries: ‘Pandas’ for dataset organization and cleansing; ‘Hyperopt’ for hyperparameter tuning; ‘Sci-kit Learn’ for data normalization, loading ML algorithms (MLR, RF, SVM, GBM), feature importance and predictive metrics computation, and stratified cross validation; ‘Xgboost’ for XGB model load; ‘Tensorflow’ for DNN model development; ‘SHAP’ for candidate features generation and individual feature interpretation; ‘Matplotlib’ for model result visualization (ROC curve analyses, calibration plot, decision curve analysis plot, and reclassification plot); ‘Seaborn’ for violin and bar plots. ‘dynamo’ for DDR Tree generation.

**Comprehensive list of abbreviations and their definitions used throughout the study**

|  |  |
| --- | --- |
| **Study Design and Reporting** | |
| DC | development cohort |
| VC | validation cohort |
| **Machine Learning and Statistical Methods** | |
| ASMD | absolute standardized mean difference |
| CDSS | clinical decision support system |
| CV | cross-validation |
| ML | machine learning |
| **Feature Interpretation Methods** | |
| DDRTree | discriminative dimensionality reduction via learning a tree |
| SHAP | shapley additive exPlanations |
| **Feature Engineering Methods** | |
| FEMs | feature engineering methods |
| RF-MDIFI | random forest-based mean decreases in Gini impurity feature importance |
| RF-PFI | random forest-based permutation feature importance |
| RF-Shapley | random forest-based Shapley values |
| XGB-BFI | eXtreme Gradient Boosting-based built-in feature importance |
| XGB-PFI | eXtreme Gradient Boosting-based permutation feature importance |
| XGB-Shapley | eXtreme Gradient Boosting-based Shapley values |
| **Machine Learning Algorithms** | |
| DNN | deep neural network |
| GBM | gradient boosting machine |
| MLR | multivariable logistic regression |
| RF | random forest |
| SVM | support vector machine |
| XGB | eXtreme gradient boosting |
| **Clinical and Laboratory Biomarkers** | |
| ALC | absolute lymphocyte counts |
| ANC | absolute neutrophil count |
| CRP | c-reactive protein |
| DM | diabetes mellitus |
| LDH | lactate dehydrogenase |
| PLT | platelet counts |
| RR | respiratory rate |
| SPO2 | saturation of peripheral oxygen |
| WBC | white blood cell counts |
| **Model Performance Metrics** | |
| AUROC | area under receiver operating characteristic curve |
| CI | confidence intervals |
| DCA | decision curve analysis |
| DOR | diagnostic odds ratio |
| ICI | integrated calibration index |
| LRP | likelihood ratio positive |
| LRN | likelihood ratio negative |
| NPV | negative predicted values |
| PPV | positive predicted values |