ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Robustness of FTIR-Based Ultrarapid COVID-19 Diagnosis Using PLS-DA

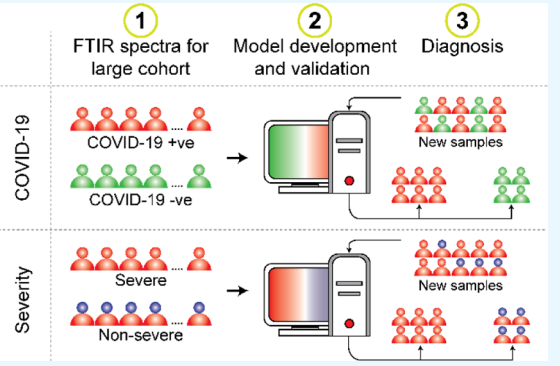
Sreejith Remanan Pushpa, Rajeev Kumar Sukumaran, and Sivaraman Savithri*

Cite This: *ACS Omega* 2022, 7, 47357−47371

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The World Health Organization (WHO) declared the Omicron variant (B.1.1.529) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the pathogen responsible for the Coronavirus disease 2019 (COVID-19) pandemic, as a variant of concern on 26 November 2021. By this time, 42% of the world's population had received at least one dose of the vaccine against COVID-19. As on 1 October 2022, only 68% of the world population got the first dose of the vaccine. Although the vaccination is incredibly protective against severe complications of the disease and death, the highly contagious Omicron variant, compared to the Delta variant (B.1.617.2), has led the whole world into more chaotic situations. Furthermore, the virus has a high mutation rate, and hence, the possibility of a new variant of concern in the future cannot be ruled out. To face such a challenging situation, paramount importance should be given to rapid diagnosis and isolation of the infected patient. Current diagnosis methods, including reverse transcription-polymerase chain reaction and rapid antigen tests, face significant burdens during a COVID-19 wave. However, studies reported ultrarapid, reagent-free, cost-efficient, and non-destructive diagnosis methods based on chemometrics for COVID-19 and COVID-19 severity diagnosis. These studies used a smaller sample cohort to construct the diagnosis model and failed to discuss the robustness of the model. The current study systematically evaluated the robustness of the diagnosis models trained using smaller (real and augmented spectra) and larger (augmented spectra) datasets. The Monte Carlo cross-validation and permutation test results suggest that diagnosis using models trained by larger datasets was accurate and statistically significant (Q^2 > 99% and AUROC = 100%).



INTRODUCTION

In December 2019, China reported an outbreak of Coronavirus Disease 2019 (COVID-19), a typical viral pneumonia caused by Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1] Although studies reported similar viral pneumonia caused by SARS-CoV and the Middle East respiratory syndrome coronavirus (MERS-CoV) with a higher mortality rate than that of COVID-19, SARS-CoV-2 spreads significantly faster than MERS-CoV and SARS-CoV,[2] leading the World Health Organization (WHO) to declare COVID-19 as a global pandemic on 11 March 2020.[3] Also, coronaviruses were genetically prone to a high rate of mutations mainly because of their RNA polymerase with a restricted proofreading mechanism.[4] Therefore, the chances of a new variant of SARS-CoV-2 with a high transmission rate than that of the Delta or the Omicron variants responsible for the previous COVID-19 waves cannot be ruled out. In that case, rapid diagnosis and isolation of the COVID-19-positive (+ve) patients are pivotal to preventing transmission of infection.

The nucleotide-based reverse transcription-polymerase chain reaction (RT-PCR) is a globally accepted diagnosis method.[5] RT-PCR is the best method for diagnosis, but it is time-consuming and can only be performed in a certified laboratory with trained health professionals. Also, the paramount demand is the costly equipment, reagents, primers, and probes.[5] Moreover, the primers and probes used for the process were vulnerable to mutations in the target genes, leading to more false-positive and false-negative diagnoses.[6] Similarly, getting an RT-PCR result may take approximately 24 h or longer,[7] increasing the chance of viral spread. However, most of the challenges faced by RT-PCR were reduced significantly after introducing the immunoassay-based rapid detection kits that were more accessible to the general public.[7,8] Still, during a COVID-19 wave, the shortage of health professionals, RT-PCR reagents, and rapid detection kits worsens the scenario.

To overcome these problems in diagnosis, Barauna et al. (2021) proposed an ultrarapid, reagent-free, and non-destructive diagnosis method using Fourier-transform infrared (FTIR) spectroscopy coupled with chemometrics.[9] The

Received: October 21, 2022
Accepted: November 29, 2022
Published: December 8, 2022

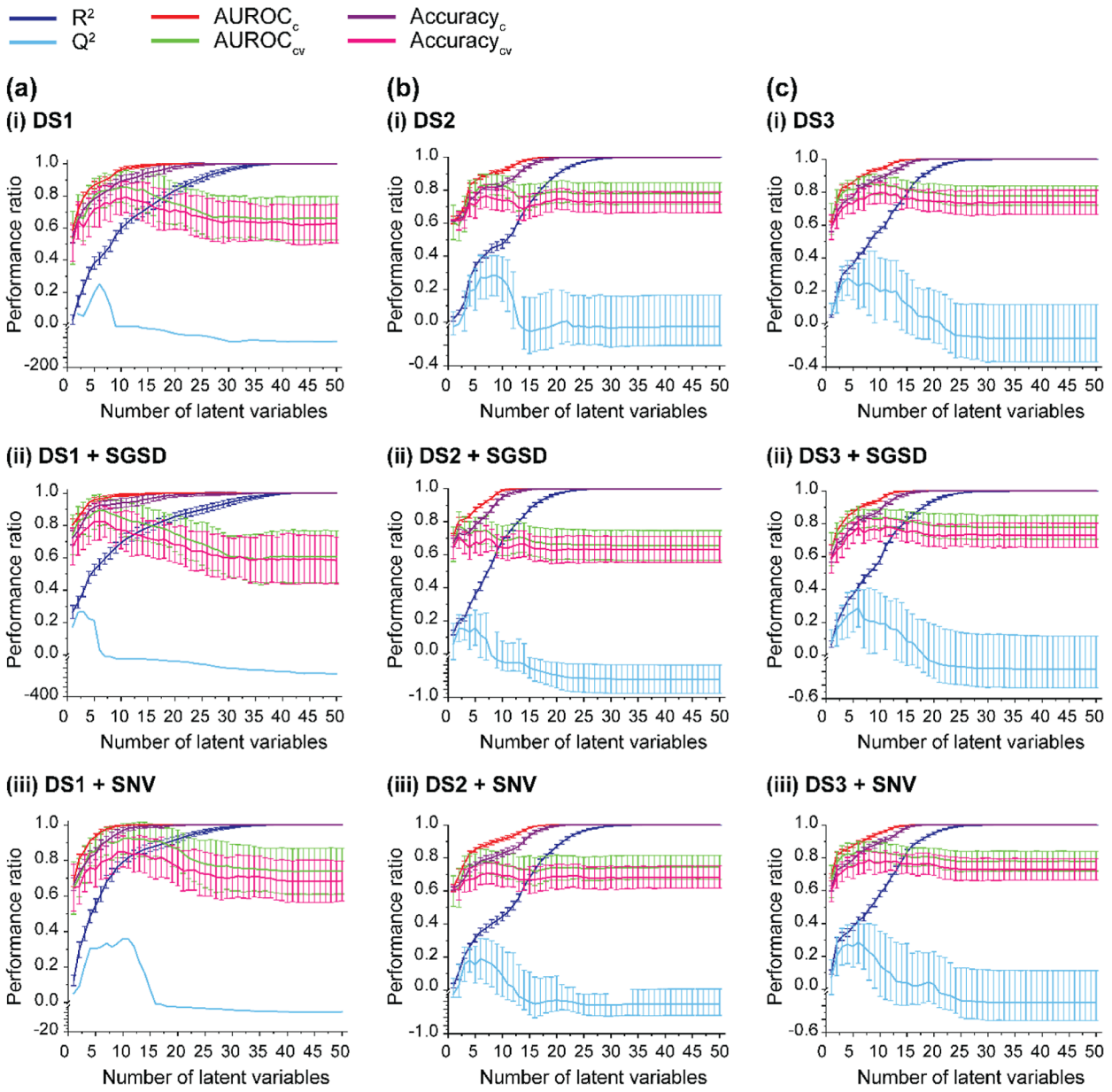


Figure 1. MCCV performance plot. a(i), b(i), and c(i) show the performance plots for the PLS-DA model using three real datasets DS1, DS2, and DS3, respectively. a(ii), b(ii), and c(ii) represent the performance of the PLS-DA models using the Savitzky−Golay second derivative (SGSD) spectra of DS1, DS2, and DS3, respectively. Similarly, a(iii), b(iii), and c(iii) represent the performance for the models using standard normal variate (SNV)-transformed DS1, DS2, and DS3, respectively. The vertical bar shows the standard deviation value of the 50 iterations during MCCV. In a(i), a(ii), and a(iii), standard deviation for Q^2 is not shown because larger values make the plot difficult to interpret. R^2, $AUROC_c$, and $accuracy_c$ show the calibration matrics. Q^2, $AUROC_{cv}$, and $accuracy_{cv}$ represent the CV metrics.

authors successfully developed a genetic algorithm-linear discriminant analysis (GA-LDA) classifier trained using the FTIR spectra acquired from the pharyngeal swab to screen COVID-19-positive (+ve) and COVID-19-negative (−ve) patients.[9] Later, Wood et al. (2021) developed a similar high-throughput diagnosis model using the partial least squares-discriminant analysis (PLS-DA) model. Here, authors used FTIR spectra of saliva samples collected from COVID-19 +ve and COVID-19 −ve patients, later determined using RT-PCR to train the PLS-DA model. The model achieved a sensitivity of 93% and a specificity of 82%.[10] Also, Banerjee et al. (2021) showed that FTIR-based PLS-DA models were paramount in diagnosing COVID-19 severe and non-severe patients. These models trained using the FTIR spectra of blood plasma samples and clinical information demonstrated an area under the receiver operating characteristic (AUROC) of 85.7%.[11] Moreover, several other studies discussed the advantage of the rapid diagnosis using chemometrics and machine learning models trained using the FTIR spectra of biofluids.[12−15]

All these studies used a small set of sample cohorts (approximately less than 200 samples) to construct the diagnosis model, and the models demonstrated significant efficiency. Still, a limitation is that these studies failed to explain the robustness of the diagnosis models, that is, determining the overfitting trend, estimating the best models, and, importantly, identifying the statistical significance of the models.[16−18] Since PLS-DA models were prone to overfitting

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

Table 1. MCCV Performance Metrics (Mean ± Standard Deviation) for the Best Diagnosis Models Trained Using Real Datasets

dataset/preprocessed dataset	LV	Q^2 (%)	MCR (%)	AUROC (%)	MCC (%)
PLS-DA using DS1					
DS1	6	25.24 ± 17.08	26.90 ± 12.43	82.52 ± 9.91	47.88 ± 25.49
DS1 + SGSD	2	26.37 ± 14.82	27.25 ± 11.11	82.62 ± 9.18	47.05 ± 23.60
DS1 + SNV	3	20.78 ± 21.33	27.25 ± 10.35	77.83 ± 10.53	48.83 ± 20.37
DS1 + EMSC	6	26.44 ± 55.37	18.75 ± 10.57	91.36 ± 7.20	63.96 ± 20.43
PLS-DA using DS2					
DS2	3	6.34 ± 12.50	36.82 ± 6.18	65.62 ± 10.78	18.01 ± 14.34
DS2 + SGSD	2	15.43 ± 8.18	29.10 ± 7.10	77.39 ± 8.64	37.13 ± 16.63
DS2 + SNV	3	15.67 ± 10.38	33.84 ± 6.99	73.78 ± 7.40	27.59 ± 15.19
DS2 + EMSC	4	19.27 ± 7.16	32.13 ± 6.63	76.90 ± 6.52	34.55 ± 14.00
PLS-DA using DS3					
DS3	3	24.35 ± 10.36	30.08 ± 5.83	79.64 ± 5.85	38.94 ± 12.35
DS3 + SGSD	3	20.01 ± 10.38	33.25 ± 7.26	76.37 ± 7.62	32.28 ± 15.64
DS3 + SNV	3	25.93 ± 11.13	28.56 ± 6.09	81.05 ± 6.38	42.07 ± 12.78
DS3 + EMSC	2	24.29 ± 11.09	28.61 ± 6.54	80.08 ± 6.55	41.87 ± 13.68
PLS-DA using DS2-FR					
DS2-FR	4	17.65 ± 9.13	29.69 ± 7.3	78.42 ± 8.01	38.82 ± 15.97
DS2-FR + SGSD	2	15.21 ± 8.10	29.10 ± 7.4	77.25 ± 8.48	37.11 ± 17.36
DS2-FR + SNV	4	18.24 ± 7.17	34.28 ± 7.38	74.95 ± 7.14	30.98 ± 16.28
DS2-FR + EMSC	3	20.13 ± 8.31	31.79 ± 8.45	77.73 ± 7.51	35.21 ± 18.18
PLS-DA using DS3-FR					
DS3-FR	3	26.88 ± 9.84	27.44 ± 6.16	81.05 ± 6.05	44.29 ± 12.95
DS3-FR + SGSD	3	19.81 ± 10.32	32.86 ± 7.29	76.37 ± 7.49	32.98 ± 15.72
DS3-FR + SNV	3	23.12 ± 11.32	28.32 ± 5.24	79.54 ± 6.70	42.48 ± 10.97
DS3-FR + EMSC	3	21.67 ± 12.29	29.25 ± 6.49	78.27 ± 7.01	40.60 ± 13.55

and producing over-optimistic results,[19] the current study systematically evaluated the robustness of models trained using real and augmented datasets. Augmented datasets contain artificial spectra calculated using extended multiplicative scatter augmentation (EMSA).[20] The advantage of using EMSA is that the n number of new spectra similar to the original spectrum could be augmented. Also, augmented spectra constitute only the physical variations (baseline shift, multiplicative effect, and instrumental and scattering effect) associated with the original spectra.[20] For biochemical variations in the spectra, spectra of new samples have to be added. However, this study exploited EMSA to generate varying-sized augmented datasets to investigate the influence of sample size on the prediction efficiency of the diagnosis models. Also, the robustness analysis was carried out for each of these models. For robustness analysis, the study was divided into two sections. First, the overfitting trend and the best classification models were estimated using Monte Carlo cross-validation (MCCV),[21] and then, the statistical significance of the best models was evaluated using a permutation test.[22,23]

RESULTS AND DISCUSSION

Significance of COVID-19 Diagnosis Models Using Real Datasets. The overfitting trend and the best diagnosis models were identified using the MCCV performance plot. The advantage of MCCV is that the n number of random unbiased splits of the datasets was possible and hence reduces the risk of overfitting.[17,21] Also, Xu and Liang (2001) showed that MCCV is a good approach when dealing with a larger dataset.[21] Similarly, Wood et al., 2021 adopted the MCCV approach to evaluate the COVID-19 diagnosis models trained using a smaller dataset.[10] Thus, it is evident that MCCV could be used for smaller and larger datasets to estimate the best models.[10,17,21] Hence, the MCCV performance metrics of the best model describe the most robust and unbiased estimates of a model to diagnose a new patient.[10]

Figure 1 shows the MCCV performance plot for the models constructed using the three real datasets, dataset 1 (DS1), dataset 2 (DS2), and dataset 3 (DS3), and the corresponding preprocessed datasets. Models constructed using DS1 and DS2 were used to diagnose COVID-19 +ve or COVID-19 −ve patients, while models using DS3 were used to diagnose severe or non-severe COVID-19 +ve patients. From Figure 1a(i−iii),b(i−iii),c(i−iii), it is evident that the prediction error for calibration (R^2) (the ability of the model to diagnose a known patient) increases with the addition of latent variables (LVs) to construct the model, while prediction error for cross-validation (Q^2) (the model's capability to diagnose a new similar sample) first increases and then decreases with the addition of LVs; that is, the diagnosis efficiency of the model decreases when the number of LVs increases. This observation indicates that the classification model is overfitting the data.[24,25] Therefore, the number of LVs to construct the best model (models without overfitting) was optimized such that $R^2 \approx Q^2$. For instance, consider Figure 1a(i). A minimum error between R^2 and Q^2 was observed for the model constructed using the first six LVs (optimal LVs), which was considered the best diagnosis model. A similar trend could be observed for AUROC from calibration ($AUROC_c$) versus AUROC from MCCV ($AUROC_{cv}$) curves. The same observation holds with accuracy from calibration ($accuracy_c$) versus accuracy from MCCV ($accuracy_{cv}$) curves.

Furthermore, the performance metrics were evaluated for the best models trained using DS1, DS2, and DS3 and the corresponding preprocessed datasets. Table 1 represents the MCCV performance metrics of the best diagnosis models. The

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

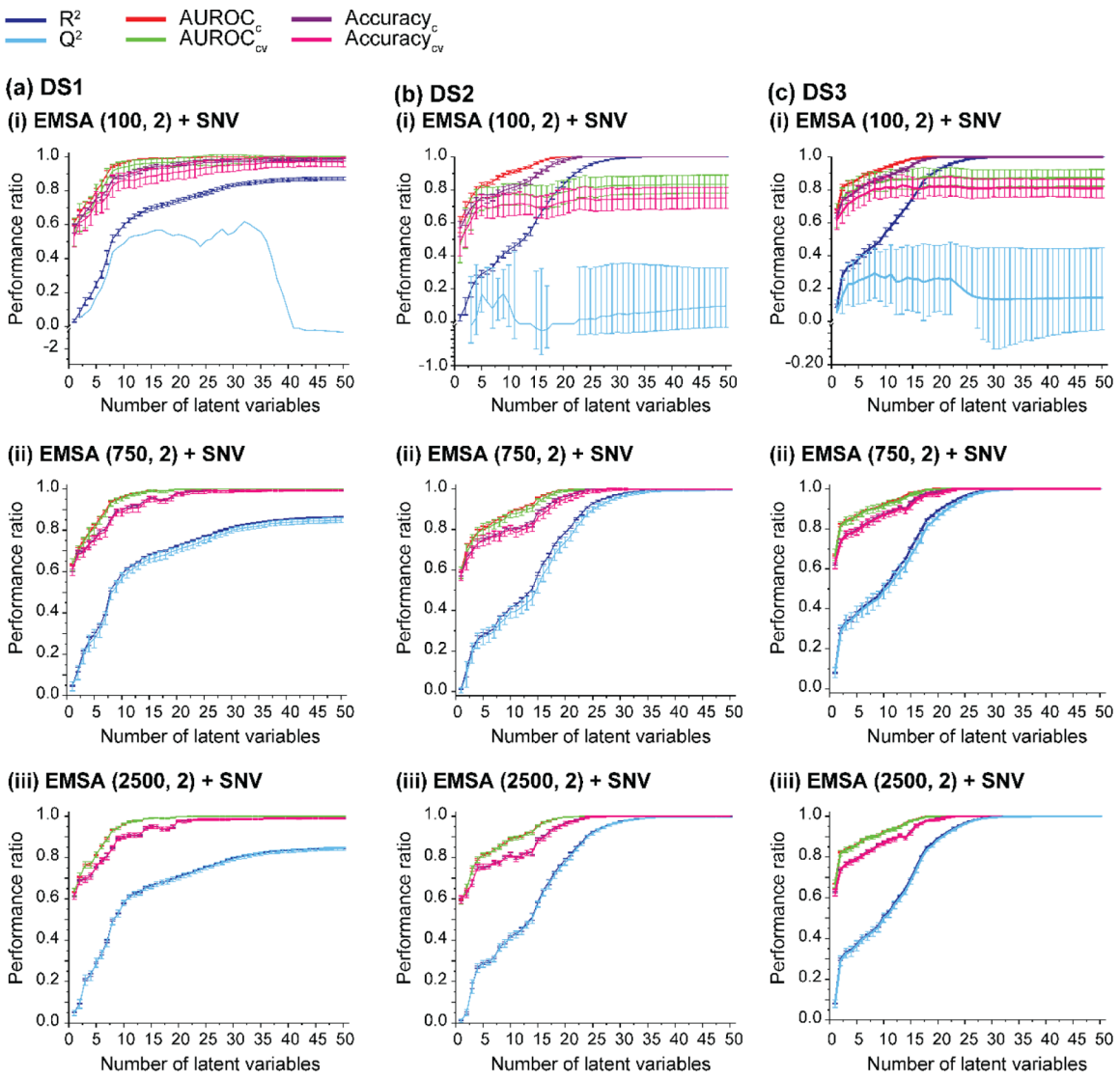


Figure 2. MCCV performance plot of the PLS-DA diagnosis model trained using SNV-transformed augmented datasets DS1, DS2, and DS3. (a) (i), (b) (i), and (c) (i) show the performance of the model using the augmented datasets with 100 spectra in each class. Similarly, (a) (ii), (b) (ii), and (c) (ii) represent the performance of the model using the augmented datasets with 750 spectra in each class. Likewise, (a) (iii), (b) (iii), and (c) (iii) show the performance of the model using the augmented datasets with 2500 spectra in each class. R^2, $AUROC_c$, and $accuracy_c$ show the calibration matrics. Q^2, $AUROC_{cv}$, and $accuracy_{cv}$ represent the CV metrics.

table shows that the model trained using DS1 acquired an AUROC of 82.52 ± 9.91%. Also, for the models trained using extended multiplicative scatter correction (EMSC)-corrected DS1 (DS1 + EMSC), one could observe an AUROC of 91.36 ± 7.20%, a misclassification rate (MCR) of 18.75 ± 10.57%, and a Mathews correlation coefficient (MCC) of 63.96 ± 20.43%. Similarly, models trained using Savitzky−Golay second derivative (SGSD)- and EMSC-corrected DS2 show similar performance with an AUROC of 77.39 ± 8.64% and 76.90 ± 6.52%. Likewise, the COVID-19 severity diagnosis models trained using standard normal variate (SNV)-transformed DS3 and EMSC-corrected DS3 show similar performance with an AUROC of 81.05 ± 6.38 and 80.08 ± 6.55%, respectively.

For a small sample cohort, the performance of the COVID-19 and the severity diagnosis models was exceptional. Notably, the models trained using EMSA spectra of DS1 outperformed the DS2- and DS3-trained models. This finding was interesting because DS1 uses only the fingerprint region (FR) from 800 to 1300 cm^{-1} for training the models. DS2 and DS3 use the complete spectral range (CSR) from 650 to 4000 cm^{-1}. Therefore, to understand the influence of the FR on model performance, the FR (650−1800 cm^{-1}) of DS2 (DS2-FR)- and DS3 (DS3-FR)-trained models was constructed, as shown in Table 1. As seen from Table 1, the models using DS2-FR showed an AUROC of 78.42 ± 8.01%, which was significantly improved compared to that of the models trained using DS2 (65.62%). Also, the models using DS3-FR showed an AUROC of 81.05 ± 6.05%. The results demonstrate a significant performance improvement between the models trained using the CSR and FR. Here, one could conclude that for the smaller dataset, the models constructed using the FR (DS1, DS2-FR, and DS3-FR) show improved performance. However, all the models show a high standard deviation and low Q^2 values. A Low Q^2 values indicates the poor class prediction capability of

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

Table 2. MCCV Performance Metrics for the Best Diagnosis Models Trained Using the SNV-Transformed Augmented Datasets

dataset	LV	Q^2 (%)	MCR (%)	AUROC (%)	MCC (%)
PLS-DA using augmented spectra from DS1					
EMSA (50, 2) + SNV	7	27.47 ± 20.14	24.90 ± 7.84	85.06 ± 8.44	51.86 ± 15.88
EMSA (100, 2) + SNV	7	31.13 ± 12.06	23.54 ± 7.40	86.33 ± 6.75	53.66 ± 14.90
EMSA (250, 2) + SNV	10	56.45 ± 4.69	10.60 ± 2.73	95.21 ± 1.73	78.96 ± 5.41
EMSA (500, 2) + SNV	40	82.71 ± 1.64	1.37 ± 0.87	99.95 ± 0.05	97.31 ± 1.73
EMSA (750, 2) + SNV	40	83.64 ± 1.13	0.83 ± 0.41	100.0 ± 0.02	98.34 ± 0.81
EMSA (1000, 2) + SNV	40	83.84 ± 1.32	1.07 ± 0.52	100.0 ± 0.02	97.85 ± 1.02
EMSA (2500, 2) + SNV	40	83.15 ± 0.77	1.12 ± 0.34	99.95 ± 0.02	97.80 ± 0.68
PLS-DA using augmented spectra from DS2					
EMSA (50, 2) + SNV	2	−11.21 ± 14.45	53.59 ± 8.89	45.34 ± 10.18	−7.57 ± 18.73
EMSA (100, 2) + SNV	5	16.77 ± 10.61	29.30 ± 6.33	74.90 ± 7.36	43.50 ± 12.79
EMSA (250, 2) + SNV	5	24.94 ± 5.28	26.71 ± 4.33	77.93 ± 4.18	49.29 ± 8.78
EMSA (500, 2) + SNV	40	99.22 ± 1.09	0.05 ± 0.28	100.0 ± 0.02	99.90 ± 0.55
EMSA (750, 2) + SNV	40	99.32 ± 0.42	0.13 ± 0.23	100.0 ± 0.00	99.76 ± 0.46
EMSA (1000, 2) + SNV	40	99.46 ± 0.23	0.04 ± 0.10	100.0 ± 0.00	99.90 ± 0.21
EMSA (2500, 2) + SNV	40	99.56 ± 0.09	0.00 ± 0.00	100.0 ± 0.00	100.0 ± 0.00
PLS-DA using augmented spectra from DS3					
EMSA (50, 2) + SNV	2	14.00 ± 17.29	39.31 ± 10.44	72.22 ± 11.83	21.83 ± 21.28
EMSA (100, 2) + SNV	2	15.41 ± 10.86	33.20 ± 7.10	73.68 ± 7.46	34.06 ± 14.23
EMSA (250, 2) + SNV	8	41.21 ± 6.73	17.14 ± 4.17	89.06 ± 2.93	66.02 ± 8.29
EMSA (500, 2) + SNV	40	99.95 ± 0.03	0.00 ± 00.00	100.0 ± 0.00	100.0 ± 0.00
EMSA (750, 2) + SNV	40	99.95 ± 0.01	0.00 ± 00.00	100.0 ± 0.00	100.0 ± 0.00
EMSA (1000, 2) + SNV	40	99.95 ± 0.00	0.00 ± 00.00	100.0 ± 0.00	100.0 ± 0.00
EMSA (2500, 2) + SNV	40	99.95 ± 0.00	0.00 ± 00.00	100.0 ± 0.00	100.0 ± 0.00

the PLS-DA diagnosis models. However, a Q^2 value for a good model is unknown.[18]

Increasing Dataset Size Reduces the Overfitting of Diagnosis Models. Conventionally, PLS-DA classification was adopted for colinear datasets with fewer samples and many variables.[26,27] Also, it is challenging to provide an appropriate sample size criterion to construct an accurate PLS-DA classification model.[18] Another major weakness of PLS-DA is the trend of overfitting.[16,18] At this juncture, the study discusses the influence of dataset size on the overfitting trend of models. Therefore, augmented datasets with varying sizes, EMSA(50, 2) (represent an EMSA dataset with 50 spectra in each class), EMSA(100, 2), EMSA(250, 2), EMSA(500, 2), EMSA(750, 2), EMSA(1000, 2), and EMSA-(2500, 2), and the corresponding SGSD- and SNV-transformed datasets were used to construct PLS-DA models. Furthermore, the MCCV performance plot of models using augmented datasets was used to understand the overfitting trend. The EMSC preprocessing of the datasets was not considered because EMSC is an intermediate step to generate the augmented spectra and therefore influences these spectra.[20]

The performance plots for all the models are shown in Figures 2 and S1−S6. For instance, consider Figure 2a(i),b-(i),b(i),c(i), the performance of the model using SNV-transformed EMSA(100, 2) of DS1, DS2, and DS3. Here, the trend of R^2 versus Q^2 demonstrates overfitting, which is comparable to the observations from the real datasets DS1, DS2, and DS3 (Figure 1). However, as seen in Figure 2(ii),b(ii),c(ii), models trained using SNV-transformed EMSA-(750, 2) of DS1, DS2, and DS3, the overfitting trend starts diminishing or is absent. Furthermore, if one observes R^2 versus Q^2 for models using an extensive dataset, EMSA(2500, 2), shown in Figure 2a(iii),b(iii),c(iii), overfitting is absent. Here, one could argue that the number of samples was greater for EMSA (2500, 2) than the number of variables, hence no overfitting. However, overfitting is absent for models trained using the SNV-transformed EMSA(500, 2), where the number of samples is less than the number of variables [Figure S3a(iv),b(iv),c(iv)]. Also, from Figures 2a(i),b(i),c(i), S1, and S2, it could be observed that models constructed using a smaller dataset (≈< 500 spectra in each class) show trends of overfitting. Conversely, as seen from Figures 2a(ii,iii),b(ii,iii),c-ii,iii),c(ii,iii) and S3−S6, increasing the dataset size (≈≥ 500 spectra in each class) reduces overfitting and improves the performance of the models.

Increasing Dataset Size Eliminates the Need for Variable Selection. Studies showed that for smaller datasets, variable selection improves COVID-19 diagnosis model performance,[10,12−14] which is evident in Table 1. However, the performance of models using a larger dataset contradicts this observation. As seen from Figure 2a(ii), for the performance of models using the SNV-transformed EMSA (750, 2) of DS1, which contains only the FR as variables (selected variables), the Q^2 curve approaches ≈85% and then flattens. Also, for the models using SNV-transformed EMSA (750, 2) of DS2 and DS3, which contain the whole spectral region (complete variables) shown in Figure 2b(ii),c(ii), Q^2 reaches ≈99.9% and flattens. Similar patterns were observed for models using the remaining datasets (Figures S1−S6). Therefore, models constructed using the entire spectral region outperform the models using selected variables. Thus, it was evident that variable selection was inessential for modes constructed using large datasets, and PLS-DA inherently identifies the variables of utmost importance.

Increasing Dataset Size Improves Model Performance. The best diagnosis models constructed using the augmented datasets were identified, and the performance of

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

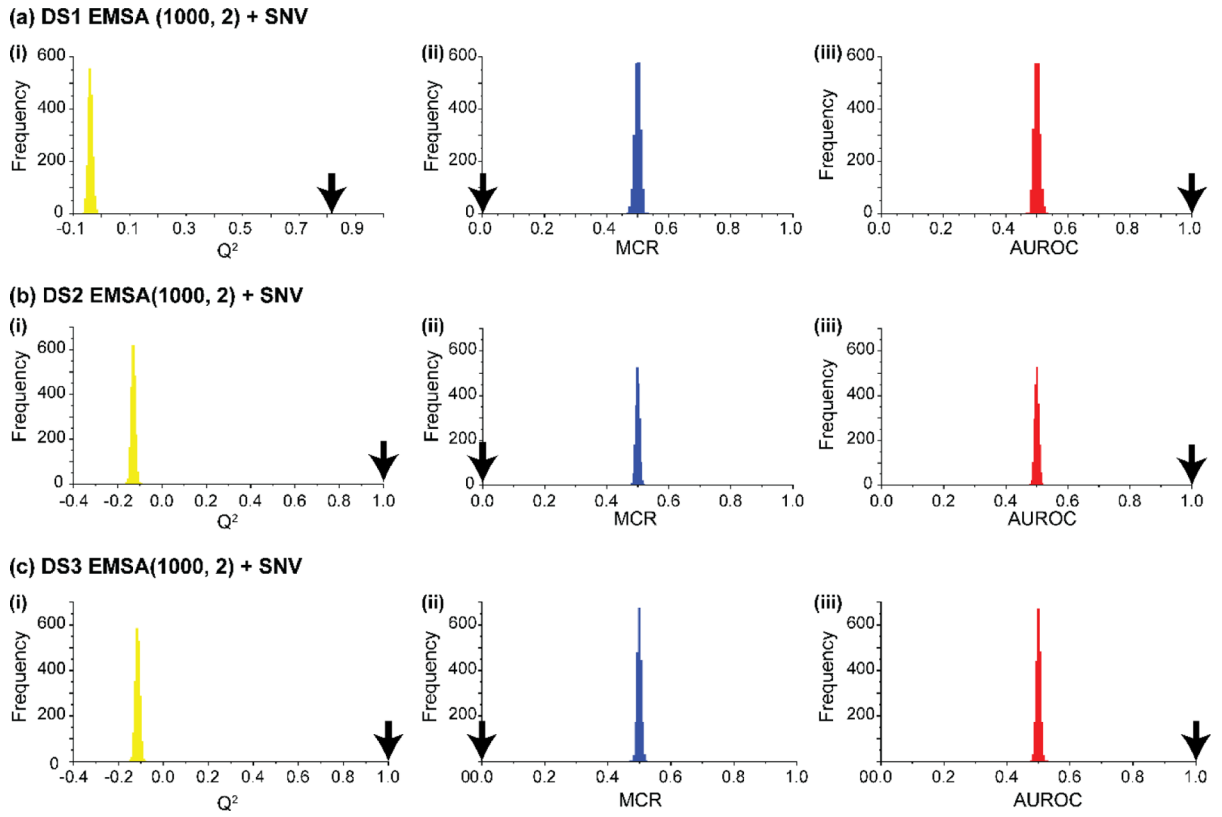


Figure 3. Quality assessment of the best PLS-DA diagnosis model based on MCCV using a permutation test. (a−c) represent the distribution for the test statistics Q^2, AUROC, and the MCR from the permuted model compared to that from the actual model (bold down arrow) trained using the SNV-transformed EMSA (1000, 2) of DS1, DS2, and DS3.

each model was computed. Table 2 shows the performance of models trained using the SNV-transformed augmented datasets of DS1, DS2, and DS3. Similarly, models trained using all the augmented datasets and the preprocessed augmented datasets of DS1, DS2, and DS3 are shown in Tables S1−S3. From Tables 2 and S1−S3, one could observe a low Q^2 value for the models trained using small augmented datasets EMSA(50, 2), EMSA(100, 2), and EMSA(250, 2) of DS1, DS2, and DS3 and the corresponding preprocessed datasets. Interestingly, this observation correlates with the Q^2 values obtained for the real datasets DS1, DS2, and DS3 (Table 1). Also, if one compares the trend of the Q^2 value for models trained using smaller and larger datasets, models using larger datasets show promising performance. More precisely, models using the SNV-transformed larger augmented datasets were the best.

For instance, consider Table 2; the diagnosis model using the SNV-transformed smaller dataset EMSA(50, 2) of DS1 shows a Q^2 of 27.47% with a major deviation of ±20.14%. Furthermore, one could observe an increase in the Q^2 value and a reduction in the deviation with an increase in the size of the dataset. Also, from the diagnosis model using SNV-transformed larger datasets of DS1, one could notice a fractional change in the Q^2 value (≈83%) and negligible deviations. A similar pattern could be seen in the models using the SNV-transformed larger datasets of DS2 and DS3 with a Q^2 value greater than 99% with negligible deviation. However, one could observe a difference in the Q^2 value between models using DS1 (≈83%) and the other two datasets (≈99%). The difference is due to the dissimilarity in the spectral region used to train the models. DS1 used the FR from 800 to 1300 cm^{-1}. DS2 and DS3 used the CSR from 650 to 4000 cm^{-1}. High Q^2 indicates that the classifier predicts class labels precisely with a minor variation from the original label. For example, if Q^2 is high (≈99%), then the classifier predicts the actual label of 1 as ≈0.99 or ≈1.01. Since the PLS-DA classifier predicts the class label based on PLS regression, attaining a Q^2 of 100% (predicting the same class label) was challenging.[28]

Contrary to Q^2, all three models show an AUROC of 100% with negligible or no deviation. Therefore, AUROC suggests that the models constructed using the larger datasets were exceptional. A similar pattern was observed for the MCR and MCC in the three models. For models trained using a larger dataset, the MCR ≈ 0%. The MCR indicates the fraction of the wrongly classified. Also, studies pointed out that too few misclassifications were achieved when the cross-validation method was implemented wrongly or the model had a poor quality.[18] Therefore, it is essential to check the quality of the model using a permutation test.[18,29]

Quality Assessment of the Best PLS-DA Diagnosis Models. A permutation test was adopted from refs 181923, and 30 to validate the best diagnosis models. Figure S7 shows the distribution of test statistics Q^2, AUROC, and the MCR for the permutation models and the actual model (bold down arrow) based on the MCCV for the real dataset. Similarly, Figures S8 and S9 show the distribution for the permutation models for smaller augment datasets EMSA (50, 2) and EMSA (100, 2) of DS1, DS2, and DS3. In Figure S9a(i), the actual diagnosis model shows a Q^2 of 31.13 ± 12.06%. Here, out of

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

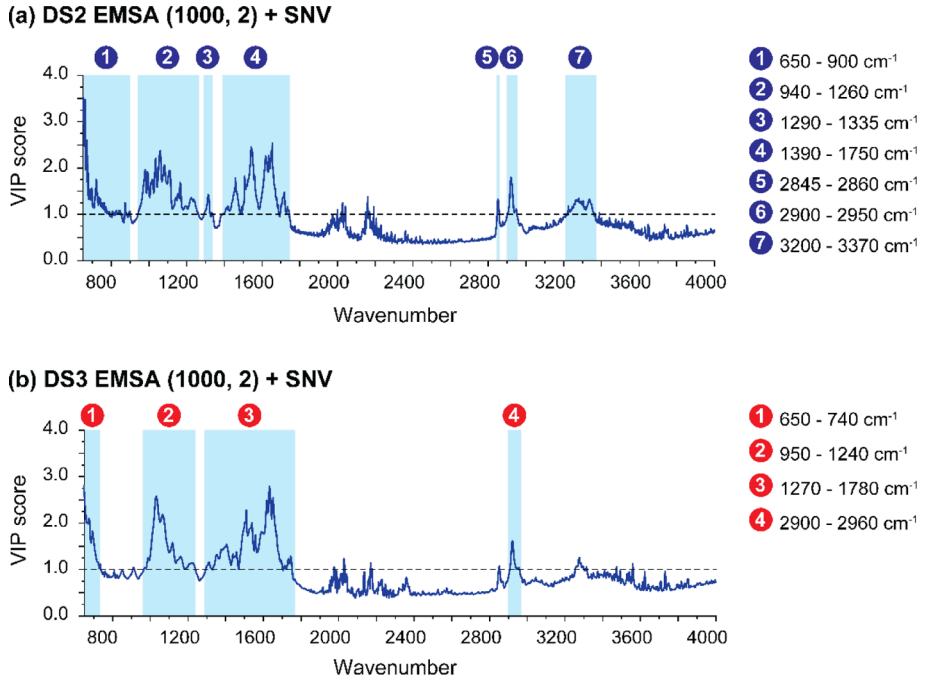


Figure 4. VIP score plots for COVID-19 diagnosis model (a) and COVID-19 severity diagnosis model (b) using SNV-transformed EMSA (1000, 2) of DS2 and DS3.

Table 3. MCCV Performance Metrics for the Best Diagnosis Models Trained Using the FR (650−1800 cm^{-1}) of Augmented Spectra

dataset	LV	Q^2 (%)	MCR (%)	AUROC (%)	MCC (%)
PLS-DA using augmented spectra from DS2-FR					
EMSA (1000, 2)	40	90.19 ± 0.86	0.05 ± 0.17	100.0 ± 0.00	99.95 ± 0.35
EMSA (1000, 2) + SGSD	40	89.21 ± 0.87	0.10 ± 0.32	100.0 ± 0.00	99.80 ± 0.64
EMSA (1000, 2) + SNV	40	93.21 ± 0.65	0.05 ± 0.22	100.0 ± 0.00	99.90 ± 0.44
PLS-DA using augmented spectra from DS3-FR					
EMSA (1000, 2)	40	92.53 ± 0.83	0.10 ± 0.30	100.0 ± 0.00	99.85 ± 0.59
EMSA (1000, 2) + SGSD	40	86.28 ± 1.31	0.59 ± 0.81	100.0 ± 0.06	98.88 ± 1.59
EMSA (1000, 2) + SNV	40	93.75 ± 0.75	0.10 ± 0.32	100.0 ± 0.00	99.80 ± 0.63

the 2000 permutation models, none had Q^2 higher than 31.13% leading to a *P*-value of < 0.0005.[29] Similar results could be observed in the Q^2 distributions for SNV-transformed EMSA (100, 2) of DS2 [Figure S9b(i)] and DS3 [Figure S9c(i)].

Comparable results were observed in the distribution of the MCR. The MCR was employed as a metric because of its known expected value in the randomly permuted scenario. The average number of misclassifications for a randomly permuted two-class problem should be half of the samples; that is, the average MCR equals 0.5 (50%),[18] which is evident from the distribution of the MCR, shown in Figure S9a(ii). As seen from the figure, none of the permutation models shows an MCR lower than 23.54 ± 7.40% leading to a significant *P*-value of <0.0005. Similarly, the models trained using the SNV-transformed EMSA(100, 2) of DS2 [Figure S9b(ii)] and DS3 [Figure S9c(ii)] provide a similar *P*-value of <0.0005, showing that these models are significant. Furthermore, consistent results were observed while comparing the permuted distribution for the metric AUROC.

All these findings were in correlation with the real dataset (Figure S7) and the smaller augmented datasets EMSA(50, 2) (Figure S8). Thus, it is crucial to note that models using smaller datasets (real and augmented) are of good quality. However, the drawback is that these models have low performance. Conversely, if one investigates the distribution of the permuted models for larger augmented datasets, shown in Figures 3 and S10−S12, all these models show a significant *P*-value of <0.0005 for Q^2, the MCR, and AUROC. The findings suggest that these models were of good quality. Also, the results were predominant because the COVID-19 diagnosis model and the COVID-19 severity diagnosis models using larger datasets demonstrated excellent performance (Q^2 > 99%).

VIP Scores Highlight the Most Relevant Spectral Variables. Variable importance in projection (VIP) scores[11,31] were employed to identify the spectral variables of utmost importance. It describes the relative influence of the *X* variable (wavenumber) on the dependent variables (*Y*) and the LVs. On the other hand, some studies used regression coefficients for the same purpose.[29] However, VIP scores are advantageous compared to regression coefficients because vital spectral variables contributing to inter-class variations could be efficiently determined for models trained using preprocessed spectra.[31] Thus, the importance of *X* could be evaluated using the VIP score. Variables with a VIP score equal to 1

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

correspond to an equal contribution of these variables, and a score greater than 1 demonstrates a higher significance to those variables.

The VIP score for the best models trained using SNV-transformed EMSA(1000, 2) of DS2 and DS3 is shown in Figure 4a,b. From the figures, one could observe that the (FR) (650−1800 cm^{-1}) accounts for the most significant vibrations (VIP ≥ 1) in both models. Furthermore, models were constructed using the FR to investigate the influence of this region on the model performance. Table 3 shows the performance of the models constructed using the FR of EMSA(1000, 2). The table suggests that the class prediction capability of the model in terms of Q^2 was less (≈93%) compared to that of the models using the complete region of spectra (Q^2 ≈ 99%, shown in Table 2). Therefore, it is evident that the variation in the spectral regions other than the FR, that is, the regions 2845−2860, 2900−2950, and 3200−3370 cm^{-1} in Figure 4a and the region 2900−2960 cm^{-1} in Figure 4b, influences the class prediction efficiency. However, both the models constructed using the FR show an AUROC of 100% and a less than 0.1% MCR.

Furthermore, closely observing spectral regions in Figure 4a (labeled 1, 2, ..., 7 in blue circles) shows that region 1 from 650 to 900 cm^{-1} provides significant variation to the COVID-19 diagnosis model. However, in the region 1 (red circle) in Figure 4b, the VIP plot for the COVID-19 severity model was influenced only by 650 to 740 cm^{-1}. These observations show that the fundamental vibrations influencing the performance of COVID-19 diagnosis and severity diagnosis models were not similar. Also, these observations were valid for the remaining regions. However, to understand the biochemical explanation, an extensive database comprising the spectral band assignments of FTIR spectra of biological and viral samples from 600 to 4000 cm^{-1} was constructed based on a literature survey[10,13−15,32−42] and is available in Table S4. The following discusses the biomolecules responsible for the significant vibrations.

650−750 cm^{-1}. Amide V vibrations of proteins; vibrations of protein secondary structures α-helix, β-sheets, β-turns, 3_{10} helices.

750−900 cm^{-1}. DNA or RNA conformational-related vibrations, including nucleotide vibrations; ring, C−C and C−O vibrations of ribose and deoxyribose sugar. Phosphate and sugar-phosphate vibrations; C−C and C−O vibrations of carbohydrates and fatty acids.

900−1000 cm^{-1}. Mainly, DNA, RNA, and DNA-RNA hybrid vibrations; P−OH bending and symmetrical stretching (ν_s) of PO_4^{2-} in phosphorylated proteins and nucleic acids.

1000−1100 cm^{-1}. This region was majorly influenced by ν_s of PO_4^{2-} present in DNA, RNA, phosphorylated proteins, and molecules in energy metabolism and membrane phospholipids; C−C and C−O vibrations of polysaccharides; ν(C−N) and C−H deformation (δ) vibrations of amino acids histidine and tryptophan, CH_2 vibration of proline, and C−O vibration of threonine.

1100−1200 cm^{-1}. Includes ν(P−O−C), ν(C−O), ν(C−OH), and ν (C═O) vibrations of the ribose sugar; asymmetrical stretching vibrations (ν_{as}) of CO−O−C in DNA, RNA, and glycans; ν(C−O), ν(C−C), and ν(C−O−C) vibrations of polysaccharide rings, phospholipids, triglycerides, and cholesterol esters; ν(C−N) and δ(C−H) vibrations of deprotonated and protonated histidine; ν(C−O) and ν(C−OH) groups of serine, threonine, and tyrosine in proteins.

1200−1300 cm^{-1}. $\nu_{as}(PO_4^{2-})$ of the phosphodiester linkage in A-DNA, B-DNA, RNA, and phospholipids; ν_{as}(P═O) of the phosphorylated molecule; ring C−O−C, C−O vibrations of polysaccharides; amide III vibrations of nucleic acids and proteins; ν(C−C), ν(C−O) tryptophan, tyrosine; δ(C−H), ν(C−N), δ(N−H) of histidine, tryptophan; δ(C−OH) vibrations of tyrosine, aspartate, glutamate; protein secondary structure vibrations of α-helix, β-sheets, β-turns, and 3_{10} helices.

1300−1400 cm^{-1}. δ(C−H) in polysaccharide rings; amide III vibrations of proteins and nucleic acids; wagging(ω) and twisting (τ) vibrations of CH_2 in proteins; $\nu_s(COO^-)$ vibrations of fatty acids and amino acid side chains. $\delta_s(CH_3)$ and $\delta_s(CH_2)$ of lipids and proteins.

1400−1500 cm^{-1}. ν(C−N), δ(N−H), δ(C−H) in proteins; $\nu_s(COO^-)$ of lipids, polysaccharides, and proteins; δ(C−H) of DNA, amino acids, and proteins; $\delta(CH_2)$ of amino acids, lipids, fatty acids, and polysaccharides; $\delta_{as}(CH_3)$ of proteins; scissoring vibrations of (σ) CH_2 in phospholipids.

1500−1600 cm^{-1}. Mainly, amide II vibrations of nucleic acids, proteins, α-helix, β-sheets, β-turns, and 3_{10} helices; δ(C═N), NH_2 of nucleic acids; δ(C−H), δ(C−C) ring, ν(C−C), ring-OH, δ(C═N), δ(C═C), ν(C═C), $\nu_s(COO^-)$ of amino acids and proteins; $\delta_s(NH_3^+)$ of lysine; $\delta(-NH_2)$ of N-terminal.

1600−1700 cm^{-1}. Mainly, amide I vibrations of nucleic acids, proteins, α-helix, β-sheets, β-turns, and 3_{10} helices; C═C ring, ν(C═O), ν(C═N), ν(C═C), ν(N−H), and NH_2 of nucleobases in single-stranded (ss) and double-stranded (ds) DNA and RNA; ν(C═O) of nucleic acids, lipids, fatty acids, and proteins; $\nu_{as}(CN_3H_5^+)$, $\delta_{as}(-NH_3^+)$, ν(C−C), and δ(C−H), δ(C−C ring), δ(C−H), and $\delta_s(-NH_2)$ in amino acids and proteins.

1700−1750 cm^{-1}. Amide I vibrations; ν_s(C═O) and ν_s(C−N) of proteins; ν(C═O) of nucleotides, amino acids, DNA, RNA, fatty acid, lipids, phospholipids, and carbonyl esters of lipids.

2900−2960 cm^{-1}. ν(C−H) and ν(N−H) vibrations; $\nu_{as}(CH_2)$ and $\nu_{as}(CH_3)$ of lipid acyl chains.

Validation of the Best Diagnosis Models. The best diagnosis model constructed using the SNV-transformed larger augmented datasets was validated using an independent validation dataset which was not used for calibrating the models. The validation dataset was constructed using the EMSA approach with 50 spectra each in COVID-19 +ve and COVID-19 −ve cases. Table 4 represents the performance metrics obtained during the validation process. The results show that increasing the number of samples after a certain threshold does not improve prediction efficiency. For example, the models constructed using the SNV-transformed larger dataset EMSA(1000, 2) of DS1 show a prediction error for validation (V^2) value of 87.23%, and the models constructed using a more extensive set, that is, the SNV-transformed EMSA(2500, 2) of DS1, show a V^2 value of 86.91%. Also, the models constructed using the SNV-transformed larger dataset of DS3 show the same V^2 value of 99.97%. Furthermore, the PLS-DA model constructed using SNV-transformed EMSA-(1000, 2) of DS1, DS2, and DS3 shows an MCR of 0%, an AUROC of 100%, and an MCC of 100%. Therefore, the

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

Table 4. Validation of the Best Diagnosis Models

dataset	LV	V^2 (%)	MCR (%)	AUROC (%)	MCC (%)
PLS-DA using augmented spectra from DS1					
EMSA (500, 2) + SNV	40	86.37	0.00	100.00	100.00
EMSA (750, 2) + SNV	40	86.81	0.00	100.00	100.00
EMSA (1000, 2) + SNV	40	87.23	0.00	100.00	100.00
EMSA (2500, 2) + SNV	40	86.91	0.00	100.00	100.00
PLS-DA using augmented spectra from DS2					
EMSA (500, 2) + SNV	40	99.14	0.00	100.00	100.00
EMSA (750, 2) + SNV	40	98.82	1.00	98.02	100.00
EMSA (1000, 2) + SNV	40	99.21	0.00	100.00	100.00
EMSA (2500, 2) + SNV	40	99.03	0.00	100.00	100.00
PLS-DA using augmented spectra from DS3					
EMSA (500, 2) + SNV	40	99.97	0.00	100.00	100.00
EMSA (750, 2) + SNV	40	99.97	0.00	100.00	100.00
EMSA (1000, 2) + SNV	40	99.97	0.00	100.00	100.00
EMSA (2500, 2) + SNV	40	99.97	0.00	100.00	100.00

validation results suggest that the models constructed using larger datasets accurately diagnose COVID-19 and its severity.

CONCLUSIONS

The study critically evaluated the robustness of the FTIR-based COVID-19-infected and COVID-19 severity diagnosis model trained using real datasets (for a smaller cohort) and smaller augmented, larger augmented, and the corresponding preprocessed datasets. The MCCV performance plot of models trained using the real and the smaller augmented dataset shows a trend of overfitting. Furthermore, the best diagnosis models were identified, and a permutation test was conducted for each of these models using the test statistics Q^2, the MCR, and AUROC. The permutation test shows a P-value of less than 0.0005 for Q^2, the MCR, and AUROC, demonstrating that the models were significant. However, all these models show a low Q^2 value, indicating a poor class prediction capability. Moreover, a positive correlation between the MCCV performances of the models trained using real and smaller augmented datasets could be observed.

Furthermore, the study investigated the robustness of models trained using larger augmented datasets. Interestingly, the trend of overfitting is absent for these models, which is validated via MCCV. Also, the best diagnosis models show Q^2 and V^2 greater than 99%, an AUROC of 100%, and an MCR of less than 0.1%, indicating that the diagnosis is accurate. Also, the permutation test shows a P-value of less than 0.0005, implying that the models were significant, and the results suggest that variable selection is unimportant for models trained using a larger dataset. PLS-DA inherently identifies the variables of utmost importance.

Moreover, the VIP score plot obtained from these models shows that variables corresponding to the fundamental vibrations associated with phosphate, nucleotide, DNA, RNA, polysaccharide, lipid, protein, and amide I to V vibrations of these biomolecules predominantly contribute to the performance of the diagnosis model. Still, the relative contribution of these molecules in COVID-19 and COVID-19 severity diagnosis models was not the same. Therefore, by considering these results, it could be concluded that the PLS-DA diagnosis model trained using larger datasets opens up new prospects for rapid, accurate, significant, and cost-effective clinical diagnosis of COVID-19-infected and COVID-19 severe patients. However, the study used only artificial spectra computed using EMSA by introducing physical variations for models trained using larger datasets. Incorporating chemical variation into the model requires spectra from a real extensive cohort. Also, since the model only depends on the difference between the molecular signatures of healthy individuals and infected patients, similar models could be implemented to rapidly diagnose new variants or even for an unknown viral outbreak in the future.

METHODS

COVID-19 FTIR Spectroscopy Dataset 1 (DS1). DS1 was obtained from ref 10. For the study, Wood et al. (2021) collected the saliva of the patients admitted to the Royal Melbourne Hospital with COVID-19-like symptoms using a viral transport medium (VTM). Furthermore, these patients were determined to be COVID-19 +ve or COVID-19 −ve using RT-quantitative (q)PCR. Later, a PerkinElmer Spectrum Two spectrometer with a customized reflecting accessory designed for transflection slides was used to collect the FTIR spectra of these saliva samples. The phosphodiester region (800 cm^{-1} to 1300 cm^{-1}) of the FTIR spectra was publicly available, which includes the data for 30 COVID-19 −ve and 31 COVID-19 +ve patients. The phosphodiester region is crucial for the bands associated with many essential RNA and glycoprotein markers, and these regions were less affected by the interference from VTM. The FTIR spectra of the COVID-19 −ve and COVID-19 +ve patients are shown in Figure S13a.

COVID-19 FTIR Spectroscopy Dataset 2 (DS2). DS2 contains the FTIR dataset obtained from ref 9. Barauna et al. (2021) collected samples from six hospitals that participated in the study. All the individuals were first identified as COVID-19 −ve or COVID-19 +ve using RT-qPCR following the standard protocols using a nasopharyngeal swab collected and stored in a VTM. Later, a pharyngeal swab from the same participants was collected and stored in ice for FTIR data acquisition. Furthermore, the attenuated total reflection (ATR)-FTIR spectra were acquired using a portable Agilent Cary 630 FTIR spectrometer equipped with an ATR ZnSe crystal from 650 to 4000 cm^{-1} with a spectral resolution of 1.86 cm^{-1}. Thus, DS2 contains the ATR-FTIR spectra from 111 COVID-19 −ve and 70 COVID-19 +ve patients.

COVID-19 FTIR Spectroscopy Dataset 3 (DS3). DS3, collected from ref 11, includes the ATR-FTIR spectra of severe and non-severe COVID-19 +ve patients. Blood samples from patients admitted to Kasturba Hospital, Bombay, were collected by following the standard protocols for infection control provided by the WHO. Air-dried, ethanol-treated plasma isolated from the blood samples was used for spectral acquisition. A portable Agilent Cary 630 FTIR spectrometer equipped with a diamond crystal was used for spectral acquisition from 650 to 4000 cm^{-1}. Thus, DS3 contains FTIR data of 69 severe and 91 non-severe patients. The severity for a patient was determined primarily based on an O_2 saturation level of less than 90% and other clinical complications.

Spectral Derivatives. Derivatives of each spectrum were computed using the SG method.[43] The advantage of a derivative spectrum is that it can remove additive effects, enhance signal properties, resolve overlapping signals, and suppress unwanted spectral features that arise from samples and instruments. The SG first derivative (SGFD) and second

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

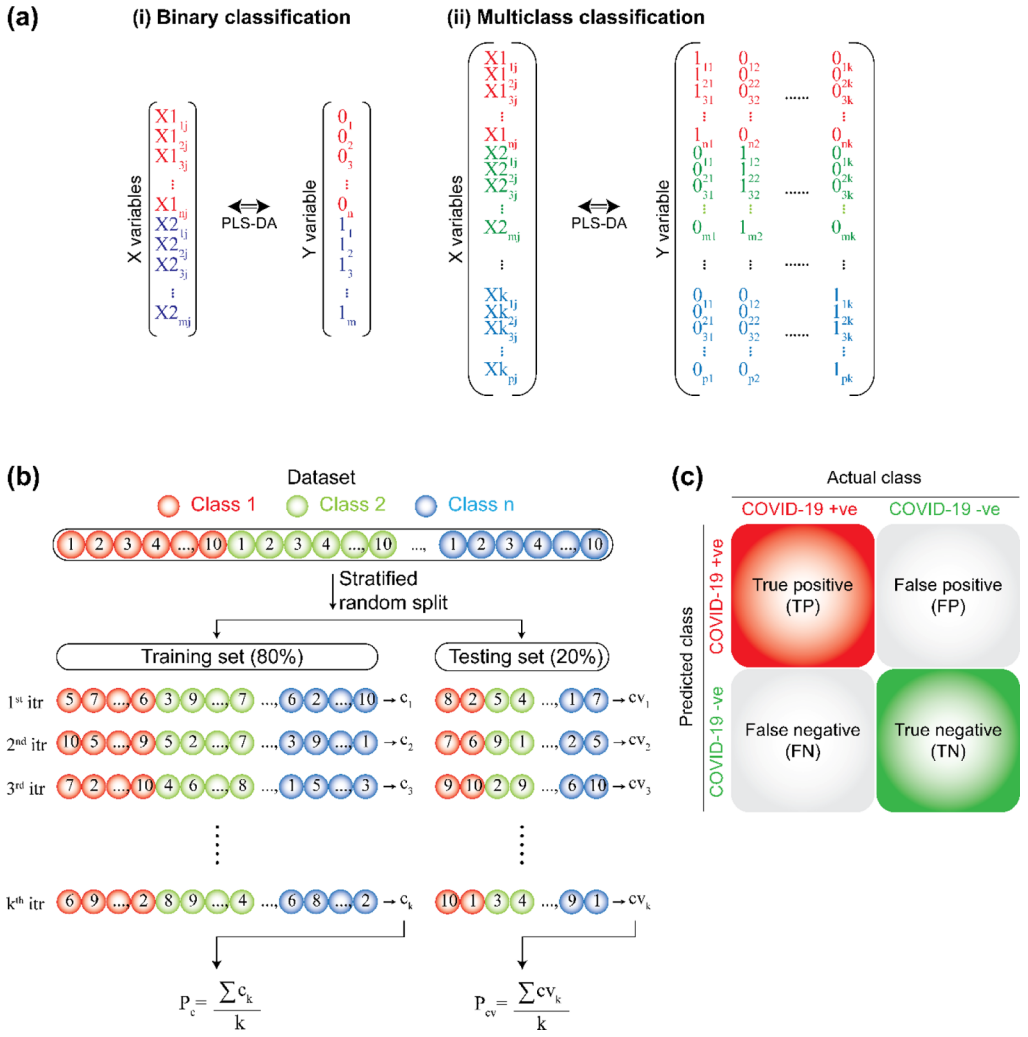


Figure 5. (a) Schematics of PLS-DA models for binary and multiclass classification. A(i) depicts the binary classification model. The X variables, a matrix of size $((n + m) \times j)$, represent the spectroscopic dataset (j wavenumbers in each spectrum) for class 1 and class 2 samples, with n number of spectra ($X1_{1j}$, $X1_{2j}$, $X1_{3j}$ ..., $X1_{nj}$) in class 1 and the m number of spectra ($X2_{1j}$, $X2_{2j}$, $X2_{3j}$ ..., $X2_{mj}$) in class 2 samples. The Y variable represents a vector containing the corresponding class labels 1 and 0. Similarly, a(ii) shows the multiclass classification model using the k class of samples. The X variable contains the spectroscopic dataset for the k class, a matrix of size $(n + m + ... + p) \times j)$, where n, m, ..., p are the number of spectra in the 1st, 2nd, ..., kth class. Similarly, the Y variable is an array of size $(n + m + ... + p) \times k)$ containing the categorical values corresponding to each class. (b) Schematic showing the MCCV with k random iterations, MCCV (k). Consider a sample dataset with n classes of samples where each class contains 10 spectra. In the first step, 80% of the data will be randomly divided into a training set, and the remaining 20% will be assigned as a testing set. The splitting is done in a stratified manner. Hence, in this case, each class will have eight spectra in the training set and two spectra in the testing set. This process will be iterated k times. The calibration and cross-validation performance will be recorded in each iteration. Finally, the average performance for the calibration (P_c) and cross-validation (P_{cv}) will be estimated to evaluate the performance of the model. (c) Schematic representation of a confusion matrix for the COVID-19 classification model.

derivative (SGSD) of the spectrum were computed using a second-order polynomial and 19 smoothing points.[44] A first derivative and second derivative spectrum for COVID-19 −ve and COVID-19 +ve spectra from DS1 can be observed in Figure S13b,c.

Standard Normal Variate Transformation. SNV was used because of its effectiveness in excluding the multiplicative effects of scattering and particle size.[45] SNV for an FTIR spectrum x_i could be defined using the formula $x_{i,\mathrm{SNV}} = \frac{(x_i - \bar{x})}{\sigma}$, where $x_{i,\mathrm{SNV}}$ is the SNV-transformed spectra; $\boldsymbol{x_i}$, absorbance for the i^{th} wavenumber; $\bar{x}$, the average of x_i; and σ, the standard deviation of the spectrum. A visualization of the SNV-transformed spectra for COVID-19 −ve and COVID-19 +ve patients from DS1 is shown in Figure S13d.

Extended Multiplicative Scatter Correction (EMSC). Multiplicative scatter correction (MSC) is a model-based method efficient in removing both additive and multiplicative interferant effects due to scattering and particle size. The advantage of the model is that the additive and multiplicative effects were parameterized before being removed from the spectra.[46] However, MSC does not consider wavenumber-dependent effects due to light scattering variation. The wavenumber-dependent effects are unknown but considered to be nonlinear but smooth functions of the wavenumber ($\overline{\nu}$). The EMSC model accounts for these effects by adding a liner and a quadratic wavenumber-dependent effect. Thus, the MSC

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

Table 5. Performance Metrics Used for Evaluating the Classification Model[a]

performance metrics	formula	worst value	best value
accuracy, CR	$\frac{TP + TN}{TP + TN + FP + FN}$	0	1
MCR	$\frac{FP + FN}{TP + TN + FP + FN}$	1	0
sensitivity, recall, TPR	$\frac{TP}{TP + FN}$	0	1
specificity, TNR	$\frac{TN}{TN + FP}$	0	1
1-specificity, fallout, FPR	$\frac{FP}{TN + FP}$	1	0
precision, PPV	$\frac{TP}{TP + FP}$	0	1
NPV	$\frac{TN}{TN + FN}$	0	1
MCC	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$	−1	1

[a]CR-classification rate; MCR-misclassification rate; TPR-true positive rate; TNR-true negative rate; FPR-false positive rate; PPV-positive predictive value; NPV-negative predictive value; and MCC-Mathew's correlation coefficient.

model could be modified to the EMSC model and could be written as $x_i(\bar{\nu}) = a_i + b_i \cdot \bar{x}(\bar{\nu}) + d_i \cdot \bar{\nu} + e_i \cdot \bar{\nu}^2$. Next, the parameters a, b, d and e describing the scatter effect of each spectrum could be estimated using the ordinary least squares regression. After estimating the parameters, each spectrum was corrected using the formula $x_{i,EMSC}(\bar{\nu}) = [\bar{x}(\bar{\nu}) - a - d \cdot \bar{\nu} - e \cdot \bar{\nu}^2]/b$.[46] EMSC models constructed for infrared spectroscopy generally consider the wavenumber-dependent parameters up to the second quadratic term.[20] Examples for EMSC-corrected spectra are shown in Figure S13f.

Data Augmentation. The EMSA method proposed by Blazhko et al. (2021)[20] helps increase the number of spectral samples in the dataset by augmenting artificial spectrums using a measured spectrum by introducing the physical variations. The method first estimates the physical parameters for the scattering and instrumental effect $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{d}$ and $\boldsymbol{e}$ using the EMSC model from the measured FTIR dataset and the corresponding standard deviations for each parameter. Next, a set of random deviations were generated for each parameter ($\Delta\boldsymbol{a}$, $\Delta\boldsymbol{b}$, $\Delta\boldsymbol{d}$ and $\Delta\boldsymbol{e}$) using a normal distribution with zero mean and the computed standard deviation. Furthermore, new artificial parameters $\boldsymbol{a}'$, $\boldsymbol{b}'$, $\boldsymbol{d}'$ and $\boldsymbol{e}'$ were obtained by adding the random deviations to the parameters of each measured spectrum. Thus, the augmented spectrum could be written as $\hat{\mathbf{x}}(\bar{\nu}) = \boldsymbol{a}' + \mathbf{b}' \cdot \bar{\mathbf{x}}(\bar{\nu}) + \mathbf{d}' \cdot \bar{\nu} + \mathbf{e}' \cdot \bar{\nu}^2 + \frac{\mathbf{e}(\bar{\nu}) \cdot \mathbf{b}'}{\mathbf{b}}$, where $\mathbf{e}(\bar{\nu})$ are the residuals. An augmented spectral dataset with n spectral samples each in the m class is represented as EMSA(n, m). For example, Figure S13g represents a small augmented spectral dataset computed using DS1 with 50 COVID-19 −ve and 50 COVID-19 +ve spectral samples, EMSA(50,2). Similarly, a more extensive set of augmented spectra computed using DS1 with 500 augmented spectra each in the COVID-19 −ve and COVID-19 +ve case (EMSA(500,2)) are shown in Figure S13h.

Partial Least Squares Discriminant Analysis. PLS-DA[27] uses the standard partial least squares regression (PLSR) method[47] to construct both binary and multiclass classification models. In the case of binary classification, the PLSR1 method is used where the independent variable (X) will be the FTIR spectroscopic data, and the dependent variables (Y) will be a vector of class labels. Usually, the control (COVID-19 −ve) will be assigned a class label 0, and the cases (COVID-19 +ve) will be given 1. Figure 5a(i) shows a schematic representation of the binary classifier. Here, the X variables $X1_{nj}$ and $X2_{mj}$ represent the FTIR spectra with j variables (wavenumbers) of the class 1 (n spectra) and class 2 (m spectra) sample, and the Y variables represent the corresponding vector of class labels 0 and 1. Likewise, Figure 5a(ii) represents the schematics of a multiclass (k classes) classifier. A multiclass PLS-DA uses the PLSR2 method, where the Y variables will be an array of multiple dependent categorical variables (the value one would like to predict for a particular class or group of spectral samples). For example, in case of a three-class classifier, the Y variable will be an array of the categorical variables [1, 0, 0], [0, 1, 0], and [0, 0, 1] for class 1, class 2, and class 3 samples, respectively.

The approach for binary and multiclass classification using PLS-DA is the same. Instead of directly relating the highly collinear X and Y variables, PLS-DA decomposed the X and Y variables to a new set of uncorrelated variables called scores and loadings, thus reducing the dimensionality of the original dataset and finding a linear subspace of explanatory variables such that the Y scores have maximum covariance with X scores. The new subspace called LVs allows the prediction of the class labels.[27] Since PLS-DA utilizes a regression method for predicting class labels, the predicted value $\hat{y}$ will be continuous (real number). Therefore, $\hat{y}$ is transformed into a class label (i.e., 0 for COVID-19 +ve or 1 for COVID-19 −ve) using a discrimination threshold (τ). The discrimination threshold is set to 0.5 because the two classes have similar

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

sizes, and Y is a vector of 0 and 1.[48] Furthermore, the class label will be assigned in such a way that

$$\text{class labels} = \begin{cases} 0, \text{ if} \hat{y} < \tau \\ 1, \text{ if} \hat{y} > \tau \end{cases}$$

Next, these class labels will be compared with the actual class labels to evaluate the model's performance.

Monte Carlo Cross-Validation. The PLS-DA model performance was evaluated using the MCCV method.[21] During MCCV, the training dataset was split into training (80% of data) and testing sets (the remaining 20%) at random using the stratified split method. The stratified method ensures that both train and test sets have a similar proportion of samples in each target class. Furthermore, PLS-DA classifiers were constructed using the training set (calibration) and evaluated the model's classification efficiency using the testing set (cross-validation). For each classification model discussed in this study, 50 random iterations were conducted, and an average of the performance metrics for calibration and cross-validation was recorded. A schematic representation of the MCCV method is shown in Figure 5b. Xu and Liang (2001)[21] suggest that if the calibration uses a fewer number of samples (n), a greater number of iterations (k) $k = n^2$ are needed, and the computational complexity could be reduced compared to the leave-k-out cross-validation method. An MCCV with k random iterations was represented as MCCV (k).

Performance Metrics for Classification. Classification performance metrics are error measures that help distinguish correctly and incorrectly classified labels. These measures were derived using the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) determined from the confusion matrix. In the context of a COVID-19 classifier, the TP value describes the number of COVID-19 +ve patients correctly identified as +ve by the trained classifier. The FP describes the number of COVID-19 −ve patients incorrectly classified as +ve. The TN explains the number of COVID-19 −ve patients correctly classified as −ve. Similarly, the FN expounds on the number of COVID-19 +ve patients incorrectly classified as −ve. A schematic representation of a typical confusion matrix for the COVID-19 classifier is shown in Figure 5c, and the performance metrics discussed in the study are explained in Table 5.

Furthermore, an ROC curve was also used. ROC is an established method to evaluate a classification model's performance. An ROC curve is a two-dimensional graph where the FPR (1-specificity) is plotted on the x-axis and the TPR (sensitivity) on the y-axis for classification models constructed with different thresholds, τ.[49] Thus, ROC combines the two critical measures, sensitivity and 1-specificity. Sensitivity (TPR) describes the ratio of COVID-19 +ve patients correctly classified by the classifier as COVID-19 +ve out of the total COVID-19 +ve patients. Conversely, 1-specificity (FPR) shows the ratio of COVID-19 −ve individuals wrongly classified as COVID-19 +ve by the classifier out of the total COVID-19 −ve individuals. Furthermore, the quality of the classification model could be assessed using the AUROC.[18] For the best classification model, the value of the AUROC reaches 1; for the worst classification model, the value will be 0.5.

Performance Metrics for Prediction. Performance metrics for prediction evaluate the prediction error measures, which help conclude whether the predicted value of the class label is wrong or very wrong. For example, an incorrectly predicted class label value of 0.7 (wrong) and 1.2 (very wrong) for an actual class label of 0 will be penalized differently. The established prediction error metrics were the prediction error for cross-validation (Q^2), when the metric is determined in the space of the testing set. When the metric is calculated from the space of calibration (training set), it is termed R^2. Thus, Q^2 demonstrates how well the classification model predicts the class labels for a new set of spectra, and R^2 measures the goodness of fit.[24] Furthermore, one could estimate the performance metric Q^2 and R^2 as follows

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2}$$

where y_i is the original class label for a spectrum i that was not used in the training process, $\hat{y}_i$ the predicted value using the spectra i, and $\overline{y}_i$ the mean of y_i.

$$R^2 = 1 - \frac{\sum_k (y_k - \hat{y}_k)^2}{\sum_k (y_k - \overline{y}_k)^2}$$

where y_k is the original class label for a spectrum k used in the training process, $\hat{y}_k$ the predicted value using the spectra k, and $\overline{y}_k$ the mean of y_k. A good classifier predicts class labels close to the reference class labels. A Q^2 value close to 1 refers to a good classification model, whereas a value close to 0 describes the worst model. A Q^2 value equal to 1 is hard to obtain because this happens only when all predicted class labels are equal to the respective reference class labels. However, variation among the spectra of the same class of samples limits the classifier from predicting the exact class labels.[18]

Best PLS-DA Models. The best PLS-DA classification model could be determined by optimizing the number of latent variables (a) and should be free from overfitting; that is, the classifier should perform well with both known (calibration) and unknown (cross-validation) sets of data. Hence, for the best classifier, the error between R^2 and Q^2 will be less ($R^2 \approx Q^2$). Conversely, the error will be high ($R^2 \gg Q^2$) for an overfitted model.[24] Also, there is a direct correlation between overfitting and the number of latent variables used to construct a model.[16] Therefore, in this study, optimal a is determined such that the model constructed using a particular a should have a minimum error between R^2 and Q^2, and the model is considered the best model. Furthermore, the best models were validated using a third set of data (validation set) which is not used for calibrating the models, and the prediction error for validation (V^2) is estimated as follows

$$V^2 = 1 - \frac{\sum_l (y_l - \hat{y}_l)^2}{\sum_l (y_l - \overline{y}_l)^2}$$

where y_l is the original class label for a spectrum l in the validation set, $\hat{y}_l$ the predicted value using the spectra l from the validation set, and $\overline{y}_l$ the mean of y_l.

Permutation Test. A permutation test is a data-driven approach in PLS-DA classification rather than a theoretical approach (like a t-test or F-test). It is introduced to check the

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

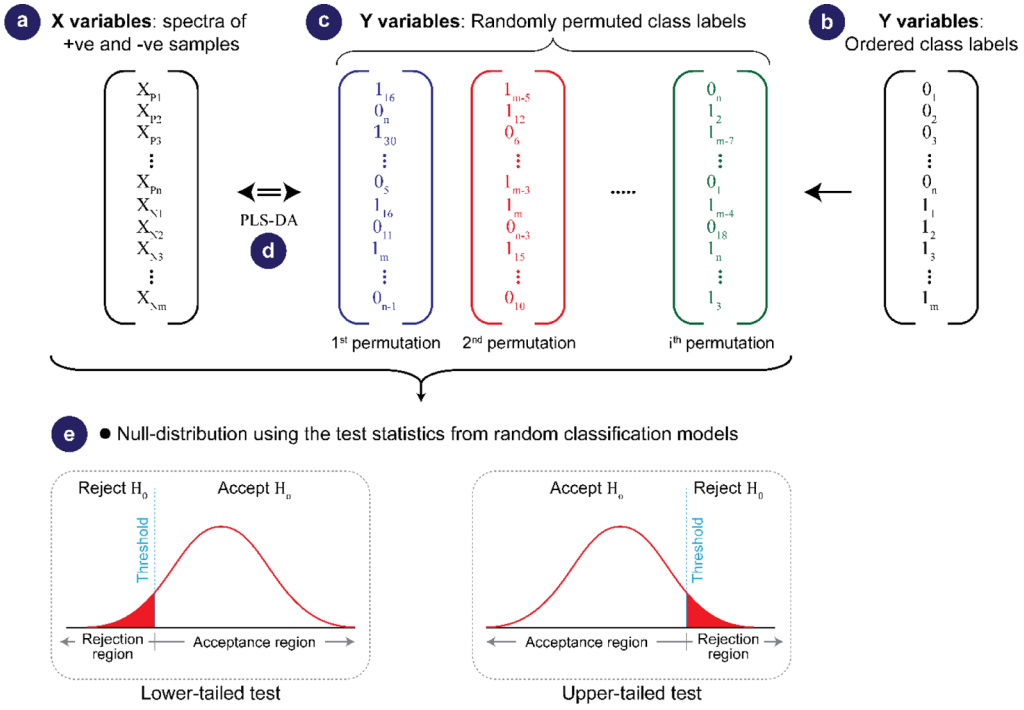


Figure 6. Rationale behind the permutation test. (a,b) represent the original X, and the Y variables used to construct the actual COVID-19 classification model. (c) represents the i random Y variables generated by permuting the original Y variables. (d) i random classifiers were constructed using the original X and permuted Y variables. (e) The null distribution for the test statistics will be constructed to evaluate H_0. A lower-tailed test is used to assess the robustness of the evidence against H_0 for the test statistic misclassification, and an upper-tailed test is used to assess the robustness against Q^2 and AUROC. The blue line in the lower- and upper-tailed test shows the null distribution's 95% confidence boundary (threshold), and H_0 will be rejected if the test statistic from the actual classifier falls outside the threshold of the null distribution.

possibility that the performance metrics of the actual classification models are not due to chance and thus estimates the statistical significance for the test statistics, the p-value.[22,23] An actual classifier is the one that is trained using the original X and Y variables. Conversely, a random classification model is constructed using the original X variables and the permuted set of Y variables. The logic is that the random classifiers should fail to relate the association between the X variables and the permuted Y variables, thus failing to predict the class labels correctly since the respective test statistic will be low compared to the test statistic of the actual classifier.[30] In this study, during the permutation test, 2000 random classifiers were constructed using random permutations (without replacement), and the test statistic was computed for each of them. The set of 2000 newly calculated test statistics associated with the random classifiers is the distribution under the null hypothesis (H_0) that there are no significant differences between the random and actual classifiers.[17] Furthermore, the test statistic from the actual classifier is compared using the null distribution, and H_0 is rejected if the test statistic from the actual classifier falls outside the 95% confidence boundary of the null distribution.[18] Also, a p-value could be estimated as the probability of observing the test statistic from the random classifiers which is at least as extreme as the test statistics for the actual classifier considering that H_0 is true. The smaller the p-value (close to 0), the more robust the evidence against H_0, leading to rejecting H_0. Hence, to deduce the p-value using the null distribution of the misclassification, a lower-tailed test is used. Similarly, for the test statistics Q^2 and AUROC, an upper-tail test will be considered. A schematic representation of the rationale behind the permutation test is shown in Figure 6.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c06786.

Performance plot of calibration vs MCCV, quality assessment of the best PLS-DA diagnosis models, visualization of real and preprocessed FTIR spectra, MCCV performance metrics for the best PLS-DA diagnosis models, and database of spectral band assignments (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Sivaraman Savithri − *Material Science and Technology Division, CSIR-National Institute for Interdisciplinary Science and Technology, Thiruvananthapuram 695019 Kerala, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India;* orcid.org/0000-0003-2220-0583; Email: ssavithri@niist.res.in

Authors

Sreejith Remanan Pushpa − *Material Science and Technology Division, CSIR-National Institute for Interdisciplinary Science and Technology, Thiruvananthapuram 695019 Kerala, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India;* orcid.org/0000-0002-5389-2394

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371

Rajeev Kumar Sukumaran − *Microbial Processes and Technology Division, CSIR-National Institute for Interdisciplinary Science and Technology, Thiruvananthapuram 695019 Kerala, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India;* orcid.org/0000-0002-1315-3650

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c06786

Author Contributions

S.R.P.: conceptualization, methodology, formal analysis, investigation, software, visualization, validation, project administration, writing—original draft, writing—review and editing. **R.K.S.:** supervision, validation, writing—original draft. **S.:** methodology, supervision, validation, funding acquisition, writing—original draft, writing—review and editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.R.P. acknowledges and thanks the Council of Scientific and Industrial Research (CSIR), New Delhi, India, and the University Grants Commission (UGC), New Delhi, India, for financial assistance through the CSIR-UGC Junior Research Fellowship (2019−2021) and CSIR-UGC Senior Research Fellowship (2021−2023). Also, the authors thank CSIR-NIIST, Thiruvananthapuram, for providing the research facility. S.R.P., R.K.S., and S.S. acknowledge and thank Bayden R. Wood et al., Valério G. Barauna et al., and Arghya Banerjee et al. for making the datasets public and permitting the reuse of the datasets.

■ REFERENCES

(1) Kim, D.; Lee, J. Y.; Yang, J. S.; Kim, J. W.; Kim, V. N.; Chang, H. The Architecture of SARS-CoV-2 Transcriptome. *Cell* **2020**, *181*, 914−921.e10.

(2) Xia, H.; Cao, Z.; Xie, X.; Zhang, X.; Chen, J. Y. C.; Wang, H.; Menachery, V. D.; Rajsbaum, R.; Shi, P. Y. Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep.* **2020**, *33*, 108234.

(3) Ciotti, M.; Ciccozzi, M.; Terrinoni, A.; Jiang, W. C.; Wang, C. B.; Bernardini, S. The COVID-19 Pandemic. *Crit. Rev. Clin. Lab. Sci.* **2020**, *57*, 365−388.

(4) Barr, J. N.; Fearns, R.Genetic Instability of RNA Viruses. *Genetic Instability of RNA Viruses*; Elsevier Inc., 2016. DOI: 10.1016/B978-0-12-803309-8.00002-1.

(5) Sharma, A.; Balda, S.; Apreja, M.; Kataria, K.; Capalash, N.; Sharma, P. COVID-19 Diagnosis: Current and Future Techniques. *Int. J. Biol. Macromol.* **2021**, *193*, 1835−1844.

(6) Wang, R.; Hozumi, Y.; Yin, C.; Wei, G. W. Mutations on COVID-19 Diagnostic Targets. *Genomics* **2020**, *112*, 5204−5213.

(7) Xu, L.; Li, D.; Ramadan, S.; Li, Y.; Klein, N. Facile Biosensors for Rapid Detection of COVID-19. *Biosens. Bioelectron.* **2020**, *170*, 112673.

(8) Pokhrel, P.; Hu, C.; Mao, H. Detecting the Coronavirus (CoVID-19). *ACS Sens.* **2020**, *5*, 2283−2296.

(9) Barauna, V. G.; Singh, M. N.; Barbosa, L. L.; Marcarini, W. D.; Vassallo, P. F.; Mill, J. G.; Ribeiro-Rodrigues, R.; Campos, L. C. G.; Warnke, P. H.; Martin, F. L. Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Anal. Chem.* **2021**, *93*, 2950−2958.

(10) Wood, B. R.; Kochan, K.; Bedolla, D. E.; Salazar-Quiroz, N.; Grimley, S. L.; Perez-Guaita, D.; Baker, M. J.; Vongsvivut, J.; Tobin, M. J.; Bambery, K. R.; Christensen, D.; Pasricha, S.; Eden, A. K.; Mclean, A.; Roy, S.; Roberts, J. A.; Druce, J.; Williamson, D. A.; McAuley, J.; Catton, M.; Purcell, D. F. J.; Godfrey, D. I.; Heraud, P. Infrared Based Saliva Screening Test for COVID-19. *Angew. Chem., Int. Ed.* **2021**, *60*, 17102−17107.

(11) Banerjee, A.; Gokhale, A.; Bankar, R.; Palanivel, V.; Salkar, A.; Robinson, H.; Shastri, J. S.; Agrawal, S.; Hartel, G.; Hill, M. M.; Srivastava, S. Rapid Classification of COVID-19 Severity by ATR-FTIR Spectroscopy of Plasma Samples. *Anal. Chem.* **2021**, *93*, 10391−10396.

(12) Guleken, Z.; Jakubczyk, P.; Wiesław, P.; Krzysztof, P.; Bulut, H.; Öten, E.; Depciuch, J.; Tarhan, N. Characterization of Covid-19 Infected Pregnant Women Sera Using Laboratory Indexes, Vibrational Spectroscopy, and Machine Learning Classifications. *Talanta* **2022**, *237*, 122916.

(13) Nascimento, M. H. C.; Marcarini, W. D.; Folli, G. S.; da Silva Filho, W. G.; Barbosa, L. L.; de Paulo, E. H.; Vassallo, P. F.; Mill, J. G.; Barauna, V. G.; Martin, F. L.; de Castro, E. V. R.; Romão, W.; Filgueiras, P. R. Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. *Anal. Chem.* **2022**, *94*, 2425−2433.

(14) Zhang, L.; Xiao, M.; Wang, Y.; Peng, S.; Chen, Y.; Zhang, D.; Zhang, D.; Guo, Y.; Wang, X.; Luo, H.; Zhou, Q.; Xu, Y. Fast Screening and Primary Diagnosis of COVID-19 by ATR-FT-IR. *Anal. Chem.* **2021**, *93*, 2191−2199.

(15) Martinez-Cuazitl, A.; Vazquez-Zapien, G. J.; Sanchez-Brito, M.; Limon-Pacheco, J. H.; Guerrero-Ruiz, M.; Garibay-Gonzalez, F.; Delgado-Macuil, R. J.; de Jesus, M. G. G.; Pereyra-Talamantes, M. A.; Mata-Miranda, A.; Mata-Miranda, M. M. ATR-FTIR Spectrum Analysis of Saliva Samples from COVID-19 Positive Patients. *Sci. Rep.* **2021**, *11*, 1−14.

(16) Defernez, M.; Kemsley, E. K. The Use and Misuse of Chemometrics for Treating Classification Problems. *TrAC, Trends Anal. Chem.* **1997**, *16*, 216−221.

(17) Faber, N. M.; Rajkó, R. How to Avoid Over-Fitting in Multivariate Calibration-The Conventional Validation Approach and an Alternative. *Anal. Chim. Acta* **2007**, *595*, 98−106.

(18) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA Cross Validation. *Metabolomics* **2008**, *4*, 81−89.

(19) Rodríguez-Pérez, R.; Fernández, L.; Marco, S. Overoptimism in Cross-Validation When Using Partial Least Squares-Discriminant Analysis for Omics Data: A Systematic Study. *Anal. Bioanal. Chem.* **2018**, *410*, 5981−5992.

(20) Blazhko, U.; Shapaval, V.; Kovalev, V.; Kohler, A. Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra. *Chemom. Intell. Lab. Syst.* **2021**, *215*, 104367.

(21) Xu, Q. S.; Liang, Y. Z. Monte Carlo Cross Validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1−11.

(22) Szymańska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. Double-Check: Validation of Diagnostic Statistics for PLS-DA Models in Metabolomics Studies. *Metabolomics* **2012**, *8*, 3−16.

(23) Golland, P.; Fischl, B. Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies. *Lect. Notes Comput. Sci.* **2003**, *2732*, 330−341.

(24) Bevilacqua, M.; Bro, R. Can We Trust Score Plots? *Metabolites* **2020**, *10*, 278.

(25) Abdi, H. Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 97−106.

(26) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(27) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17*, 166−173.

(28) Westerhuis, J. A.; van Velzen, E. J. J.; Hoefsloot, H. C. J.; Smilde, A. K. Discriminant Q2 (DQ2) for Improved Discrimination in PLSDA Models. *Metabolomics* **2008**, *4*, 293−296.

(29) Hobro, A. J.; Kuligowski, J.; Döll, M.; Lendl, B. Differentiation of Walnut Wood Species and Steam Treatment Using ATR-FTIR and Partial Least Squares Discriminant Analysis (PLS-DA). *Anal. Bioanal. Chem.* **2010**, *398*, 2713−2722.
(30) Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. Model validation by permutation tests: Applications to variable selection. *J. Chemom.* **1996**, *10*, 521−532.
(31) Kuligowski, J.; Quintás, G.; Herwig, C.; Lendl, B. A Rapid Method for the Differentiation of Yeast Cells Grown under Carbon and Nitrogen-Limited Conditions by Means of Partial Least Squares Discriminant Analysis Employing Infrared Micro-Spectroscopic Data of Entire Yeast Cells. *Talanta* **2012**, *99*, 566−573.
(32) Banyay, M.; Sarkar, M.; Gräslund, A. A Library of IR Bands of Nucleic Acids in Solution. *Biophys. Chem.* **2003**, *104*, 477−488.
(33) Goormaghtigh, E.; Cabiaux, V.; Ruysschaert, J. M. Determination of Soluble and Membrane Protein Structure by Fourier Transform Infrared Spectroscopy. *Subcell. Biochem.* **1994**, *23*, 363−403.
(34) Hansen, L.; De Beer, T.; Pierre, K.; Pastoret, S.; Bonnegarde-Bernard, A.; Daoussi, R.; Vervaet, C.; Remon, J. P. FTIR Spectroscopy for the Detection and Evaluation of Live Attenuated Viruses in Freeze Dried Vaccine Formulations. *Biotechnol. Prog.* **2015**, *31*, 1107−1118.
(35) Stukalov, A.; Girault, V.; Grass, V.; Karayel, O.; Bergant, V.; Urban, C.; Haas, D. A.; Huang, Y.; Oubraham, L.; Wang, A.; Hamad, M. S.; Piras, A.; Hansen, F. M.; Tanzer, M. C.; Paron, I.; Zinzula, L.; Engleitner, T.; Reinecke, M.; Lavacca, T. M.; Ehmann, R.; Wölfel, R.; Jores, J.; Kuster, B.; Protzer, U.; Rad, R.; Ziebuhr, J.; Thiel, V.; Scaturro, P.; Mann, M.; Pichlmair, A.*Multilevel Proteomics Reveals Host Perturbations by SARS-CoV-2 and SARS-CoV*; Springer US, 2021; Vol. *594*.
(36) Abassi Joozdani, F. A.; Yari, F.; Abassi Joozdani, P. A.; Nafisi, S. Interaction of Sulforaphane with DNA and RNA. *PLoS One* **2015**, *10*, e0127541−14.
(37) Kitane, D. L.; Loukman, S.; Marchoudi, N.; Fernandez-Galiana, A.; El Ansari, F. Z.; Jouali, F.; Badir, J.; Gala, J. L.; Bertsimas, D.; Azami, N.; Lakbita, O.; Moudam, O.; Benhida, R.; Fekkak, J. A Simple and Fast Spectroscopy-Based Technique for Covid-19 Diagnosis. *Sci. Rep.* **2021**, *11*, 1−11.
(38) Movasaghi, Z.; Rehman, S.; ur Rehman, I. U. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **2008**, *43*, 134−179.
(39) Nogueira, M. S.; Leal, L. B.; Marcarini, W.; Pimentel, R. L.; Muller, M.; Vassallo, P. F.; Campos, L. C. G.; dos Santos, L.; Luiz, W. B.; Mill, J. G.; Barauna, V. G.; de Carvalho, L. F. d. C. e. S.; das, L. F. C. e. S. Rapid Diagnosis of COVID-19 Using FT-IR ATR Spectroscopy and Machine Learning. *Sci. Rep.* **2021**, *11*, 1−13.
(40) Santos, M. C. D.; Morais, C. L. M.; Lima, K. M. G. ATR-FTIR Spectroscopy for Virus Identification: A Powerful Alternative. *Biomed. Spectrosc. Imag.* **2021**, *9*, 103−118.
(41) Serec, K.; Šegedin, N.; Krajačić, M.; Dolanski Babić, S. D. Conformational Transitions of Double-Stranded Dna in Thin Films. *Appl. Sci.* **2021**, *11*, 2360.
(42) Tatulian, S. A.Structural Characterization of Membrane Proteins and Peptides by FTIR and ATR-FTIR Spectroscopy. *Lipid-protein interactions*; Springer, 2013; Vol. *974*.
(43) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627−1639.
(44) Zimmermann, B.; Kohler, A. Optimizing Savitzky-Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 892−902.
(45) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(46) Kohler, A.; Kirschner, C.; Oust, A.; Martens, H. Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-Sections of Beef Loin. *Appl. Spectrosc.* **2005**, *59*, 707−716.
(47) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(48) Triba, M. N.; Le Moyec, L.; Amathieu, R.; Goossens, C.; Bouchemal, N.; Nahon, P.; Rutledge, D. N.; Savarin, P. PLS/OPLS Models in Metabolomics: The Impact of Permutation of Dataset Rows on the K-Fold Cross-Validation Quality Parameters. *Mol. Biosyst.* **2015**, *11*, 13−19.
(49) Fawcett, T.*ROC Graphs: Notes and Practical Considerations for Researchers*; Kluwer Academic Publishers, 2004. No. March.

https://doi.org/10.1021/acsomega.2c06786
ACS Omega 2022, 7, 47357−47371