Tohoku J. Exp. Med., 2024, **262**, 33-41

Evaluation of Statistical Approaches in Developing a Predictive Model of Severe COVID-19 during Early Phase of Pandemic with Limited Data Resources

Tetsuya Akaishi,[1] Yasunori Tadano,[1] Yoshitaka Kimura,[2] Nobuo Yaegashi[2,3] and Tadashi Ishii[1]

[1]Department of Education and Support for Regional Medicine, Tohoku University, Sendai, Miyagi, Japan
[2]Department of Obstetrics and Gynecology, Tohoku University Hospital, Sendai, Miyagi, Japan
[3]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Miyagi, Japan

As evidence of risk factors for severe cases of coronavirus disease 2019 (COVID-19) was uncertain in early phases of the pandemic, the development of an efficient predictive model for severe cases to triage high-risk individuals represented an urgent yet challenging issue. It is crucial to select appropriate statistical models when available data and evidence are limited. This study was conducted to assess the accuracy of different statistical models in predicting severe cases using demographic data from patients with COVID-19 prior to the emergence of consequential variants. We analyzed data from 929 consecutive patients diagnosed with COVID-19 prior to March 2021, including their age, sex, body mass index, and past medical histories, and compared areas under the receiver operating characteristic curve (ROC AUC) between different statistical models. The random forest (RF) model, deep learning (DL) models with not too many neurons, and naïve Bayes model exhibited AUC measures of > 0.70 with the validation datasets. The naïve Bayes model performed the best with the AUC measures of > 0.80. The accuracies in RF were more robust with narrower distribution of AUC measures compared to those in DL. The benefit of performing feature selection with a training dataset before building models was seen in some models, but not in all models. In summary, the naïve Bayes and RF models exhibited ideal predictive performance even with limited available data. The benefit of performing feature selection before building models with limited data resources depended on machine learning methods and parameters.

Keywords: coronavirus disease 2019 (COVID-19); deep learning; naïve Bayes; neural network; random forest
Tohoku J. Exp. Med., 2024 January, **262** (1), 33-41.
doi: 10.1620/tjem.2023.J090

Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has raised global public health concerns since 2019 (Huang et al. 2020; Zhu et al. 2020). In early stages of the pandemic, each country had to optimize public health strategies to quarantine infected individuals. In Japan, each prefectural government has converted pre-existing facilities, such as non-governmental hotels, into quarantine stations (Akashi et al. 2022; Machida and Wada 2022). Many studies have been conducted to build effective prediction models for the identification of potential severe cases of COVID-19 (Gallo Marin et al. 2021). The effective pre-admission triage of infected individuals to hospitals for high-risk patients or quarantine facilities for low-risk patients represented an important issue during the pandemic. Consequently, appropriate statistical models had to be developed to predict severe COVID-19 cases from the limited data available during early stages of the pandemic. To date, most predictive models have incorporated clinical or laboratory COVID-19 infection data (Li et al. 2020; Sun et al. 2020; Wang et al. 2020; Gude-Sampedro et al. 2021; Meng et al. 2021). Many of these models have achieved moderate to high predictive accuracy with areas under the receiver operating characteristic curve (ROC AUC) exceed-

Received September 30, 2023; revised and accepted October 22, 2023; J-STAGE Advance online publication November 2, 2023
Correspondence: Tetsuya Akaishi M.D., Ph.D., Department of Education and Support for Regional Medicine, Tohoku University, 1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan.
e-mail: t-akaishi@med.tohoku.ac.jp

ing 0.70 (Metz 1978). However, the symptoms differ with respect to the number of days from infection (Lauer et al. 2020), and incorporating this information into predictive models requires careful consideration of how and when to collect the data. Furthermore, laboratory and imaging data were unavailable for most cases during the initial triage step. Another issue is the selection of an appropriate statistical model for the predictive task. Conventional statistical models include multivariate analysis with a logistic regression model, as well as machine learning (ML)-based models. However, few studies to date have employed real-world data to evaluate which statistical models can more accurately predict a severe clinical course given limited available data and evidence of risk factors (Xiong et al. 2022; Ustebay et al. 2023). Moreover, the accuracy and robustness of Bayesian models, such as the naïve Bayes classification model, in the prediction of severe COVID-19 cases remains unknown. Therefore, the present study was conducted to evaluate the accuracy of predicting a severe disease course in different statistical models using basic demographic data and medical histories of COVID-19 patients obtained prior to the emergence of consequential variant strains.

Materials and Methods

Data source and participants

The present study utilized data of individuals infected with COVID-19 who were designated to the largest quarantine hotel in Miyagi Prefecture between December 2020 and February 2021 (Tadano et al. 2023). Because the study period preceded the start of the mass vaccination campaign, none of the participants had previously been vaccinated against SARS-CoV-2. This period also preceded the first occurrence of the delta variant in Japan, which may have caused severe disease profiles (Hu et al. 2022; Ong et al. 2022).

All participants were assessed upon preadmission interviews by local government health workers to be (1) clinically mild and (2) without severe conditions of medical history that may predispose them to life-threatening events due to COVID-19 infection. Detailed eligibility criteria for admission to the isolation facility have been reported previously (Tadano et al. 2023). Specifically, scores were calculated for each patient based on a combination of potential risk factors including older age, pregnancy, occurrence of serious conditions, obesity, smoking habits, and past medical history of diabetes mellitus (DM), bronchial asthma (BA), chronic obstructive pulmonary disease (COPD), uncontrolled hypertension, cardiovascular diseases (CVD), chronic kidney diseases (CKD), and malignancies. Symptomatic patients were allowed to leave the facility once they lacked antipyretics or respiratory symptoms (1) 10 days after onset and (2) 72 hours after the resolution of fever. Asymptomatic patients were permitted to leave the facility 10 days after testing for SARS-CoV-2.

Body mass index (BMI) data were provided for approximately 40% of the admitted patients. Patients with BMI data were used as the original cohort, and those without BMI data were used for the sensitivity analyses. A flow diagram of the study design is presented in Fig. 1.

Collected variables and outcome

Variables considered prior to feature selection with the least absolute shrinkage and selection operator (LASSO) included age, sex, nationality, BMI, antibiotic prescription before admission, and current smoking status, and medical history of the following 14 conditions: hypertension, DM, dyslipidemia, BA, heart disease, CVD, malignancies, COPD, CKD, hyperuricemia (HU), liver disease, psychiatric disease, sleep apnea syndrome (SAS), and atopy. The evaluated outcome was the occurrence of hypoxia with a prolonged decrease in percutaneous arterial oxygen saturation (SpO_2) $\leq 93\%$.

Statistical analysis

Machine learning process in this study was consisted of the following steps: (1) data preprocessing with z-score normalization, (2) data splitting into a training and validation dataset, (3) feature selection with a training dataset, (4) model training, and (5) cross-validation with test dataset for accuracy estimation. As the data preprocessing to improve the prediction performance, each variable was standardized using z-score as $Z = (\chi - M)/SD$, where χ is each patient's raw score, M is the mean of the population, and SD is the standard deviation (Andrade 2021; Tanaka et al. 2022). Before constructing supervised ML models to predict the development of severe COVID-19, the properties to be incorporated into these models were determined using LASSO to minimize dimensionality and maximize predictive power (Yamada et al. 2014). In the LASSO, $\ell 1$ norm was used as the penalty, and features with calculated coefficients that were reduced to zero were excluded from subsequent ML models. Feature selection was performed only with the training dataset and not with the validation dataset to avoid data leakage with excessively optimistic evaluation caused by using the whole dataset before cross-validation in building models (Yagis et al. 2021). In the subsequent cross-validation step, 60% of the data were allocated as the training dataset, with the remaining 40% reserved for validation. The following supervised ML models were prepared: linear discriminant analysis (LDA), nonlinear discrimination with logistic regression (LR), a tree-based model with classification and regression trees (CART), support vector machines (SVM), random forest (RF), a single-hidden layer neural network (NN), deep learning (DL) models with multiple hidden layers, and naïve Bayes models with and without Kernel density estimation for non-parametrically estimating the probability density function (Uddin et al. 2019). In the SVM model, three-fold cross-validation was performed. In the NN models, validation accuracies were obtained using 100 epochs with 3-fold cross-validation. In the RF model, square-rooted quantities

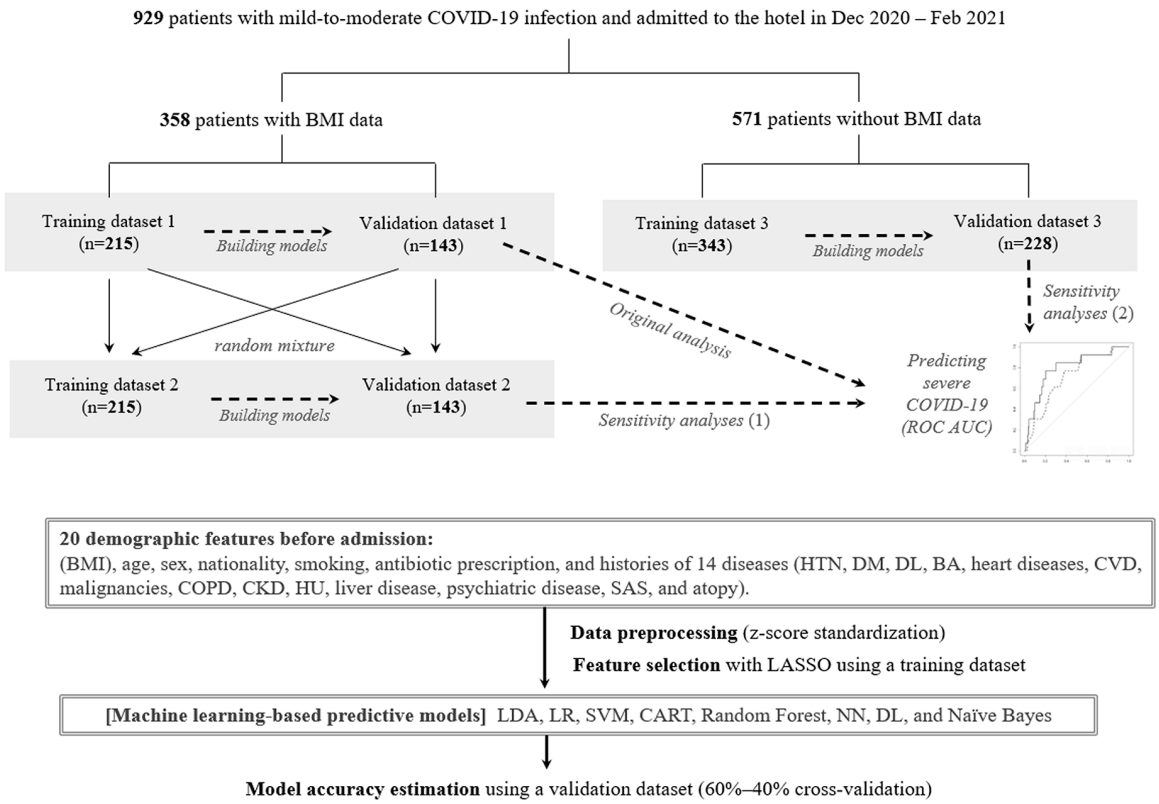

Fig. 1. Flow diagram of study design.
A total of 929 patients with mild to moderate coronavirus disease 2019 (COVID-19) symptoms, admitted to a quarantine hotel between Dec. 2020 and Feb. 2021, were enrolled in this study. Patients were divided into cohorts of 358 and 571 individuals based on the presence or absence of body mass index (BMI) data, respectively. For each cohort, several sensitivity analyses were performed to evaluate the robustness of model performance. Feature selection using least absolute shrinkage and selection operator (LASSO) before building models was performed only with the training dataset to avoid data leakage. ROC AUC, area under the receiver operating characteristic curve; HTN, hypertension; DM, diabetes mellitus; DL, dyslipidemia; BA bronchial asthma; CVD, cardiovascular disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease; HU, hyperuricemia; SAS, sleep apnea syndrome; LDA, linear discriminant analysis; LR, logistic regression; SVM, Support Vector Machines; CART, Classification and Regression Trees; NN, neutral network; DL, deep learning.

of incorporated features were randomly selected to build 500 decision trees. The number of decision trees was determined after checking for the relationship between the error rate and the number of decision trees. In the single-hidden-layer NN model, five units were prepared in the hidden layer. Multiple patterns for the numbers of hidden layers and neurons in each layer were tested for the single-hidden-layer NN and DL models, with predictive accuracies compared by ROC AUC using the DeLong test (DeLong et al. 1988). Multiple comparisons of the AUC measures were not adjusted based on the exploratory nature of the present study. For the ML model with the highest predictive accuracy, the importance of preliminary feature selection was verified by calculating AUC with and without said selection. To verify the robustness of AUC measures, 20-fold iterated measurements were performed for the NN, DL, and RF models. Comparisons of repeated AUC measures between models with and without feature selection were performed using the Mann-Whitney U test. The AUC measurements were reperformed after randomly reallocating training and validation datasets in the first cohort with BMI data. To verify the robustness of the finding, AUC measurements were further performed in the second cohort without BMI data. All statistical analyses were performed using Python 3.11.1, and R version 4.1.3 (R Foundation, Vienna, Austria).

Ethics

This study was approved by the institutional review board of Tohoku University Graduate School of Medicine (approval number: 2021-1-1178). Written informed consent was waived by the review board because of the anonymity of the present study and to prevent unnecessary risks of transmitting the infection by obtaining written forms from the participants. All process of the study was performed in accordance with the latest version of the Declaration of Helsinki as revised in 2013 (World Medical Association 2013).

Table 1. Area under the receiver operating characteristic curve (ROC AUC) measures with different machine learning models and parameters (original dataset).

Models	ML method	Parameters	AUC (95% CI)	
			Feature selection: Done	Feature selection: Not done*
Model 1	Liner discriminant analysis	-	0.671 (0.462-0.881)	0.671 (0.462-0.881)
Model 2	Nonlinear discrimination	Logistic regression model	0.522 (0.215-0.830)	0.540 (0.257-0.824)
Model 3	SVM	3-fold cross validation	0.594 (0.359-0.830)	0.594 (0.359-0.830)
Model 4	CART	-	0.763 (0.595-0.932)	0.763 (0.595-0.932)
Model 5	Random Forest	500 decision trees	0.753 (0.645-0.860)	0.723 (0.622-0.823)
Model 6.1	Single-HL NN	5 units in the hidden layer	0.664 (0.455-0.873)	0.718 (0.546-0.890)
Model 6.2	Single-HL NN	10 units in the hidden layer	0.671 (0.464-0.879)	0.721 (0.551-0.890)
Model 7.1	Two-HL NN	2 HLs with [2, 2] neurons	0.763 (0.568-0.958)	0.722 (0.512-0.932)
Model 7,2	Two-HL NN	2 HLs with [3, 2] neurons	0.735 (0.596-0.875)	0.585 (0.289-0.881)
Model 7.3	Two-HL NN	2 HLs with [5, 3] neurons	0.581 (0.313-0.849)	0.590 (0.380-0.800)
Model 7.4	Two-HL NN	2 HLs with [10, 5] neurons	0.585 (0.347-0.824)	0.705 (0.401-1.000)
Model 7.5	Deep Learning	3 HLs with [4, 3, 2] neurons	0.705 (0.536-0.873)	0.681 (0.420-0.941)
Model 7.6	Deep Learning	3 HLs with [15, 10, 5] neurons	0.676 (0.534-0.818)	0.827 (0.700-0.954)
Model 7.7	Deep Learning	3 HLs with [30, 20, 10] neurons	0.491 (0.259-0.722)	0.670 (0.411-0.930)
Model 7.8	Deep Learning	3 HLs with [45, 30, 15] neurons	0.596 (0.355-0.838)	0.568 (0.282-0.853)
Model 7.9	Deep Learning	5 HLs with [50, 40, 30, 20, 10] neurons	0.602 (0.377-0.826)	0.541 (0.252-0.829)
Model 8.1	Naïve Bayes	Using Kernel density estimation	0.856 (0.744-0.967)	0.868 (0.803-0.932)
Model 8.2	Naïve Bayes	Not using Kernel density estimation	0.825 (0.715-0.935)	0.794 (0.708-0.881)

The AUC measure in each machine learning (ML) model was obtained with or without feature selection by least absolute shrinkage and selection operator (LASSO) using the training dataset. Feature selection with LASSO was performed with the training dataset before building each ML model to reduce dimensionality, with eight eligible features identified with nonzero coefficients [i.e., age, sex, body mass index (BMI), hypertension (HTN), diabetes mellitus (DM), bronchial asthma (BA), hyperuricemia (HU), and antibiotics]. The ML models were built based on a training dataset of 215 individuals, and AUC measures were obtained from the validation dataset encompassing the 143 remaining individuals. For all models, the occurrence of prolonged decrement in SpO_2 measures ≤ 93% was used as the binary outcome.

CART, Classification and Regression Trees; HL, hidden layer; NN, neural network; SVM, Support Vector Machines.

*All 20 features before feature selection by LASSO were used in these models.

Results

Overall participants

Of the 929 patients with reliable daily SpO_2 measurements enrolled during the study period, 358 had BMI data and the remaining 571 did not. Fifteen patients (1.6%) were transferred from hospitals to quarantine facilities for continued isolation. Previously reported demographic data (Tadano et al. 2023) indicate that although none of the patients were hypoxic (with $SpO_2 \leq 93\%$) on admission to the isolation facility, 63 (6.8%) developed hypoxia at a median of 8 days (interquartile range: 6-10 days) from the clinical onset.

First cohort

The first cohort included data from 358 individuals (197 males and 161 females), including 96 current smokers, with evaluated variables and reliable SpO_2 measurement results. The median and interquartile range (IQR; 25-75 percentile) of age at hotel admission were 39 and 24-52 years, respectively. Among them, 341 were of Japanese nationality, 15 were Asian from countries other than Japan, and 2 were Caucasians. The following prevalence of medical histories was reported: 60 with hypertension, 16 with diabetes mellitus, 34 with dyslipidemia, 24 with BA, 14 with heart disease, 3 with CVD, 9 with malignancies, 2 with COPD, 2 with CKD, 6 with HU, 3 with liver diseases, and 4 with psychiatric diseases. The median BMI was 22.67 (IQR; 20.30-26.35). There were 23 patients (6.4%) who developed hypoxia following clinical onset. The overall cohort was randomly divided into a training dataset (215 individuals, including 15 who developed hypoxia) and a validation dataset (143 individuals, including 8 who developed hypoxia).

Feature selection

In the initial feature selection stage using LASSO with the training dataset, the following pre-infection variables produced non-zero coefficients and were used in the subsequent ML models: age (Wald $\chi^2 = 4.65$; $p = 0.0310$), BMI ($\chi^2 = 5.43$; $p = 0.0198$), sex ($\chi^2 = 2.27$; $p = 0.1318$), history of hypertension ($\chi^2 = 0.12$; $p = 0.7249$), history of DM ($\chi^2 = 2.89$; $p = 0.0893$), history of BA ($\chi^2 = 0.16$; $p = 0.6866$), history of HU ($\chi^2 = 0.46$; $p = 0.4967$), and antibiotic pre-

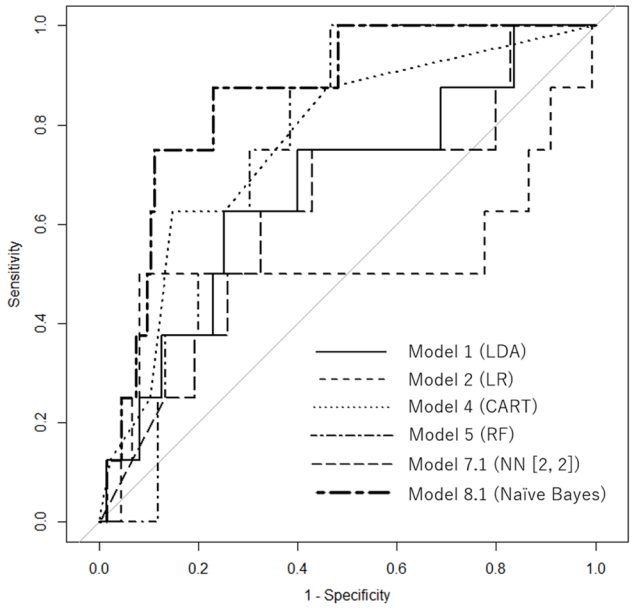

Fig. 2. The receiver operating characteristic (ROC) curves with evaluated machine learning (ML) models for predicting severe COVID-19 cases.
The conventional linear regression model (Model 2) exhibited poor AUC measures below 0.60. Other ML models produced AUC measures greater than 0.70 when the parameters were appropriately assigned. Especially, the naïve Bayes models exhibited AUC measures of > 0.80. CART, classification and regression trees; LDA, linear discriminant analysis; LR, logistic regression; NN, neural network; RF, random forest.

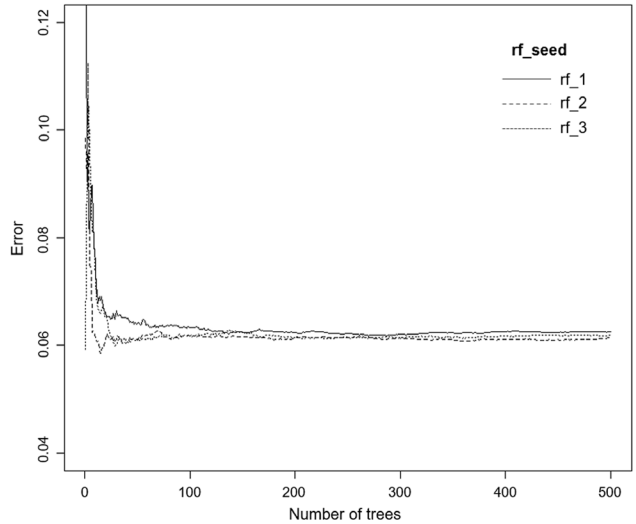

Fig. 3. Relationship between the number of trees in random forest (RF) model and predictive error.
As the predictive error was suggested to decrease with an increased number of trees in RF model, the relationship between the number of trees and the predictive error rate was evaluated for three times to determine the optimal number of trees for a stable prediction. The obtained line graphs suggested that approximately 300 trees realize minimal error rate, and the error rate was stabilized above this number of trees.

scription before admission (χ^2 = 0.55; p = 0.4577). All other variables produced coefficients of zero and were eliminated from the subsequent ML models. For the naïve Bayes models, variables producing variance errors were further excluded.

AUC measures for ML models (original data configuration)

Table 1 lists the obtained AUC measures and 95% confidence intervals obtained for each ML model using the validation dataset for the first cohort. The linear discriminator (AUC, 0.671; p = 0.0525), logistic regression (AUC, 0.522; p = 0.5852), SVM (AUC, 0.594; p = 0.1863), and DL models with too many hidden layers or neurons failed to show a significant prediction accuracy. The highest AUC measures were obtained with the naïve Bayes model, with both configurations with and without kernel density estimation exhibiting AUC measures greater than 0.80 (p = 0.0004 with kernel density estimation and p = 0.0010 without it). In the NN and DL models, AUC measures largely differed according to parameters such as the numbers of hidden layers and neurons, with too many of either resulting in lower AUC measures, possibly reflecting overfitting of the models. The ROC curves obtained by the evaluated models are presented in Fig. 2, suggesting the superiority of the naïve Bayes model over the other models.

Next, to decide the optimal number of trees in RF model, the relationship between the number of trees and the errors in prediction were evaluated for three times with different random seeds to determine the optimal number of trees for a stable prediction (Fig. 3). In obtaining the errors in these analyses, the outcome was used as a dummy variable. The obtained results indicated that approximately 300 trees would realize minimal error rate, and the error rate will not decrease by further increasing the number of trees. Based on this finding, the number of trees in RF was set with 500 in the subsequent sensitivity analyses.

Repeated AUC measures with NN, DL, and RF models

To determine the variability of AUC measures with the NN, DL, and RF models, 20-fold repeated AUC measurements with the validation dataset were performed with the NN models with two hidden layers (of [3, 2] neurons or [10, 5] neurons), a DL model with three hidden layers (of [4, 3, 2] neurons), and an RF model. The distribution of AUC measurements is depicted in Fig. 4. These measures were more widely distributed with the NN and DL models irrespective of parameters than with the RF model, suggesting the high reliability and reproducibility of the RF model. The benefit of performing feature selection with the training dataset before building models depended on the types of ML model and the parameters.

Sensitivity analysis with randomly reassigned training and validation datasets

Next, a sensitivity analysis with randomly changed

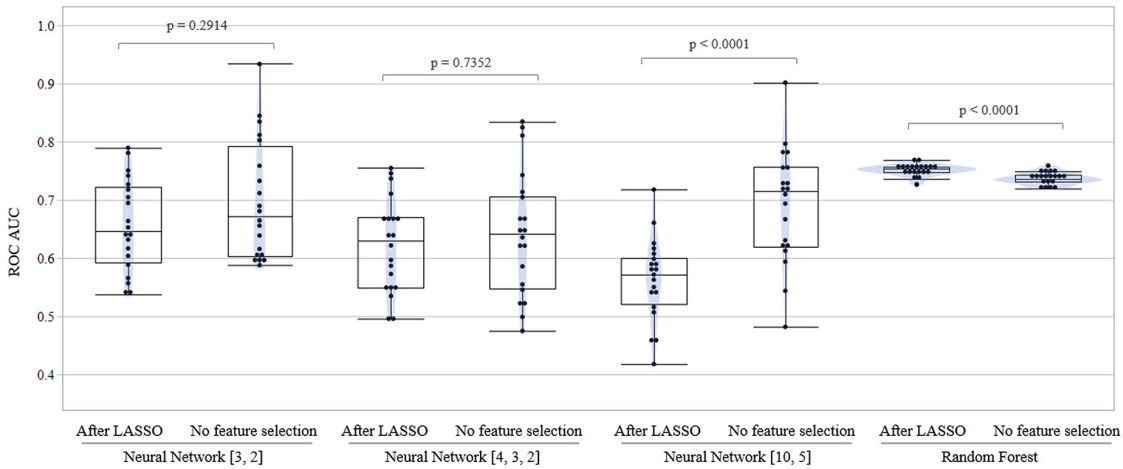

Fig. 4. Variability of AUC measures based on different machine learning (ML) models.
The distributions shown with eight different ML models were obtained from 20 iterations of ML simulations among 358 patients with body mass index (BMI) data. The AUC measures obtained with random forest (RF) exhibited narrower distributions than those obtained via deep learning (DL) models, implying greater robustness of the former. The AUC measure obtained with the RF method was higher when preliminary feature selection was performed with least absolute shrinkage and selection operator (LASSO) ($p < 0.0001$), demonstrating the importance of feature selection in this model. Meanwhile, the benefit of preliminary feature selection could not be confirmed with the neutral network (NN) and DL models. The p-values are results of the Mann-Whitney U test. NN [4,3,2] denotes a DL model with three hidden layers comprising 4, 3, and 2 neurons. Validation accuracies in NN models were obtained using 100 epochs with 3-fold cross-validation. The number of trees in RF was 500.

datasets for training (215 individuals, including 13 with the primary outcome episode) and validation (143 individuals, including 10 with the primary outcome episode) was performed to verify the robustness of the present findings. The AUC measurements obtained using the changed datasets are listed in Table 2. Again, the RF model, DL models with small number or neurons, and naïve Bayes models exhibited robust predictive accuracies with moderate-to-high levels of AUC measures greater than 0.70. In particular, the naïve Bayes models again achieved the highest AUC measures, exceeding 0.80. These models were the only models that showed AUC measures greater than 0.80 for both data configurations.

Second cohort

Finally, to further evaluate the reproducibility of the results, another sensitivity analysis was performed with the 571 patients without available BMI data. Among this cohort, 40 patients (7.0%) developed hypoxia following clinical onset. Sixty percent of this cohort was randomly allocated for the training dataset (343 individuals, including 23 patients who developed hypoxia), with the remaining 40% reserved for the validation dataset (228 individuals, including 17 patients who developed hypoxia). First, feature selection using LASSO with the training dataset identified the following six characteristics with non-zero coefficients: age, sex, dyslipidemia, heart disease, liver disease, and psychiatric disease. The AUC measures obtained with different ML-based models by using these characteristics are listed in Table 3. As in the previous analyses, the naïve Bayes models, NN models with not too many hidden layers or neurons, and RF model exhibited moderate-to-high AUC measures of > 0.70, with the naïve Bayes models showing the highest AUC.

Discussion

In the present study, different types of prediction models based on the conventional logistic regression model and other ML models were comprehensively evaluated, and their accuracies were compared in the prediction of severe conditions using only pre-infection data. Although the models in this study did not use data directly pertaining to COVID-19-related symptoms, most of them exhibited moderate-to-high AUC measures exceeding 0.70. In particular, the naïve Bayes models exhibited the highest AUC measures, exceeding 0.80, in all evaluated data configurations. The findings suggested that some ML-based models, including the RF, DL, and naïve Bayes models, would realize higher AUC measures compared to conventional logistic regression model even with limited data resources in size and variables. These ML-based predictive models may contribute to the initial triage stage of public health agencies when predicting outcomes in the absence of reliable data and evidence of risk factors, especially in early phases of a pandemic. Other notable findings of the present study include the wide distribution of expected AUC measures with NN and DL models depending on their parameters, such as the numbers of hidden layers and neurons. These findings collectively indicate the excellent usability of the RF and naïve Bayes models in predicting severe COVID-

Table 2. Sensitivity analysis for AUC measures with randomly reassigned training and validation datasets following feature selection.

Models	ML method	Parameters	AUC (95% CI)
Model 1	Liner discriminant analysis		0.793 (0.708-0.877)
Model 2	Nonlinear discrimination	Logistic regression model	0.753 (0.585-0.921)
Model 3	SVM	3-fold cross validation	0.802 (0.672-0.932)
Model 4	CART		0.734 (0.577-0.890)
Model 5	Random Forest	500 decision trees	0.795 (0.670-0.920)
Model 6.1	Single-HL NN	5 units in the hidden layer	0.824 (0.733-0.916)
Model 6.2	Single-HL NN	10 units in the hidden layer	0.819 (0.729-0.908)
Model 7.1	Two-HL NN	2 HLs with [2, 2] neurons	0.815 (0.733-0.898)
Model 7,2	Two-HL NN	2 HLs with [3, 2] neurons	0.708 (0.527-0.888)
Model 7.3	Two-HL NN	2 HLs with [5, 3] neurons	0.834 (0.751-0.917)
Model 7.4	Two-HL NN	2 HLs with [10, 5] neurons	0.721 (0.525-0.912)
Model 7.5	Deep Learning	3 HLs with [4, 3, 2] neurons	0.824 (0.710-0.939)
Model 7.6	Deep Learning	3 HLs with [15, 10, 5] neurons	0.793 (0.616-0.969)
Model 7.7	Deep Learning	3 HLs with [30, 20, 10] neurons	0.497 (0.289-0.705)
Model 7.8	Deep Learning	3 HLs with [45, 30, 15] neurons	0.698 (0.474-0.922)
Model 7.9	Deep Learning	5 HLs with [50, 40, 30, 20, 10] neurons	0.730 (0.516-0.945)
Model 8.1	Naïve Bayes	Using Kernel density estimation	0.839 (0.749-0.929)
Model 8.2	Naïve Bayes	Not using Kernel density estimation	0.811 (0.716-0.905)

Sensitivity analysis was performed by randomly selecting another pair of datasets for training (215 patients) and validation (143 patients) using the features selected by least absolute shrinkage and selection operator (LASSO). The feature selection was performed with the new training dataset, and the following 11 features were with nonzero coefficients: age, sex, body mass index (BMI), hypertension (HTN), diabetes mellitus (DM), bronchial asthma (BA), heart diseases, cardiovascular disease (CVD), chronic obstructive pulmonary disease (COPD), hyperuricemia (HU), and atopy. Validation accuracies in neural network (NN) models were obtained using 100 epochs with 3-fold cross-validation.
ML, machine learning; SVM, Support Vector Machines; CART, Classification and Regression Trees; HL, hidden layer.

19 cases when reliable clinical or laboratory data are unavailable. In clinical studies, it is crucial to determine whether a predictive model can be structurally interpreted in view of distinct risk factors. In this respect, the conventional logistic regression model has an advantage over ML-based models. Conversely, in view of the practical usability of predictive models in actual triage processes, predictive accuracy may be prioritized over interpretability by certain public health policies, in which case ML models are more desirable than those derived from conventional logistic regression. Although the naïve Bayes models employed fewer features, they exhibited promising potential as predictive models for severe COVID-19 cases. Subsequent attempts to develop predictive models for severe cases may benefit from using the naïve Bayes classifier, particularly given a relatively small quantity of training set.

To date, few studies have evaluated the accuracy of the naïve Bayes classifiers in the prediction of severe COVID-19 cases. The naïve Bayes classifier uses Bayes' theorem with a strong assumption of independence between features, with parameters approximated using the maximum likelihood method. Despite its strong independence assumption with input features that appear to be oversimplified for real-world data, this model has been reported to exhibit excellent performance compared to logistic regression and even other supervised ML models (Awan et al. 2020; Golpour et al. 2020; Mfateneza et al. 2022). One strength of the naïve Bayes classifier is that it often works well with a relatively small amount of training data, as demonstrated in a previous study (Sardesai et al. 2021). This strength may be derived from the assumption of independence between features, as the dimension of calculation, or number of data points, used to estimate the optimal parameters is well-suppressed to a lower level than in other ML models, including DL models. This is particularly important when the data dimensionality is high for a number of training datasets, which results in the curse-of-dimensionality problem. In the early stages of a pandemic, with limited direct data and evidence, the issue of high dimensionality with a low training data size is a frequent occurrence. In such situations, the naïve Bayes model represents a promising approach for the prediction of outcomes along with the conventional logistic regression model or penalized feature selection methods, such as LASSO, for extracting significant risks.

This study had several limitations. First, the number of available data points was relatively small, and the incidence of the primary outcomes was relatively low at less

Table 3. AUC measures following feature selection using data from another cohort of 571 patients.

Models	ML method	Parameters	AUC (95% CI)
Model 1	Liner discriminant analysis		0.749 (0.659-0.839)
Model 2	Nonlinear discrimination	Logistic regression model	0.777 (0.683-0.872)
Model 3	SVM	3-fold cross validation	0.550 (0.383-0.717)
Model 4	CART		0.749 (0.629-0.869)
Model 5	Random Forest	500 decision trees	0.779 (0.670-0.888)
Model 6.1	Single-HL NN	5 units in the hidden layer	0.789 (0.705-0.874)
Model 6.2	Single-HL NN	10 units in the hidden layer	0.796 (0.713-0.878)
Model 7.1	Two-HL NN	2 HLs with [2, 2] neurons	0.673 (0.538-0.808)
Model 7,2	Two-HL NN	2 HLs with [3, 2] neurons	0.741 (0.639-0.844)
Model 7.3	Two-HL NN	2 HLs with [5, 3] neurons	0.822 (0.751-0.894)
Model 7.4	Two-HL NN	2 HLs with [10, 5] neurons	0.737 (0.588-0.886)
Model 7.5	Deep Learning	3 HLs with [4, 3, 2] neurons	0.693 (0.546-0.839)
Model 7.6	Deep Learning	3 HLs with [15, 10, 5] neurons	0.733 (0.587-0.879)
Model 7.7	Deep Learning	3 HLs with [30, 20, 10] neurons	0.749 (0.624-0.873)
Model 7.8	Deep Learning	3 HLs with [45, 30, 15] neurons	0.749 (0.595-0.903)
Model 7.9	Deep Learning	5 HLs with [50, 40, 30, 20, 10] neurons	0.738 (0.604-0.872)
Model 8.1	Naïve Bayes	Using Kernel density estimation	0.821 (0.743-0.900)
Model 8.2	Naïve Bayes	Not using Kernel density estimation	0.798 (0.718-0.879)

Another sensitivity analysis was conducted on a different cohort of 571 patients without available body mass index (BMI) data. Sixty percent of the patients were randomly allocated for the training dataset (n = 343), with the remaining 40% reserved for the validation dataset (n = 228). Feature selection was performed using the least absolute shrinkage and selection operator (LASSO) method for the training dataset, and the following six features were with nonzero coefficients: age, sex, deep learning (DL), heart disease, liver disease, and psychiatric disease.
ML, machine learning; SVM, Support Vector Machines; CART, Classification and Regression Trees; HL, hidden layer; NN, neural network.

than 10%. Consequently, the generalizability of our findings to other demographics, including other variants of COVID-19, remains uncertain. Further studies using data from patients with these variants are required to confirm our hypotheses. Another limitation pertains to the patterns of parameters used for the NN and DL models, with 1-5 hidden layers and 2-50 neurons in each layer. Parameter optimization is an essential but difficult issue in developing DL models, and the advantages of these models could not be statistically evaluated in this study. Finally, because the present study encompassed the period before the development of vaccines against COVID-19, the collected features did not include a history of vaccination. In similar future studies, the vaccination status of each patient must also be considered as an important demographic feature, as vaccination status is known to suppress the incidence of severe COVID-19 (Ng et al. 2022).

In conclusion, this study evaluated the robustness and accuracy of clinical predictive models based on logistic regression and ML models given a limited sample size and set of variables. Several ML-based models, including the naïve Bayes, RF, NN, and DL models, performed better than the conventional logistic regression model for this task. Conversely, an excessive number of hidden layers or neurons in a DL model resulted in suboptimal predictive accuracy. The benefit of performing feature selection before building models depended on the types of ML models and their parameters. Overall, this study demonstrated a high usability of the naïve Bayes model in building a prediction model when the data resources and the evidence of risk factors are limited.

Acknowledgments

The authors appreciate all medical staffs and local government staffs (Miyagi Prefecture) who cooperated to the management of the quarantine facility where the present study was performed.

This study was funded by JSPS KAKENHI Grant Number JP21K10367.

Author Contributions

T.A., Y.T., and T.I. contributed to the concept, design, and data collection of this study. T.A. and Y.T. performed statistical analyses. Y.K. and N.Y. verified the machine learning processes. Y.T. played a primary role in data collection. T.A. drafted the manuscript and prepared figures and tables. T.I. and N.Y. supervised this study. All authors critically revised and approved the final version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

Akashi, H., Kodoi, H., Noda, S., Tamura, T., Baba, H., Chinda, E., Thandar, M.M., Naito, K., Watanabe, Y., Suzuki, Y., Narita, T. & Shimazu, T. (2022) Reporting on the implementation to set up a "care and isolation facility" for mild COVID-19 cases in Tokyo. *Glob. Health Med.*, **4**, 71-77.

Andrade, C. (2021) Z scores, standard scores, and composite test scores explained. *Indian J. Psychol. Med.*, **43**, 555-557.

Awan, F.M., Saleem, Y., Minerva, R. & Crespi, N. (2020) A comparative analysis of machine/deep learning models for parking space availability prediction. *Sensors* (*Basel*), **20**, 322.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837-845.

Gallo Marin, B., Aghagoli, G., Lavine, K., Yang, L., Siff, E.J., Chiang, S.S., Salazar-Mather, T.P., Dumenco, L., Savaria, M.C., Aung, S.N., Flanigan, T. & Michelow, I.C. (2021) Predictors of COVID-19 severity: a literature review. *Rev. Med. Virol.*, **31**, 1-10.

Golpour, P., Ghayour-Mobarhan, M., Saki, A., Esmaily, H., Taghipour, A., Tajfard, M., Ghazizadeh, H., Moohebati, M. & Ferns, G.A. (2020) Comparison of support vector machine, naïve Bayes and logistic regression for assessing the necessity for coronary angiography. *Int. J. Environ. Res. Public Health*, **17**, 6449.

Gude-Sampedro, F., Fernández-Merino, C., Ferreiro, L., Lado-Baleato, Ó., Espasandín-Domínguez, J., Hervada, X., Cadarso, C.M. & Valdés, L. (2021) Development and validation of a prognostic model based on comorbidities to predict COVID-19 severity: a population-based study. *Int. J. Epidemiol.*, **50**, 64-74.

Hu, Z., Huang, X., Zhang, J., Fu, S., Ding, D. & Tao, Z. (2022) Differences in clinical characteristics between delta variant and wild-type SARS-CoV-2 infected patients. *Front. Med.* (*Lausanne*), **8**, 792135.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., et al. (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, **395**, 497-506.

Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G. & Lessler, J. (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.*, **172**, 577-582.

Li, X., Marmar, T., Xu, Q., Tu, J., Yin, Y., Tao, Q., Chen, H., Shen, T. & Xu, D. (2020) Predictive indicators of severe COVID-19 independent of comorbidities and advanced age: a nested case-control study. *Epidemiol. Infect.*, **148**, e255.

Machida, M. & Wada, K. (2022) Public health responses to COVID-19 in Japan. *Glob. Health Med.*, **4**, 78-82.

Meng, Z., Wang, M., Zhao, Z., Zhou, Y., Wu, Y., Guo, S., Li, M., Zhou, Y., Yang, S., Li, W. & Ying, B. (2021) Development and validation of a predictive model for severe COVID-19: a case-control study in China. *Front. Med.* (*Lausanne*), **8**, 663145.

Metz, C.E. (1978) Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283-298.

Mfateneza, E., Rutayisire, P.C., Biracyaza, E., Musafiri, S. & Mpabuka, W.G. (2022) Application of machine learning methods for predicting infant mortality in Rwanda: analysis of Rwanda demographic health survey 2014-15 dataset. *BMC Pregnancy Childbirth*, **22**, 388.

Ng, O.T., Marimuthu, K., Lim, N., Lim, Z.Q., Thevasagayam, N.M., Koh, V., Chiew, C.J., Ma, S., Koh, M., Low, P.Y., Tan, S.B., Ho, J., Maurer-Stroh, S., Lee, V.J.M., Leo, Y.S., et al. (2022) Analysis of COVID-19 incidence and severity among adults vaccinated with 2-dose mRNA COVID-19 or inactivated SARS-CoV-2 vaccines with and without boosters in Singapore. *JAMA Netw. Open*, **5**, e2228900.

Ong, S.W.X., Chiew, C.J., Ang, L.W., Mak, T.M., Cui, L., Toh, M., Lim, Y.D., Lee, P.H., Lee, T.H., Chia, P.Y., Maurer-Stroh, S., Lin, R.T.P., Leo, Y.S., Lee, V.J., Lye, D.C., et al. (2022) Clinical and virological features of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) variants of concern: a retrospective cohort study comparing B.1.1.7 (Alpha), B.1.351 (Beta), and B.1.617.2 (Delta). *Clin. Infect. Dis.*, **75**, e1128-e1136.

Sardesai, A.U., Tanak, A.S., Krishnan, S., Striegel, D.A., Schully, K.L., Clark, D.V., Muthukumar, S. & Prasad, S. (2021) An approach to rapidly assess sepsis through multi-biomarker host response using machine learning algorithm. *Sci. Rep.*, **11**, 16905.

Sun, L., Song, F., Shi, N., Liu, F., Li, S., Li, P., Zhang, W., Jiang, X., Zhang, Y., Sun, L., Chen, X. & Shi, Y. (2020) Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. *J. Clin. Virol.*, **128**, 104431.

Tadano, Y., Akaishi, T., Suzuki, S., Ono, R., Saito, N., Arita, R., Kanno, T., Tanaka, J., Kikuchi, A., Ohsawa, M., Takayama, S., Abe, M., Onodera, K. & Ishii, T. (2023) Predictors for the development of hypoxia or prolonged acute symptoms among non-hospitalized mild-to-moderate patients with coronavirus disease 2019. *Tohoku J. Exp. Med.*, **260**, 231-244.

Tanaka, T., Nambu, I., Maruyama, Y. & Wada, Y. (2022) Sliding-window normalization to improve the performance of machine-learning models for real-time motion prediction using electromyography. *Sensors* (*Basel*), **22**, 5005.

Uddin, S., Khan, A., Hossain, M.E. & Moni, M.A. (2019) Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.*, **19**, 281.

Ustebay, S., Sarmis, A., Kaya, G.K. & Sujan, M. (2023) A comparison of machine learning algorithms in predicting COVID-19 prognostics. *Intern. Emerg. Med.*, **18**, 229-239.

Wang, F., Hou, H., Wang, T., Luo, Y., Tang, G., Wu, S., Zhou, H. & Sun, Z. (2020) Establishing a model for predicting the outcome of COVID-19 based on combination of laboratory tests. *Travel Med. Infect. Dis.*, **36**, 101782.

World Medical Association (2013) World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, **310**, 2191-2194.

Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C. & Huang, L.; National Traditional Chinese Medicine Medical Team (2022) Comparing different machine learning techniques for predicting COVID-19 severity. *Infect. Dis. Poverty*, **11**, 19.

Yagis, E., Atnafu, S.W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L. & Diciotti, S. (2021) Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. Rep.*, **11**, 22544.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E.P. & Sugiyama, M. (2014) High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Comput.*, **26**, 185-207.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., et al. (2020) A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.*, **382**, 727-733.