Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 146 (2022) 105659

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning

Alexandre de Fátima Cobre [a], Monica Surek [a], Dile Pontarolo Stremel [b], Mariana Millan Fachi [a], Helena Hiemisch Lobo Borba [c], Fernanda Stumpf Tonin [a,d], Roberto Pontarolo [c,*]

[a] Pharmaceutical Sciences Postgraduate Program, Universidade Federal Do Paraná, Curitiba, Brazil
[b] Department of Forest Engineering and Technology, Universidade Federal Do Paraná, Curitiba, Brazil
[c] Department of Pharmacy, Universidade Federal Do Paraná, Curitiba, Brazil
[d] H&TRC- Health & Technology Research Center, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Lisbon, Portugal

ARTICLE INFO

Keywords:
COVID 19
Fatality
Severity
Diagnosis
Biomarker
Machine learning

ABSTRACT

Objective: To implement and evaluate machine learning (ML) algorithms for the prediction of COVID-19 diagnosis, severity, and fatality and to assess biomarkers potentially associated with these outcomes.
Material and methods: Serum (n = 96) and plasma (n = 96) samples from patients with COVID-19 (acute, severe and fatal illness) from two independent hospitals in China were analyzed by LC-MS. Samples from healthy volunteers and from patients with pneumonia caused by other viruses (i.e. negative RT-PCR for COVID-19) were used as controls. Seven different ML-based models were built: PLS-DA, ANNDA, XGBoostDA, SIMCA, SVM, LREG and KNN.
Results: The PLS-DA model presented the best performance for both datasets, with accuracy rates to predict the diagnosis, severity and fatality of COVID-19 of 93%, 94% and 97%, respectively. Low levels of the metabolites ribothymidine, 4-hydroxyphenylacetoylcarnitine and uridine were associated with COVID-19 positivity, whereas high levels of N-acetyl-glucosamine-1-phosphate, cysteinylglycine, methyl isobutyrate, L-ornithine and 5,6-dihydro-5-methyluracil were significantly related to greater severity and fatality from COVID-19.
Conclusion: The PLS-DA model can help to predict SARS-CoV-2 diagnosis, severity and fatality in daily practice. Some biomarkers typically increased in COVID-19 patients' serum or plasma (i.e. ribothymidine, N-acetyl-glucosamine-1-phosphate, L-ornithine, 5,6-dihydro-5-methyluracil) should be further evaluated as prognostic indicators of the disease.

1. Introduction

The COVID-19 outbreak has been met by variable responses across countries, especially regarding the adoption of prevention measures. Yet, although science has uncovered much about SARS-CoV-2 and made unprecedented progress in the development of vaccines, there is still great uncertainty as the pandemic continues to evolve (i.e. new active cases and deaths reported worldwide), and no globally recognized effective treatment is available [1,2]. In this scenario, the implementation of further sensitive, accurate, low-cost screening and early diagnostic approaches is paramount for preventing infections and guiding disease stage monitoring [3,4].

In recent years, artificial intelligence (AI), including deep learning (DL) and machine learning-based (ML) algorithms, has emerged as a useful tool to support the decision-making process in healthcare, as well as drug discovery and disease diagnosis and monitoring [5–7]. Some studies published in the past two years have already employed these methods to guide COVID-19 diagnosis (i.e. using chest computed tomography scans and X-ray images), to characterize biomarkers of disease stage, to identify risk factors of disease severity and mortality and to forecast future outbreaks [8–12]. A recent systematic review conducted by Wang (2021) highlights that AI has the potential to improve existing medical and healthcare system efficiency during the COVID-19 pandemic by additionally assisting with surveillance and public health decision-making [27].

* Corresponding author. Department of Pharmacy, Universidade Federal do Paraná, Campus III, Av. Pref. Lothário Meissner, 632, Jardim Botânico, Curitiba, PR, 80.210-170, Brazil.
E-mail addresses: alexandrecobre@gmail.com (A. de Fátima Cobre), monicasurek13@gmail.com (M. Surek), dile.stremel@gmail.com (D.P. Stremel), marianamfachi@gmail.com (M.M. Fachi), helena.hlb@gmail.com (H.H. Lobo Borba), stumpf.tonin@ufpr.br (F.S. Tonin), pontarolo@ufpr.br (R. Pontarolo).

https://doi.org/10.1016/j.compbiomed.2022.105659
Received 20 February 2022; Received in revised form 11 May 2022; Accepted 18 May 2022
Available online 21 May 2022

Abbreviations

LC-MS	liquid chromatography coupled with mass spectrometry
PLS-DA	discriminant analysis by partial least squares
ANNDA	artificial neural networks discriminant analysis
XGBoostDA	gradient boosted tree discriminant analysis
SIMCA	soft independent modelling of class analogy
SVM	support vector machine
KNN	k-nearest neighbours
LREG	logistic regression discriminant analysis

However, although several of these ML-based models for the prediction of COVID-19 diagnosis and severity are available in the literature, limitations to their use in practice still exist. Biomarkers identified by these algorithms as potentially associated with the disease vary widely due to, among others, high heterogeneity among patients' clinical profiles (i.e. different populations of regions/countries) and small sample sizes (i.e. retrospective single-centre studies), which reduces external validity and data generalization [13,14]. Additionally, differences in sample preparation and analytical techniques, as well the increasing number of SARS-CoV-2 variants worldwide, hinder the integration of more streamlined and effective predictive modelling in this field [15–17].

Considering the increasing cases of COVID-19 in the past months, mainly caused by the rapid spread of the omicron variant [18], alongside with the limited number of kits to perform real-time PCR tests - as a consequence of the growing demand for these products worldwide, the aim of this study is to evaluate different ML-based algorithms for the prediction of COVID-19 diagnosis, severity and fatality, as well as to identify new biomarkers associated with these outcomes, using two databases with over 1300 water-soluble and fat-soluble metabolites.

2. Material and methods

2.1. Study design and databases

We evaluated public datasets from two cohorts of patients diagnosed with COVID-19 by RT-PCR in China (Wuhan), including mild and severe cases and deaths associated with the disease (https://drive.google.com/drive/folders/1R_I_gu5D3SkD_9q_J93HOA9GuKxZiGNG) [19]. Blood samples from all diagnosed patients were assessed by ultra-efficiency liquid chromatography coupled to mass spectrometry (LC-MS).

The first cohort (Dataset I) refers to samples from COVID-19 patients diagnosed in the Wuhan Jinyin Tan Hospital (China) (file name: C2_metaboanalyst_input_full.csv) [19]. A series of samples was registered during the disease course: samples collected from 14 patients with mild symptoms at two time points in the study (total of $14 \times 2 = 28$ samples), samples from 11 patients with severe symptoms (total of $11 \times 2 = 22$ samples) and samples collected at four time points from 9 patients that died during the study (total of $9 \times 4 = 36$ samples). Blood samples from 10 healthy volunteers with negative RT-PCR tests were used as controls. Plasma samples from all these patients were also collected and analyzed by ultra-efficiency liquid chromatography (C18 column) coupled to quadrupole mass spectrometry and electrospray source (UPLC-ESI-MS/MS). A total of 431 metabolites (both fat-soluble and water-soluble substances) were identified and quantified using an in-house hospital database and molecular ion fragmentation profile in MS/MS mode. Fragments were compared to data from the international public literature/database. This first database, thus, included 96 samples and 431 variables (metabolites).

The second cohort (Dataset II) refers to a sample of 46 patients diagnosed with COVID-19 in the Taizhou Hospital (China) (file name: C3_metaboanalyst_input_full.csv) [19]. Blood samples from 25 healthy volunteers (negative RT-PCR) and from 25 patients with pneumonia syndrome but with negative RT-PCR for SARS-CoV-2 were respectively used as negative and positive control groups. Serum samples of all patients were analyzed by UPLC-ESI-MS/MS. This second database accounted for 96 samples and 941 metabolites (identified and quantified using both ionization modes; ESI- and ESI+).

2.2. Data preprocessing in ML

Data preprocessing is an important step for metabolomic data analysis and refers to the technique of preparing (i.e. cleaning and organizing) the raw data to make it suitable (i.e. readable) for building and training ML-based models [20,21].

In this study, both COVID-19 datasets (i.e. diagnosis and disease severity) went through different preprocessing methods aiming at selecting the one that best fit the data: (i) Imputation: missing data were replaced by the median values; (ii) Transformations: absolute value, Log10; (iii) Filtering: baseline (specified points), baseline (weighted least square), derivative (Savitzky – Golay), smoothing (Savitzky – Golay), detrend, generalized least squares weighting (GLSW), orthogonal signal correction (OSC) and external parameter orthogonalization (EPO); (iv) Normalization: normalize, standard normal variate (SNV) and multiplicative scatter correction (MSC-mean); (v) Scaling and centering: autoscale, group scale, Log decay scaling, mean center, median center, multiway center, multiway scale and sqrt mean scale. All analyzes were performed in SOLO software (Eigenvector Research).

2.3. Development of ML-based and prediction models

In this study, an unsupervised ML-based model (principal component analysis - PCA) was initially developed with both datasets aiming at identifying the structure of the data and detecting possible anomalous samples [22,23] (i.e. exploratory analyses).

For the prediction of COVID-19 diagnosis (Database II) and disease severity and fatality (Database I), several supervised ML-based models were used: support vector machine (SVM), discriminant analysis (ANNDA), k-nearest neighbours (KNN), artificial neural networks discriminant analysis (ANNDA), discriminant analysis by partial least squares (PLS-DA), soft independent modelling of class analogy (SIMCA), gradient boosted tree discriminant analysis (XGBoostDA) and logistic regression discriminant analysis (LREG). For the implementation of these classification models, 70% of the data was used for the training set (calibration) and the remaining 30% for the test set.

Sample selection for the training and testing sets was randomly performed using the Kennad Stone algorithm [24]. For the implementation of the models, the classes (groups) of samples from the datasets were individually divided into two new subsets (i.e., training and test samples), being this last subset (test samples) used to predict each specific class. Samples from Database I (Wuhan, Jinyin Tan Hospital, n = 96 samples) were grouped into the following: class 1 – healthy (accounting for 10 samples of which 5 were used for training the model and the remaining 5 for data prediction); class 2 – death (36 samples of which 25 were used for training; 11 for data prediction); class 3 – severe COVID-19 (22 samples of which 15 were used for training; 7 for data prediction); and class 4 – mild COVID-19 (28 samples; 20 used for training and 8 for data prediction). The samples from Database II (Taizhou Hospital, n = 96) were categorized into the following classes: class 1 – COVID-19 (accounting for 46 samples of which 32 were used for training the models and the remaining 14 for data prediction); class 2 – healthy (25 samples of which 15 were used for training and 10 for data prediction); and class 3 – non-COVID-19 (25 samples; 15 used for training and 10 for data prediction).

The venetian blind cross-validation method was used to select the number of latent variables (LVs) of the ML-based models [25]. LVs with lower values of cross-validation error, square root of mean

cross-validation error (RMSECV) and root mean square error of calibration (RMSEC) were selected. The root of the mean error of prediction (RMSEP) was the metric used to assess the predictive capacity of the ML-based models; models with smaller RMSEP performed better.

Model performance was evaluated considering the metrics of accuracy, sensitivity, and specificity. These metrics were calculated using the following figures of merit: false positive (FP), false negative (FN), true positive (TP) and true negative (TN), according to equations (1)–(3).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

where, FP = false positive; FN = false negative; TP = true positive; TN = true negative.

The accuracy of the models was also evaluated considering the area under the receiver operating characteristic (ROC) curve (AUC). The values of AUC ROC were calculated considering both samples' datasets (i.e., training and test sets).

A VIP (variable importance in projection) graph was built from the ML-based model presenting the best performance aiming at identifying the 'top 10' most important biomarkers for predicting the diagnosis of COVID-19 and the 'top 10' most important biomarkers for predicting disease severity and fatality. A VIP score of an original variable is calculated as a weighted sum of the squared correlations between the LV of the PLS-DA model and the original variable (e.g. metabolite). The number of terms in the sum relies on the number of LV from the PLS-DA model that were considered significant for distinguishing the groups (classes) of samples. The weights correspond to the percentage variance explained by the LV in the PLS-DA model. An original variable with a VIP score greater than 1 is considered statistically significant for classifying groups (e.g. COVID-19 group vs. healthy individuals). See below the VIP score calculation equation (4) [26,27].

$$VIP_j^2 = \sum_f w_{if}\, SSY_f\, J/(SSY_{\text{total expl.}}\, F) \quad (4)$$

where: wj = PLS-weight value; SSY = percentage of explained Y variance by each specific latent variable; F = number of latent variables of the PLS-DA model; J = number of X variables.

Analyses were carried out using SOLO (Eigenvector Research) software [28] and Metaboanalyst 5.0 web server [29]; results obtained with these different tools were qualitatively compared.

3. Results

3.1. Exploratory analyses

Fig. 1 shows the PCA model for both datasets (Dataset I and II). The preprocessing methods that best suited the model were a combination of imputation using median values, autoscale and GLSW. In both cases, the PCA model was able to discriminate all classes of samples. No outlier sample was identified.

3.2. Classification models

Table 1 shows the performance of the ML-based models built using SOLO software. The PLS-DA model showed the most promising results (high performance) for predicting the diagnosis, severity, and fatality of COVID-19 with higher figures of accuracy, sensitivity, and specificity of 93%, 94% and 97%, respectively (see ROC curve using samples from both training and test sets in Fig. 2. The ROC curves using only the training set samples are available in supplementary material - Fig. S1). Supplementary Material Figs. S2–S7 show the PLS-DA models for predicting each class of samples. The remaining ML-based models (ANN, ANNDA, XGBoostDA, SIMCA, SVM, LREG and KNN) showed poor performance in this study.

Figs. 3 and 4 (VIP graphs of the PLS-DA models) depict the most promising biomarkers for predicting the diagnosis and severity/fatality of COVID-19, respectively. The calculation of the VIP scores of these

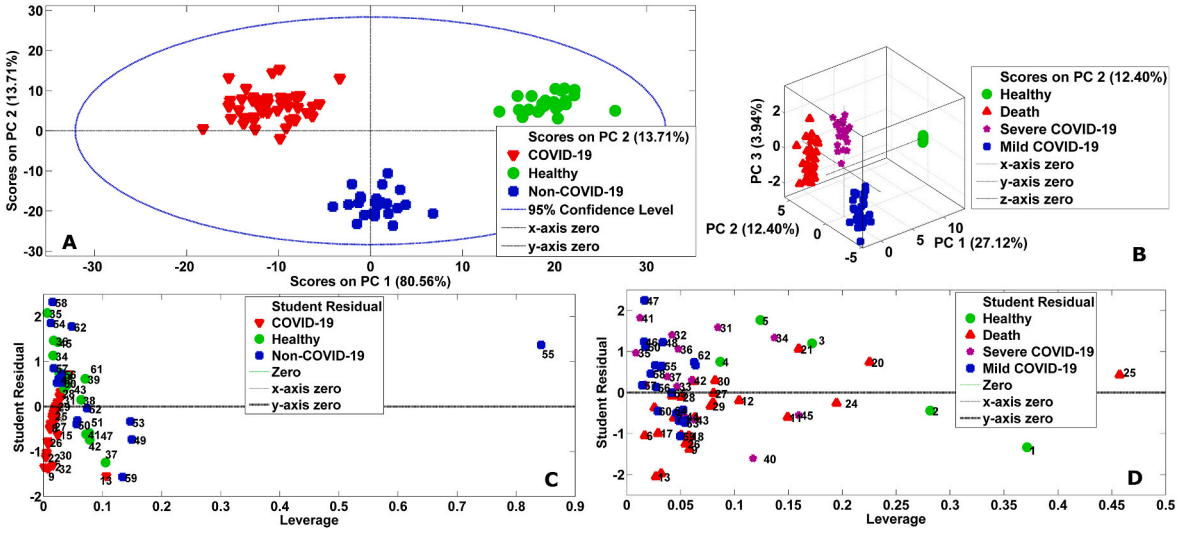

Fig. 1. Exploratory data analysis. (A) PCA model of the Thaizhou hospital patient dataset (blood samples from 46 patients with COVID-19 diagnosed by RT-PCR are represented by the red triangles; blood samples from 25 patients with pneumonia syndrome but with negative RT-PCR for SARS-CoV-2 are depicted as blue squares; blood samples from 25 healthy volunteers with negative RT-PCR are represented by green circles). (B) PCA model of the Wuhan hospital patient dataset (blood samples from 28 patients with mild COVID-19 are represented by the blue color squares; blood samples from 36 patients with COVID-19 are represented by pink stars; blood samples from 36 deaths by COVID-19 are depicted as red triangles; blood samples from 10 healthy volunteers negative by RT-PCR are depicted as green circles). (C) Graph of leverage versus student residuals for the detection of outlier samples in the Thaizhou hospital patient dataset (sample n. 55 had high leverage values but was not considered an outlier as it is within ±2.5 standard deviations of student residuals). (D) Graph of leverage versus student residuals for the detection of outlier samples in the Wuhan hospital patient dataset (sample n. 25 had high leverage values but was not considered an outlier as it is within ±2.5 standard deviations of student residuals).

Table 1
Performance of the seven ML-based models.

Wuhan Hospital dataset (Database I)								
Class	Model	TP	FN	TN	FP	Sensitivity	Specificity	Accuracy
Mild COVID-19	ANN	15	13	49	19	0.54	0.72	0.67
	KNN	20	8	53	15	0.71	0.78	0.76
	SVM	23	5	45	23	0.82	0.66	0.71
	PLS-DA	28	0	63	5	**1.00**	**0.93**	**0.95**
	SIMCA	17	11	39	29	0.61	0.57	0.58
	XGboost	25	3	60	8	0.89	0.88	0.89
	LREG	18	10	56	12	0.64	0.82	0.77
Severe COVID-19	ANN	16	6	65	9	0.73	0.88	0.84
	KNN	15	7	57	17	0.68	0.77	0.75
	SVM	18	4	60	14	0.82	0.81	0.81
	PLS-DA	19	3	71	3	**0.86**	**0.96**	**0.94**
	SIMCA	15	7	41	33	0.68	0.55	0.58
	XGboost	17	5	59	15	0.77	0.80	0.79
	LREG	16	6	66	8	0.73	0.89	0.85
Deaths by COVID-19	ANN	29	7	50	10	0.81	0.83	0.82
	KNN	24	12	53	7	0.67	0.88	0.80
	SVM	31	5	55	5	0.86	0.92	0.90
	PLS-DA	35	1	58	2	**0.97**	**0.97**	**0.97**
	SIMCA	28	8	43	17	0.78	0.72	0.74
	XGboost	33	3	51	9	0.92	0.85	0.88
	LREG	25	9	46	14	0.69	0.77	0.74
Healthy (control)	ANN	10	0	74	12	1.00	0.86	0.88
	KNN	7	3	67	19	0.70	0.78	0.77
	SVM	9	1	70	16	0.90	0.81	0.82
	PLS-DA	10	0	82	4	**1.00**	**0.95**	**0.96**
	SIMCA	10	0	73	13	1.00	0.85	0.86
	XGboost	10	0	76	10	1.00	0.88	0.90
	LREG	10	0	61	19	1.00	0.71	0.74
Taizhou Hospital dataset (Database II)								
Class	**Model**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**Accuracy**
COVID-19 positive	ANN	40	6	39	11	0.87	0.78	0.82
	KNN	33	13	40	10	0.72	0.80	0.76
	SVM	41	5	37	13	0.89	0.74	0.81
	PLS-DA	45	1	44	6	**0.98**	**0.88**	**0.93**
	SIMCA	34	12	42	8	0.74	0.84	0.79
	XGboost	39	7	41	9	0.85	0.82	0.83
	LREG	41	5	33	17	0.89	0.66	0.77
Non-COVID-19	ANN	20	5	59	12	0.80	0.83	0.82
	KNN	14	11	55	16	0.56	0.77	0.72
	SVM	22	3	44	17	0.88	0.62	0.69
	PLS-DA	23	2	64	7	**0.92**	**0.90**	**0.91**
	SIMCA	19	6	42	19	0.76	0.59	0.64
	XGboost	15	10	62	9	0.60	0.87	0.80
	LREG	17	8	49	12	0.68	0.69	0.69
Healthy (control)	ANN	16	9	56	15	0.64	0.79	0.75
	KNN	18	7	63	8	0.72	0.89	0.84
	SVM	16	9	51	20	0.64	0.72	0.70
	PLS-DA	23	2	67	4	**0.92**	**0.94**	**0.94**
	SIMCA	19	6	60	11	0.76	0.85	0.82
	XGboost	21	4	58	13	0.84	0.82	0.82
	LREG	17	8	55	16	0.68	0.77	0.75

Note: TP = true positive; TN = true negative; FP = false positive; FN = false negative.

biomarkers (see equation (4) - material and methods section) included two parameters: (i) four LVs selected for the PLS-DA model as they presented lower calibration error (RMSEC) and cross-validation (RMSECV) values (see Figs. S8–S9 in supplementary material), and (ii) and the total variance explained by these four selected LVs, which was 42.84% for block X and 83.53%% for block Y.

The VIP graph from the PLS-DA model revealed that the most important biomarkers for predicting COVID-19 diagnosis were dibutyl sulfosuccinate, ortho-cresol sulphate, beta alanine, 4-vinylguaiacol sulphate, 4-hydroxyphenylacetoylcarnitine, ribothymidine, glycerophosphoserine and uridine (see Fig. 5). These three last biomarkers were found in extremely low concentrations (decreased by factors of two, three and four) in patients diagnosed with COVID-19 when compared to negative control samples.

As for the prediction of COVID-19 severity and fatality, six different biomarkers were highlighted in the model as most probably associated with these outcomes: cyclohexylamine, methyl isobutyrate, 2-Undecanone, cysteinylglycine, N-acetyl-glucosamine-1-phosphate and 5,6-dihydro-5-methyluracil were increased by factors of three, four, five, six, seven and seven, respectively, in patients with severe COVID-19 when compared to those with acute disease (see Fig. 6).

The above mentioned analyzes were also carried out (i.e. re-run) using Metaboanalyst 5.0; results are available in supplementary material (Figs. S10–S13). In this case, the PCA model was not able to distinguish the samples from the three classes (COVID-19, non-COVID-19, healthy individuals) using the diagnostic data; the accuracy was inferior to 80% (i.e. lower than the one obtained in our study [93–94%] using SOLO software). Similarly, although the PCA model of the severity data was able to distinguish healthy patients from deaths, it was not able to classify the other two groups (mild COVID-19 vs. severe COVID-19). The accuracy of this model was of around 80%, lower that the one obtained using SOLO (94%–97%).

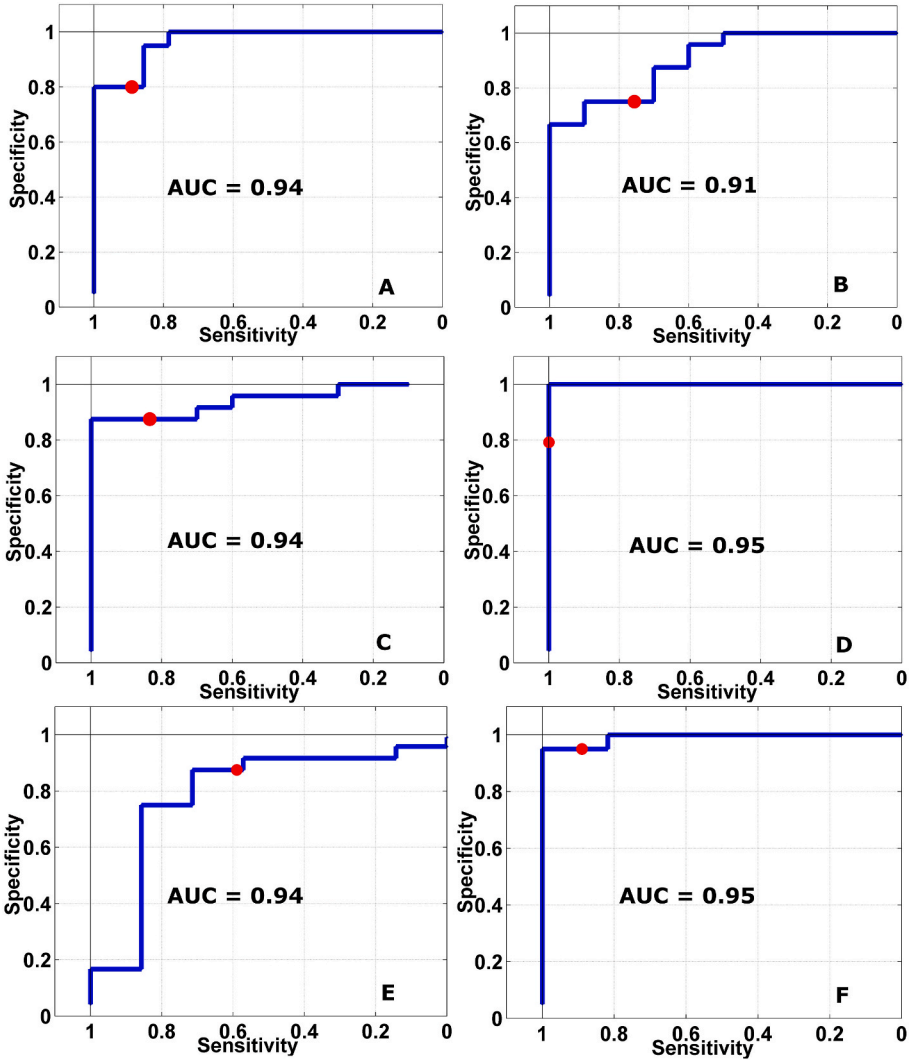

Fig. 2. Area under the ROC curve of PLS-DA model performance. Area under the curves (AUC) reflect the accuracy of PLS-DA models in predicting patients of different COVID-19 classes and healthy volunteers. Curves include both sets of samples (training and test samples). (A) Thaizhou hospital dataset (dataset II): results represent the accuracy in predicting the class of patients with COVID-19 diagnosed by RT-PCR (AUC = 0.93). (B) Thaizhou hospital dataset (dataset II): results represent the accuracy in predicting the class of patients with pneumonia syndrome but with negative RT-PCR for SARS-CoV-2 (AUC = 0.91). (C) Thaizhou hospital dataset (dataset II): results represent the accuracy in predicting the class of healthy volunteers with negative RT-PCR (AUC = 0.94). (D) Wuhan hospital dataset (dataset I): results represent the accuracy in predicting the class of patients with acute COVID-19 (AUC = 0.95). (E) Wuhan hospital dataset (dataset I): results represent the accuracy in predicting patients with severe COVID-19 (AUC = 0.94). (F) Wuhan hospital dataset (dataset I): results represent the accuracy in the classification of deaths by COVID-19 (AUC = 0.97).

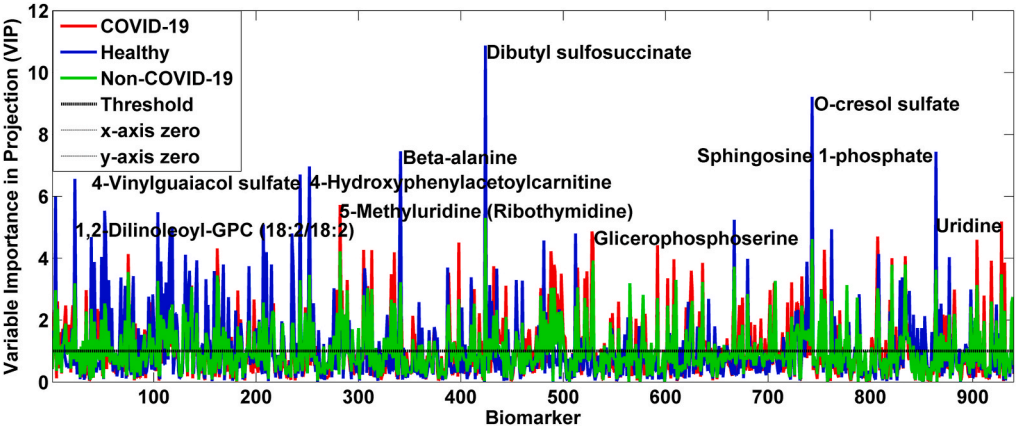

Fig. 3. Variable Importance in Projection graph of the most important biomarkers for COVID-19 diagnosis (top 10). X axis represents all analyzed metabolites; Y axis represents the VIP score that reflects the importance of each metabolite in the prediction of the different classes of the samples (COVID-19 represented by the red color, non-COVID-19 by blue, and healthy volunteers in green). The black dashed line parallel to the X axis represents the VIP score threshold (VIP score threshold = 1). Metabolites significantly contributing to the prediction of the different classes of the samples are above the threshold (VIP score> 1); the top 10 biomarkers were highlighted in the figure.

We also found differences in the identification of biomarkers from the PLS-DA models obtained using SOLO vs. Metaboanalyst 5.0 software (see Table 2). The analyzes performed in Metaboanalyst 5.0 showed the following metabolites in extremely low levels in patients with COVID-19 compared with those without the disease or healthy volunteers: linoleate, palmitate, urea, lactate, carnitine, proline, glycerophosphoethanolamine, stearate, phenylalanine (see Fig. 7). On the other hand, the metabolites 6-methylmercaptopurine, dihydroxybenzeneacetic acid, L-phenylalanine, 6-methylmercaptopurine, 4-dihydroxybenzeneacetic acid, L-phenylalanine, formylanthranilic acid, terephthalic acid and phthalic acid were found in high concentrations in patients who died from COVID-19 (see Fig. 8).

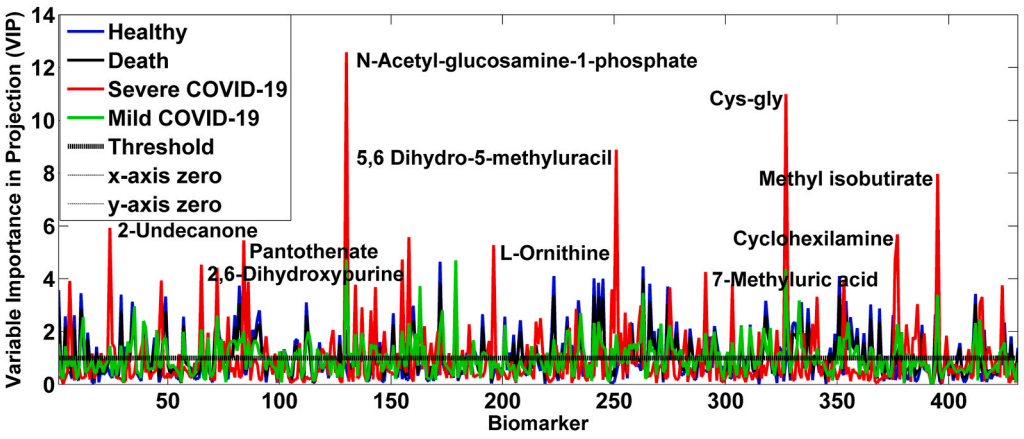

Fig. 4. Variable Importance in Projection graph of the most important biomarkers for COVID-19 severity and fatality. X axis represents all analyzed metabolites; Y axis represents the VIP score that reflects the importance of each metabolite in the prediction of the different classes of the samples (healthy individuals are depicted in blue, mild COVID-19 is green, severe COVID-19 is in red and death is colored in black). The black dashed line parallel to the X axis represents the VIP score threshold (VIP score threshold = 1). Metabolites significantly contributing to the prediction of the different classes of the samples are above the threshold (VIP score> 1); the top 10 biomarkers were highlighted in the figure.

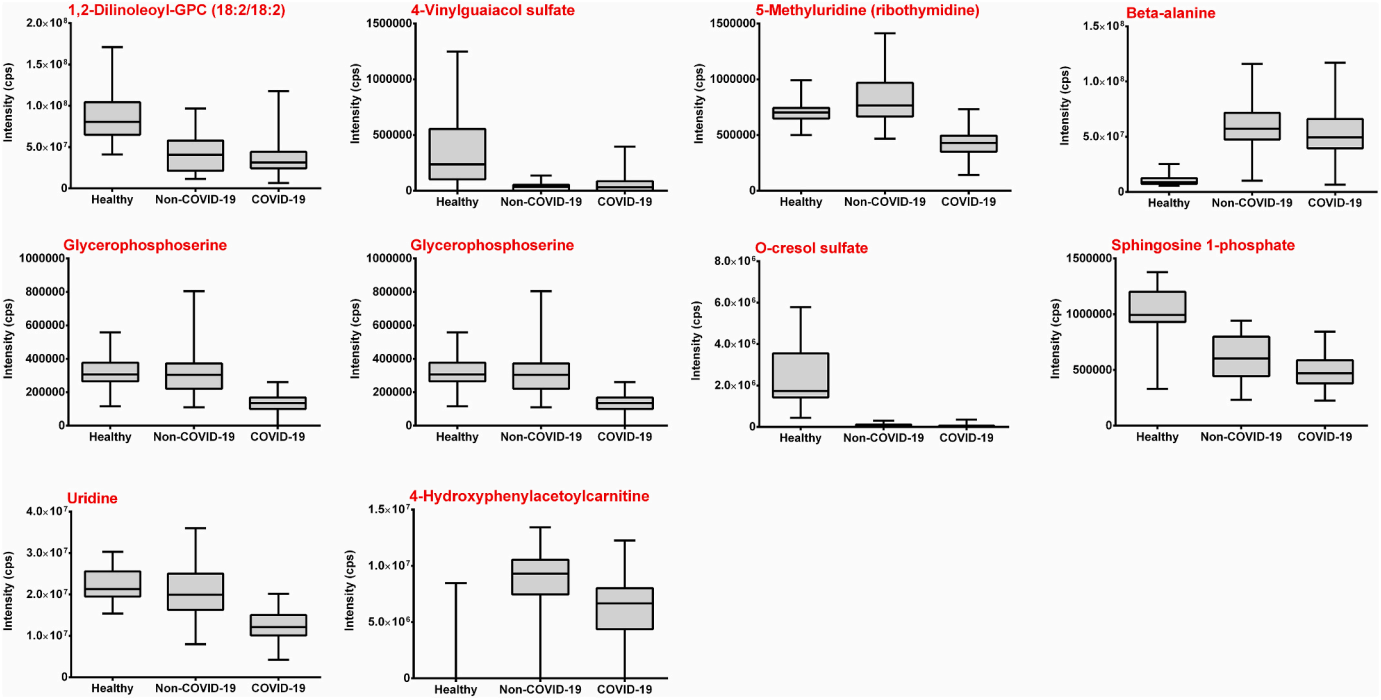

Fig. 5. Profile of the top 10 blood biomarkers associated with the diagnosis of COVID-19. Results are grouped according to the classes: healthy (n = 25), non-COVID-19 (n = 25) and COVID-19 (n = 46). Boxes indicate the interquartile ranges (median); horizontal lines indicate minimum and maximum values.

4. Discussion

We were able to develop and evaluate the performance of seven different ML-based models (ANNDA, PLS-DA, XGBoostDA, SIMCA, SVM, LREG and KNN) to predict the diagnosis of COVID-19 as well as disease severity and fatality using plasma and serum samples of patients from two reference hospitals in China. Over 1300 water-soluble and fat-soluble metabolites were assessed. As the course of COVID-19 is extremely variable among patients (especially due to emerging SARS-Cov-2 mutations and biological variability), the metabolomic profile of these cases is also uncertain, thus requiring more robust ML-based models [30–32].

In our study, the PLS-DA model presented the best performance (AUC ROC 87%–97%), with accuracy figures similar to those of other ML-based models available in the literature in this field (AUC ROC 70%–99%) [33]. The PLS-DA model is currently one of the most commonly used ML-based algorithms for analyzing data from metabolomics and other omics sciences (e.g. genomics, transcriptomics and proteomics), being recommended by experts in the field [34,35]. In fact, a systematic review assessing the number of citation of studies published between 1990 and 2018 and available in Web of Science showed an increase in publications citing PLS-DA (n = 2242), while other algorithms (e.g, ANN, SVM, RF, logistic regression, deep learning) were less commonly mentioned (n = 500) [36]. The main reasons for this include the intrinsic characteristics of the PLS-DA, considered a versatile algorithm with better predictive and descriptive advantages over other models. A recent study performed by Mendes et al. (2019) [35] compared the predictive performance of eight ML algorithms (PLS-DA, ANN, non-ANN, random forest (RF, radial basis function kernel support vector machines (SVM), logistic regression and principal components regression (PCR)), using 10 clinical metabolomics datasets available in the Metabolights and Metabolomics Workbenchre repositories, and reported the PLS-DA as the model with the best performance [35].

The PLS-DA is able to forecast highly multivariate data into a space of

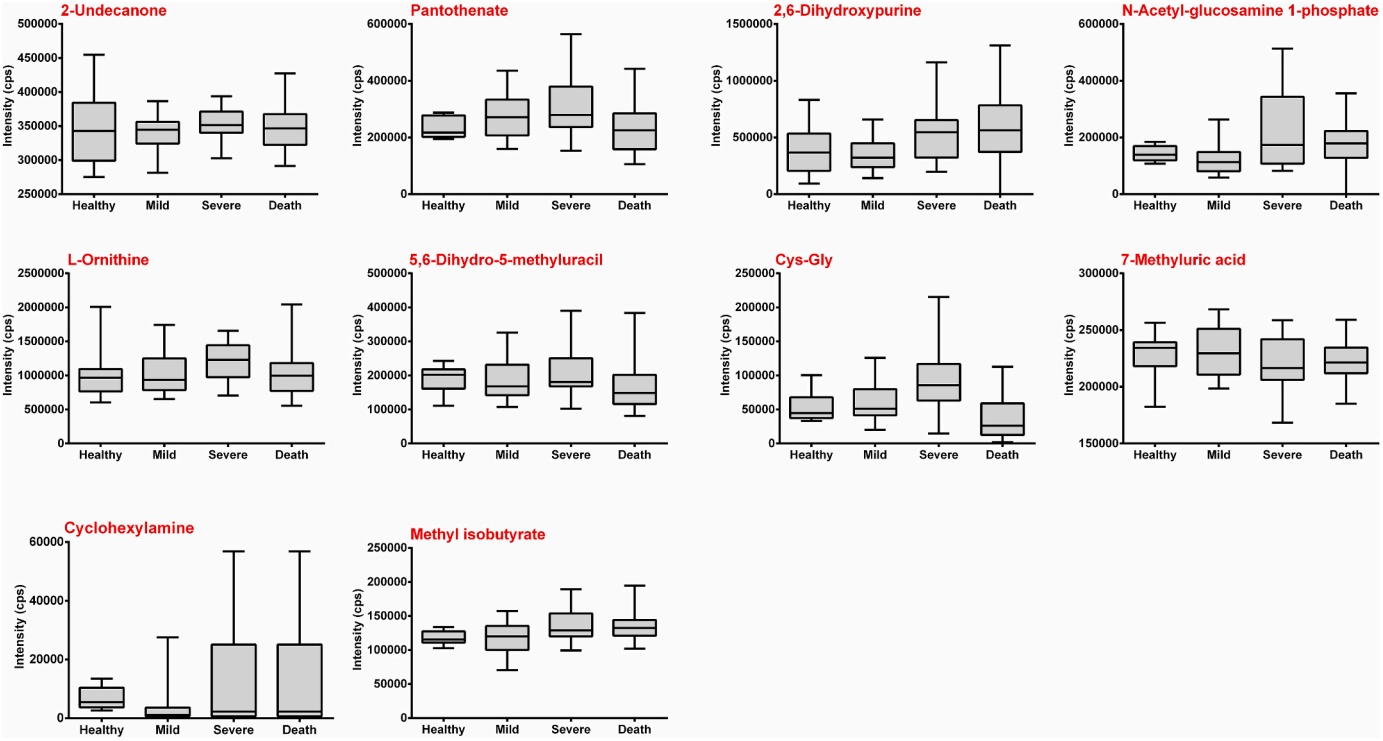

Fig. 6. Profile of the top 10 blood biomarkers associated with the COVID-19 severity and fatality. Results are grouped according to the classes: healthy (n = 10), mild COVID-19 (n = 28), severe COVID-19 (n = 22) and death (n = 36). Boxes indicate interquartile ranges (median); horizontal lines indicate minimum and maximum values.

Table 2
Biomarkers for predicting the diagnosis and severity/fatality of COVID-19 according to the PLS-DA models from SOLO software vs. and Metaboanalyst 5.0

	Diagnosis of COVID-19 (Dataset II - Thaizou Hospital)		COVID-19 severity/fatality (Dataset I - Wuhan Hospital)	
Rank	SOLO	Metaboanalyst	SOLO	Metaboanalyst
1	Dibutyl sulfosuccinate	Oleate	N-acetyl-glucosamine-1 phosphate	5-Triiodo-L-thyronine
2	O-cresol sulfate	Linoleate	Cys-glicine	L-Thyroxine
3	Beta-alanine	Palmitate	5,6Dihydro-5-methyluracil	2-Furanmethanol
4	Sphingosine 1-phosphate	Urea	Methyl isobutirate	4-Nitrophenol
5	4-Vinylguaiacol sulfate	Lactate	2-Undecanone	6-Methylmercaptopurine
6	4-Hydroxyphenylacetoylcarnitine	Carnitine	Pantothenate	4-Dihydroxybenzeneacetic acid
7	1,2 Dilinoleoyl-GPC (18:2/18:2)	Roline	L-Ornithine	L-phenylalanine
8	5-Methyluridine	Glycerophosphoethanolamine	2,6 Dihydroxypurine	L-citruline
9	Glicerophosphoserine	Stearate	Cyclohexilamine	Formylanthranilic acid
10	Uridine	phenylalanine	7-methyluric acid	Phthalic acid

smaller coordinates called LV (or principal components) that describe the variance between input data (e.g. metabolites) and output data (e.g. sample class) before regressing to a dependent variable. This allows datasets with more variables than samples to be modeled without resorting to pre-screening variables (essential for hypothesis-generating studies). Additionally, as it considers LV, problems regarding multicollinearity between the different metabolites in any biological system can be avoided (i.e. LV do not correlate with each other) [35,37]. Once optimized, the PLS-DA model can be reduced to a common linear regression model, enabling to predict the value of each metabolite/biomarker in the dataset [34,35]. Other ML-based models, including multilayer ANN, usually require larger sample sizes to achieve a high predictive performance. As a consequence, the number of variables included in these models is less than the number of samples [5,38–41], which is not the scenario of most metabolomic datasets [42–44]. In our study, we were able to use a dataset including 180 samples and 1300 variables.

Considering the great complexity of metabolomics data and the intrinsic properties of this information, missing values, heteroscedasticity, poorly informative parameters, and biological variability are commonplace. Data preprocessing is thus paramount to improve the quality of information by transforming the raw data matrix into a 'cleaner' set [29,45,46]. Several preprocessing strategies are available including missing data imputation, filtering, transformations, sample-based normalization, metabolite-based normalization, sample and metabolite-based and internal standard-based normalization [46–48]. This process may be conducted in free online tools such as MetaboAnalyst, NOREVA, ANPELA, NormalizeMets, MMEASE e Data Analysis [29,46–55]. Another challenge in metabolomic analysis is the integration of data from different experiments and the simultaneous removal of unwanted biological and experimental variations [55]. MMEASE is an online tool that allows merging this data and removing the effect of unwanted variations between samples, which increases the efficiency of statistical analyzes and leads to more robust and reliable results [55]. Data is merged according to the alignment ID for retention time (RT) and exact mass (m/z) of a given metabolite considered as a reference. If both RT and m/z of the reference metabolite fall within the tolerable range, this procedure is automatically applied to the

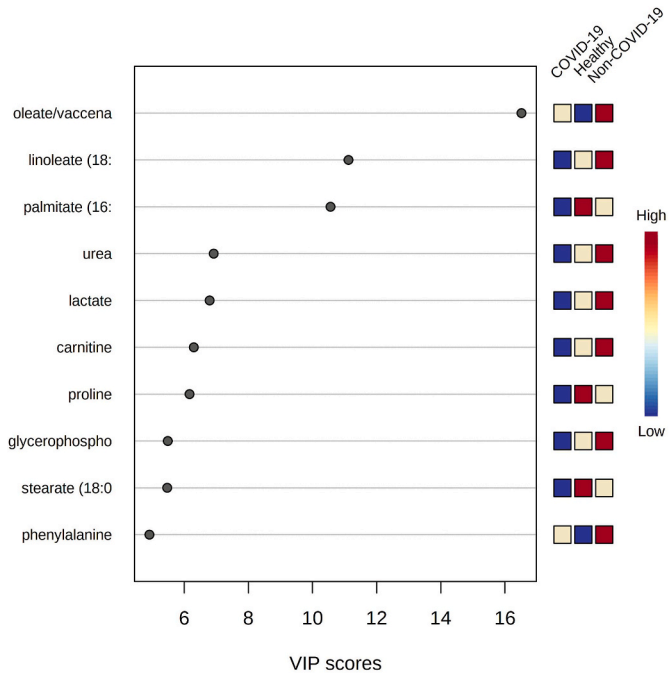

Fig. 7. Variable Importance in Projection (VIP) graph of the most important biomarkers for COVID-19 diagnosis (web server Metaboanalyst 5.0). The Y axis represents the top 10 most important metabolites in predicting COVID-19 diagnosis and the X axis represents the VIP score that reflects the importance of each metabolite in the prediction of the different classes of the samples (COVID-19, non-COVID-19, and healthy volunteers). The change from blue to red color is proportional to the increase in the intensity of the biomarker signal.

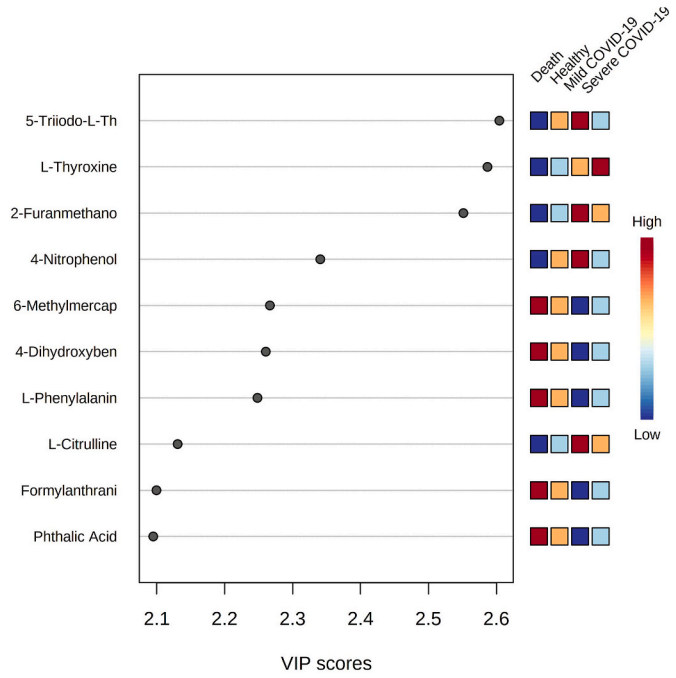

Fig. 8. Variable Importance in Projection (VIP) graph of the most important biomarkers for COVID-19 severity/fatality (web server Metaboanalyst 5.0). The Y axis represents the top 10 most important metabolites in predicting the severity of COVID-19 and the X axis represents the VIP score that reflects the importance of each metabolite in the prediction of the different classes of the samples (death, severe COVID-19, mild COVID-19, healthy individuals). The change from blue to red color is proportional to the increase in the intensity of the biomarker signal.

metabolites in the chromatograms of the remaining samples from other experiments [55]. Another alternative to eliminate problems from the batch effect is to use the Z-score method, which transforms the data to mean zero and standard deviation of 1, normalizing the distribution of analytical signals [49]. In our study, as only spectra data were available (RT of metabolites were absent), the MMEASE method was not employed. However, the datasets were standardized using the GLSW preprocessing method, which calculates a matrix of filters based on the differences between groups of samples that somehow should be 'similar' [28,56]. According to this method, in the case of classification problems, similar samples would be those whose data from the same samples were analyzed in different instruments or even in different periods [28,56]. In our study, we used two databases of patients with COVID-19 from two different experiments, whose samples were obtained by two different models of LC-MS equipments and time periods [28,56].

Methodological procedures for optimizing the processing of metabolomics data are better described in the guidelines of NOREVA (Normalization and Evaluation of MS-based Metabolomics Data), NormalizeMets, MMEASE, MetaboAnalyst and ANPELA [29,46–55]. Data preprocessing is usually performed in five steps: data filtering and missing value imputation (S1), quality control samples correction (S2), data transformation (S3), data normalization (S4) and performance assessment (S5). During S1, filtering focuses on removing uninformative features considered as intrinsic properties of the metabolomic data, while imputation seeks to replace missing or invalid values arising from technical/biological reasons with specific values based on available information, thus preserving the structure of the dataset, and reducing the imprecision or limitation of the analyses. The correction of quality control samples (S2) aims to reduce interference from harmful or uncontrollable signals in the metabolomic data to guarantee the stability and consistency of the data based on quality control samples. This allows to correct problems related to the variation in signal strength, intra- and inter-sample variability, and deviations in quality accuracy. In stages S3 and S4, the transformation and normalization of the metabolomic data aims to correct problems of heteroscedasticity and unwanted variations, transforming the distribution of asymmetric data into symmetrical ones, while preserving the existing variables. Finally, S5 consists on evaluating the performance of the pre-processing data based on five criteria: (i) ability to reduce intragroup variation among samples (metric: pooled median absolute deviation); (ii) effect on differential metabolic analysis (metric: purity); (iii) method's consistency in markers discovered from different datasets (metric: relative weighted consistency); (iv) method's influence on classification accuracy (metric: area under the curve); (v) level of correspondence between normalized and reference data (metric: log fold changes of the concentrations) [29,46–55]. In our study, the above-mentioned steps of data preprocessing were followed (i.e. imputation of missing values and filtering of data by means of GLSW were employed [28,57]; data were normalized using autoscale [52]). Median imputation is a widely used method in metabolomics, as unlike the mean, it is not affected by extreme values (outliers), which preserves the structure of the data and provides a more reliable value of the dataset [28,52]. Autoscale is an approach based on mean-centering followed by the division of each column or variable (e.g. protein or any other metabolite) by the column standard deviation, assuming that all metabolites are equally important [28,57].

Recent studies using blood or urine samples from patients diagnosed with COVID-19 highlighted that some biomarkers predict the severity and fatality of the disease. Yao et al. (2020), by using the SVM model, found that high levels of neutrophils were associated with more severe cases [58], while Patterson et al. (2020), through the random forest model, highlighted that an increase in interleukin 6 (IL-6) and interferon-gamma (IFN-γ) is related to a worse prognosis [59]. Conversely, using SOLO (Eigenvector Research) software we found different and new biomarkers potentially associated with the disease

course. High levels of ribothymidine, 4-hydroxyphenylacetoylcarnitine and uridine were associated with COVID-19 positivity, whereas high levels of N-acetyl-glucosamine-1-phosphate, cysteinylglycine, methyl isobutyrate, ornithine and 5,6-dihydro-5-methyluracil were related to COVID-19 severity and fatality. Differences among study findings may be due the different samples (i.e. type of sample, origin), the multifactorial pathophysiological course of the disease [60] that has not yet been fully elucidated [61,62], as well as the different analytical methods/models employed by the authors. Regarding this last, we also found that the analyzes conducted in SOLO software resulted in models with higher predictive performance compared to those from in Metaboanalyst 5.0 and identified different biomarkers for COVID-19 diagnosis and severity/fatality prediction (see qualitative comparison in Table 2). This may be due the differences on the preprocessing methods. While SOLO enables the combination of autoscale and GLSW, the Metaboanalyst 5.0 applies only this first approach, meaning that GLSW was, in this case, a determinant factor for obtaining more robust models. Although SOLO is not a free software, it allows the selection of different preprocessing strategies, providing further autonomy to the analysts, which should be considered when developing ML-based studies.

Currently, COVID-19 is broadly considered a viral respiratory and vascular illness. Yet, it can affect other major organs such as those of the gastrointestinal tract and the hepatobiliary, cardiovascular, renal, and central nervous systems. Recent evidence shows that SARS-CoV-2 can cause dysbiosis in the faecal microbiota and modify the oral and respiratory tract microbiome, leading to changes in the levels of several microbial metabolites in the blood or in their metabolic pathways [63–65]. Although evidence on the matter is still scarce, it has been reported that microbiota is responsible for around 50% of all blood metabolites [65], which raises questions about its role on multifactorial diseases, such as COVID-19 [63].

Li et al. (2019), by evaluating the nasopharyngeal microbiota profile of patients with COVID-19, found that positive samples were significantly enriched with the signature of two bacterial taxa (*Cutibacterium* and *Lentimonas*) and had a lower abundance of other bacterial taxa, including Prevotellaceae. The latter is a family of the phylum Bacteroidetes commonly found in the oral and faecal microbiota, recently associated with the metabolite ribothymidine (methylated nucleoside), which was increased in COVID-19-positive samples in our study. When overexpressed, these proteins actively contribute to the severity of pneumonia and pneumonia-like symptoms and are thus potential biomarkers for disease diagnosis and severity [66,67]. Similarly, high levels of 2-undecanone, a long-chain volatile organic compound usually produced during hospital-acquired bacterial infections caused by *Pseudomonas aeruginosa* [68,69], may be associated with severe cases of respiratory infections, including COVID-19. This substance can be found in patients with cystic fibrosis [70]. In fact, pulmonary fibrosis is a serious complication of some viral pneumonias, often leading to dyspnoea and impaired lung function. Patients with confirmed COVID-19 were found to have different degrees of pulmonary fibrosis at and after hospital discharge [71]. Sphingosine 1-phosphate (a product of membrane sphingolipid metabolism or secreted from cells), acts through G protein-coupled receptors and regulates immune cell trafficking, diverse immunological processes and fibrosis [72]. The pathway of this metabolite is implicated in normal pulmonary vasculature function; it appears to be impaired in acute lung dysfunction, while it is induced during chronic fibrosis. Further studies on the alteration of levels of this compound in COVID-19 are needed to elucidate its role in infection.

Another microbial metabolite, now associated with oral bacteria causing caries and periodontitis (e.g. *Porphyromonas gingivalis*, *Prevotella* sp. and *Tannerella forsythia*), is methyl isobutyrate [73]. Metagenomic analyses of patients infected with SARS-CoV-2 demonstrated high reads of cariogenic and periodontopathic bacteria, endorsing the notion of a connection between the oral microbiome and COVID-19 complications [73]. We also found high levels of cyclohexylamine (a potential carcinogenic compound eliminated in the urine) in patients with severe COVID-19. This probably occurs due to another dysbiosis caused by SARS-CoV-2, which allows the hyperproliferation of intestinal bacteria that metabolize cyclamate (an artificial sweetener still used in some food categories in China) [74,75]. Other compounds that are commonly found in foods and manufactured products (e.g. tobacco smoke) are the cresols (xenobiotics). O-cresol and 4-vinylguaiacol are converted to sulphates through phase II metabolism (i.e. a joint process between the microbiome and the host), and eliminated through the urine [76]. Previous studies demonstrated low levels of o-cresol sulphate and 4-vinylguaiacol sulphate in COVID-19 patients, which can be due to the high rates of urinary elimination of these metabolites (e.g. possible kidney damage caused by the disease) [77].

COVID-19 might also negatively impact body weight and nutritional status [78]. This may occur due to loss of appetite and reduced nutrient intake, patients' fear and stress regarding the disease and metabolic alterations in caused by the infection. For instance, the metabolite 4-hydroxyphenylacetoylcarnitine, found to be increased in patients with COVID-19 in our study, belongs to tyrosine metabolism and has been previously associated with overweight in patients with metabolic syndrome. Other studies also reported an increase in inflammation and serum levels of leptin in COVID-19 patients as in other infectious diseases that can contribute to anorexia [79–81]. These metabolites should be further investigated as potential biomarkers of viral infection severity.

Another important metabolite is uridine, a pyrimidine nucleotide for RNA synthesis that is associated with glucose homeostasis, lipid and amino acid metabolism, regulation of glycogen synthesis and lipid deposition [82]. During its catabolism, uridine is converted into β-alanine, followed by secretion to the brain and muscle tissues. Beta-alanine and histidine are components of carnosine, a molecule with proven anti-inflammatory, antioxidant and *anti*-glycating effects [83]. In our first model, levels of beta-alanine were found to be low in COVID-19 patients, while those of uridine were high. This may indicate inhibition of uridine catabolism during the course of the infection. A recent study found a significantly low ratio of arginine/ornithine among adults and children infected with SARS-Cov-2. Ornithine and citrulline are amino acids resulting from the breakdown of arginine by the arginase enzyme. The depletion of these substances may contribute to endothelial dysfunction, T-cell dysregulation and coagulopathies that are commonly observed in COVID-19 [84]. The high level of ornithine in COVID-19 patients that was reported in our study may indicate increased activity of the arginase enzyme.

N-acetyl-glucosamine-1-phosphate (GlcNAc-1-P) is a substrate of the biosynthetic pathway of hexosamines, converted by the enzyme UDP-GlcNAc pyrophosphorylase into UDP-GlcNAc (this metabolite can use the O-glycosylation route). This conversion is an important step in the production of cytokines during influenza virus infection, as demonstrated in *vivo* models (murine models) [85]. Researchers believe that inhibition of the hexosamine pathway is a mechanism used by respiratory viruses, including SARS-Cov-2, to infect host cells [86,87]. The elevated level of GlcNAc-1-P in patients with severe COVID-19 reveals a potential modification of the hexosamine biosynthetic pathway. Additionally, as GlcNAc-1-P is an intracellular component, its presence in the plasma indicates the existence of cellular damage. SARS-CoV-2 infection leads to pyroptosis, which is usually more prevalent in severe cases. More than half of hospitalized COVID-19 patients present high levels of lactate dehydrogenase, another marker of cell damage [88,89]. Regulators of oxidative stress such as cysteinylglycine, an intermediate metabolite in the glutathione metabolic pathway, have also been associated to cell damage in viral diseases. High levels of oxidized cysteinylglycine were reported in HIV-infected individuals and also related to a higher risk of lung damage in COVID-19, probably due increased oxidative stress [90,91]. Other metabolites such as 5,6-dihydro-5-methyluracil (dihydrothymine), an intermediate breakdown product of thymine, may act as markers of DNA damage [92]. We found that the levels of this substance were high in patients with severe COVID-19, but

recent studies demonstrated that the spike protein from SARS-CoV-2 can inhibit repair of damaged DNA [93]. Other metabolites identified at extremely low levels in patients with COVID-19 (e.g. linoleate, palmitate, urea, lactate, carnitine) when using the Metaboanalyst 5.0 software, or those found at high levels in patients who died from the disease (e.g. 6-methylmercaptopurine, L-phenylalanine, terephthalic acid) should also be further evaluated.

Our study has some limitations. Although we used approximately 1300 different biomarkers for model training and validation, these may not accurately represent the universe of metabolites available in the blood. Yet, it was possible to obtain models with high performance (accuracy >90%) for the prediction of diagnosis, severity and fatality of COVID-19 that can be used in daily practice. Seven different ML-based models grounded in data from two different sets from China were built in our study; however, other datasets and algorithms may lead to different findings.

5. Conclusion

In this study, seven different ML-based algorithms (PLS-DA, KNN, XGboost, SVM, ANN, SIMCA and LREG) were built to predict the diagnosis, severity and fatality of COVID-19 using two different databases. The PLS-DA model presented the best performance, with an accuracy of approximately 93%. This model can aid in the early diagnosis of COVID-19 and guide disease management with additional interventions tailored to daily practice. Finally, some of the biomarkers associated with the diagnosis and prognosis of COVID-19 found in the sample set of our study (i.e. 5,6-dihydro-5-methyluracil, cysteinylglycine, ribothymidine, sphingosine 1-phosphate, cyclohexylamine, uridine and ornithine) have previously been mentioned in the scientific literature, which reinforces their role in infection. Conversely, we reported for the first-time additional biomarkers (i.e. N-acetyl-glucosamine-1-phosphate and 4-hydroxyphenylacetoylcarnitine) that should be evaluated further as prognostic indicators of COVID-19.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Study concepts: AFC, RP, FST, MMF.
Study design: AFC, FST, RP, DPS, MMF.
Data acquisition: AFC, MS, HHLB.
Statistical analysis: AFC, DPS, MMF.
Manuscript preparation: AFC, RP, MS, FST, HHLB, DPS.
Manuscript editing: AFC, RP, MS, FST, HHLB.
Manuscript review: AFC, RP, MS, FST, HHLB, DPS.

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their gratitude to the Brazilian National Council of Technological and Scientific Development (CNPq) and CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil) for research funding - Finance Code 001.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105659.

References

[1] D. Skegg, et al., Future scenarios for the COVID-19 pandemic, Lancet 397 (2021) 777–778, https://doi.org/10.1016/S0140-6736(21)00424-4.
[2] World Health Organization, Coronavirus disease (COVID-19) dashboard. https://covid19.who.int, 2022. (Accessed 3 May 2022).
[3] P. Asrani, et al., Diagnostic approaches in COVID-19: clinical updates, Expert Rev. Respir. Med. 15 (2021) 197–212, https://doi.org/10.1080/17476348.2021.1823833.
[4] J. Majumder, T. Minko, Recent developments on therapeutic and diagnostic approaches for COVID-19, AAPS J 23 (2021) 14, https://doi.org/10.1208/s12248-020-00532-2.
[5] C. Réda, et al., Machine learning applications in drug development, Comput. Struct. Biotechnol. J. 18 (2020) 241–252, https://doi.org/10.1016/j.csbj.2019.12.006.
[6] C.H. Chang, et al., Machine learning and novel biomarkers for the diagnosis of alzheimer's disease, Int. J. Mol. Sci. 22 (2021) 2761, https://doi.org/10.3390/ijms22052761.
[7] P. Shah, et al., Artificial intelligence and machine learning in clinical development: a translational perspective, Npj Digit. Med. 2 (2019) 69, https://doi.org/10.1038/s41746-019-0148-3.
[8] A.L. Udriștoiu, et al., COVID-19 and artificial intelligence: an approach to forecast the severity of diagnosis, Life (Basel, Switzerland) 11 (2021) 1281, https://doi.org/10.3390/life11111281.
[9] R. Fusco, et al., Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: machine learning and deep learning approaches for diagnosis and treatment, J. Pers. Med. 11 (2021) 993, https://doi.org/10.3390/jpm11100993.
[10] X. Zhang, et al., Machine learning approaches for biomarker discovery using gene expression data, in: H.I. Nakaia (Ed.), Bioinformatics, Exon Publications., Brisbane (AU), 2021, pp. 53–64, https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4.
[11] X. Mi, et al., Permutation-based identification of important biomarkers for complex diseases via machine learning models, Nat. Commun. 12 (2021) 3008, https://doi.org/10.1038/s41467-021-22756-2.
[12] A. de F. Cobre, et al., Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? Comput. Biol. Med. 134 (2021) 104531, https://doi.org/10.1016/j.compbiomed.2021.104531.
[13] X. Ma, et al., Development and validation of prognosis model of mortality risk in patients with COVID-19, Epidemiol, Infect 148 (2020) 168, https://doi.org/10.1017/S0950268820001727.
[14] D. Assaf, et al., Utilization of machine-learning models to accurately predict the risk for critical COVID-19, Intern. Emerg. Med. 15 (2020) 1435–1443, https://doi.org/10.1007/s11739-020-02475-0.
[15] M. Kukar, et al., COVID-19 diagnosis by routine blood tests using machine learning, Sci. Rep. 11 (2021) 10738, https://doi.org/10.1038/s41598-021-90265-9.
[16] D.L. Kitane, et al., A simple and fast spectroscopy-based technique for Covid-19 diagnosis, Sci. Rep. 11 (2021) 16740, https://doi.org/10.1038/s41598-021-95568-5.
[17] J.M. Mir, et al., A nonclinical spectroscopic approach for diagnosing COVID-19: a concise perspective, J. Appl. Spectrosc. 88 (2021) 765–771, https://doi.org/10.1007/s10812-021-01238-9.
[18] C. Ma, et al., Drastic decline in sera neutralization against SARS-CoV-2 omicron variant in Wuhan COVID-19 convalescents, Emerg, Microbes \& Infect 11 (2022) 567–572, https://doi.org/10.1080/22221751.2022.2031311.
[19] Z. Pang, et al., Comprehensive meta-analysis of COVID-19 global metabolomics datasets, Metabolites 11 (2021) 44, https://doi.org/10.3390/metabo11010044.
[20] J. Burdack, et al., Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning, Front. Bioeng. Biotechnol. 8 (2020) 260, https://doi.org/10.3389/fbioe.2020.00260.
[21] S.B. Kotsiantis, D. Kanellopoulos, Data preprocessing for supervised leaning, Int. J. Comput. Sci. 1 (2006) 1, https://doi.org/10.1080/02331931003692557.
[22] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790–3798, https://doi.org/10.1039/c3ay40582f.
[23] B. Walczak, D.L. Massart, Multiple outlier detection revisited, Chemom. Intell. Lab. Syst. 41 (1998) 1–15. https://doi.org/10.1016/S0169-7439(98)00034-3.
[24] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.
[25] J.W. Dr-Ing, et al., Cross-validation and robustness of daylight glare metrics, Lighting. Res. Technol. 51 (2019) 983–1013, https://doi.org/10.1177/1477153519826003.
[26] M. Cocchi, et al., Chapter Ten - chemometric methods for classification and feature selection, in: J. Jaumot, C. Bedia, R. Tauler (Eds.), Data Anal. Omi. Sci. Methods Appl., Elsevier, Amsterdam, 2018, pp. 265–299. https://doi.org/10.1016/bs.coac.2018.08.006.

[27] S. Favilla, et al., Assessing feature relevance in NPLS models by VIP, Chemom. Intell. Lab. Syst. 129 (2013) 76–86. https://doi.org/10.1016/j.chemolab.2013.05.013.
[28] B.M. Wise, et al., Chemometrics tutorial for PLS toolbox_solo, Eigenvector Research 3905 (2006) 102–159. http://mitr.p.lodz.pl/raman/jsurmacki/pliki/zajecia/LMDiT/cw3/LMDiT_PLS_Manual_4.pdf.
[29] Z. Pang, et al., MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights, Nucleic Acids Res 49 (2021) 388–396, https://doi.org/10.1093/nar/gkab382.
[30] C.M. Voloch, et al., Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil, J. Virol. 95 (2021) e00119–e00121, https://doi.org/10.1128/JVI.00119-21.
[31] R. Wang, et al., Mutations on COVID-19 diagnostic targets, Genomics 112 (2020) 5204–5213. https://doi.org/10.1016/j.ygeno.2020.09.028.
[32] C.K. V Nonaka, et al., Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil, Emerg, Infect. Dis. 27 (2021) 1522–1524, https://doi.org/10.3201/eid2705.210191.
[33] L. Wang, et al., Artificial intelligence for COVID-19: a systematic review, Front. Med. 8 (2021) 704256, https://doi.org/10.3389/fmed.2021.704256.
[34] P.S. Gromski, et al., A tutorial review: metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding, Anal. Chim. Acta. 879 (2015) 10–23, https://doi.org/10.1016/j.aca.2015.02.012.
[35] K.M. Mendez, et al., A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification, Metabolomics 15 (2019) 150, https://doi.org/10.1007/s11306-019-1612-4.
[36] K.M. Mendez, et al., The application of artificial neural networks in metabolomics: a historical perspective, Metabolomics 15 (2019) 142, https://doi.org/10.1007/s11306-019-1608-0.
[37] W.B. Dunn, et al., Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy, Chem. Soc. Rev. 40 (2011) 387–426, https://doi.org/10.1039/b906712b.
[38] A. Alwosheel, et al., Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis, J. Choice Model. 28 (2018) 167–182, https://doi.org/10.1016/j.jocm.2018.07.002.
[39] K.T. Butler, et al., Machine learning for molecular and materials science, Nature 559 (2018) 547–555, https://doi.org/10.1038/s41586-018-0337-2.
[40] S.J. Qin, L.H. Chiang, Advances and opportunities in machine learning for process data analytics, Comput. Chem. Eng. 126 (2019) 465–473. https://doi.org/10.1016/j.compchemeng.2019.04.003.
[41] B. Liu, et al., Deep neural networks for high dimension, low sample size data, in: Proc. Twenty-Sixth Int. Jt. Conf. Artif. Intell. {IJCAI-17}, Melbourne, 2017, pp. 2287–2293, https://doi.org/10.24963/ijcai.2017/318.
[42] D. Ruiz-Perez, et al., So you think you can PLS-DA? BMC Bioinformatics 21 (2020) 1–10, https://doi.org/10.1186/s12859-019-3310-7.
[43] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173, https://doi.org/10.1002/cem.785.
[44] L. Eriksson, et al., Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm), Anal. Bioanal. Chem. 380 (2004) 419–429, https://doi.org/10.1007/s00216-004-2783-y.
[45] J. Tang, et al., Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains, Mol. Cell. Proteomics 18 (2019) 1683–1699, https://doi.org/10.1074/mcp.RA118.001169.
[46] J. Fu, et al., Pharmacometabonomics: data processing and statistical analysis, Brief. Bioinform. 22 (2021) 1–25, https://doi.org/10.1093/bib/bbab138.
[47] J. Tang, et al., MetaFS: performance assessment of biomarker discovery in metaproteomics, Brief. Bioinform. 22 (2020) 1–11, https://doi.org/10.1093/bib/bbaa105.
[48] F. Li, et al., POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability, Brief, Bioinform 23 (2022) 1–16, https://doi.org/10.1093/bib/bbac040.
[49] Q. Yang, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Brief. Bioinform. 21 (2020) 1058–1068, https://doi.org/10.1093/bib/bbz049.
[50] J. Fu, et al., Optimization of metabolomic data processing using NOREVA, Nat. Protoc. 17 (2022) 129–151, https://doi.org/10.1038/s41596-021-00636-9.
[51] Q. Yang, et al., NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data, Nucleic Acids Res 48 (2020) 436–448, https://doi.org/10.1093/nar/gkaa258.
[52] J. Tang, et al., Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains, Mol. Cell. Proteomics 18 (2019) 1683–1699, https://doi.org/10.1074/mcp.RA118.001169.
[53] A.M. de Livera, et al., NormalizeMets: assessing, selecting and implementing statistical methods for normalizing metabolomics data, Metabolomics 14 (2018) 54, https://doi.org/10.1007/s11306-018-1347-7.
[54] J. Tang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2020) 621–636, https://doi.org/10.1093/bib/bby127.
[55] Q. Yang, et al., MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis, J. Proteomics. 232 (2021) 104023, https://doi.org/10.1016/j.jprot.2020.104023.
[56] G. Bertol, et al., Differentiation of *Mikania glomerata* and *Mikania laevigata* species through mid-infrared spectroscopy and chemometrics guided by HPLC-DAD analyses, Rev. Bras. Farmacogn. 31 (2021) 442–452, https://doi.org/10.1007/s43450-021-00170-5.
[57] S. Serranti, et al., Classification of oat and groat kernels using NIR hyperspectral imaging, Talanta 103 (2013) 276–284, https://doi.org/10.1016/j.talanta.2012.10.044.
[58] H. Yao, et al., Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests, Front. Cell Dev. Biol. 8 (2020) 683, https://doi.org/10.3389/fcell.2020.00683.
[59] B.K. Patterson, et al., Immune-based prediction of COVID-19 severity and chronicity decoded using machine learning, Front. Immunol. 12 (2021) 700782, https://doi.org/10.3389/fimmu.2021.700782.
[60] M. Cascella, et al., Features, evaluation, and treatment of coronavirus (COVID-19). https://www.ncbi.nlm.nih.gov/books/NBK554776/, 2022. (Accessed 7 February 2022).
[61] S. Choudhary, et al., Role of genetic variants and gene expression in the susceptibility and severity of COVID-19, Ann, Lab. Med. 41 (2021) 129–138, https://doi.org/10.3343/alm.2021.41.2.129.
[62] L.J. Abu-Raddad, et al., Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants, N. Engl. J. Med. 385 (2021) 187–189, https://doi.org/10.1056/NEJMc2104974.
[63] S. Yamamoto, et al., The human microbiome and COVID-19: a systematic review, PLoS One 16 (2021) 253293, https://doi.org/10.1371/journal.pone.0253293.
[64] J. Patel, V. Sampson, The role of oral bacteria in COVID-19, the Lancet, Microbe 1 (2020) 105, https://doi.org/10.1016/S2666-5247(20)30057-4.
[65] A. Visconti, et al., Interplay between the human gut microbiome and host metabolism, Nat. Commun. 10 (2019) 4505, https://doi.org/10.1038/s41467-019-12476-z.
[66] D. Yang, et al., Implications of gut microbiota dysbiosis and metabolic changes in prion disease, Neurobiol. Dis. 135 (2020) 104704, https://doi.org/10.1016/j.nbd.2019.104704.
[67] N. Li, et al., The commensal microbiota and viral infection: a comprehensive review, Front. Immunol. 10 (2019) 1551, https://doi.org/10.3389/fimmu.2019.01551.
[68] B.J. Langford, et al., Bacterial co-infection and secondary infection in patients with COVID-19: a living rapid review and meta-analysis, Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 26 (2020) 1622–1629, https://doi.org/10.1016/j.cmi.2020.07.016.
[69] C. Garcia-Vidal, et al., Incidence of co-infections and superinfections in hospitalized patients with COVID-19: a retrospective cohort study, Clin. Microbiol. Infect. 27 (2021) 83–88, https://doi.org/10.1016/j.cmi.2020.07.041.
[70] C.M. Timm, et al., Direct growth of bacteria in headspace vials allows for screening of volatiles by gas chromatography mass spectrometry, Front. Microbiol. 9 (2018) 491, https://doi.org/10.3389/fmicb.2018.00491.
[71] J.N. Zou, et al., The characteristics and evolution of pulmonary fibrosis in COVID-19 patients as assessed by AI-assisted chest HRCT, PLoS One 16 (2021) 248957, https://doi.org/10.1371/journal.pone.0248957.
[72] A. Cartier, T. Hla, Sphingosine 1-phosphate: lipid signaling in pathology and therapy, Science 366 (2019) 5551, https://doi.org/10.1126/science.aar5551.
[73] K. Roslund, et al., Identifying volatile in vitro biomarkers for oral bacteria with proton-transfer-reaction mass spectrometry and gas chromatography-mass spectrometry, Sci. Rep. 11 (2021) 16897, https://doi.org/10.1038/s41598-021-96287-7.
[74] Y. Wang, et al., Estimated assessment of dietary exposure to artificial sweeteners from processed food in Nanjing, China, Food Addit. Contam. Part A, Chem. Anal. Control. Expo. Risk Assess. 38 (2021) 1105–1117, https://doi.org/10.1080/19440049.2021.1905883.
[75] M. Eichelbaum, et al., Pharmacokinetics, cardiovascular and metabolic actions of cyclohexylamine in man, Arch. Toxikol. 31 (1974) 243–263, https://doi.org/10.1007/BF00311057.
[76] G. Liu, et al., Metabolomic analysis identified reduced levels of xenobiotics, oxidative stress, and improved vitamin metabolism in smokers switched to vuse electronic nicotine delivery system, Nicotine Tob. Res. 23 (2021) 1133–1142, https://doi.org/10.1093/ntr/ntaa225.
[77] X. Chen, et al., Metabolite reanalysis revealed potential biomarkers for COVID-19: a potential link with immune response, Future Microbiol 16 (2021) 577–588, https://doi.org/10.2217/fmb-2021-0047.
[78] A.F. Cobre, et al., Influence of foods and nutrients on COVID-19 recovery: a multivariate analysis of data from 170 countries using a generalized linear model, Clin. Nutr. (2021), https://doi.org/10.1016/j.clnu.2021.03.018.
[79] M.K. Sikaroudi, et al., Assessment of anorexia and weight loss during the infection and recovery period of patients with coronavirus disease 2019 (COVID-19), Clin. Nutr. Open Sci. 40 (2021) 102–110, https://doi.org/10.1016/j.nutos.2021.11.001.
[80] L. Di Filippo, et al., COVID-19 is associated with clinically significant weight loss and risk of malnutrition, independent of hospitalisation: a post-hoc analysis of a prospective cohort study, Clin. Nutr. 40 (2021) 2420–2426, https://doi.org/10.1016/j.clnu.2020.10.043.
[81] P.H.J. van der Voort, et al., Leptin levels in SARS-CoV-2 infection related respiratory failure: a cross-sectional study and a pathophysiological framework on the role of fat tissue, Heliyon 6 (2020) 4696, https://doi.org/10.1016/j.heliyon.2020.e04696.
[82] Y. Zhang, et al., Uridine metabolism and its role in glucose, lipid, and amino acid homeostasis, Biomed Res. Int. 2020 (2020) 7091718, https://doi.org/10.1155/2020/7091718.
[83] J. Feehan, et al., Nutritional interventions for COVID-19: a role for carnosine? Nutrients 13 (2021) 1463, https://doi.org/10.3390/nu13051463.

[84] C.A. Rees, et al., Altered amino acid profile in patients with SARS-CoV-2 infection, Proc. Natl. Acad. Sci. U. S. A. 118 (2021) 2101708118, https://doi.org/10.1073/pnas.2101708118.
[85] Q. Wang, et al., O-GlcNAc transferase promotes influenza a virus-induced cytokine storm by targeting interferon regulatory factor-5, Sci. Adv. 6 (2020) 7086, https://doi.org/10.1126/sciadv.aaz7086.
[86] J.S. Ayres, A metabolic handbook for the COVID-19 pandemic, Nat. Metab. 2 (2020) 572–585, https://doi.org/10.1038/s42255-020-0237-2.
[87] H.A. Laviada-Molina, et al., Working hypothesis for glucose metabolism and SARS-CoV-2 replication: interplay between the hexosamine pathway and interferon RF5 triggering hyperinflammation. role of BCG vaccine? Front. Endocrinol. (Lausanne). 11 (2020) 514, https://doi.org/10.3389/fendo.2020.00514.
[88] M.M. da Silva, et al., Cell death mechanisms involved in cell injury caused by SARS-CoV-2, Rev. Med. Virol. (2021) 2292, https://doi.org/10.1002/rmv.2292.
[89] U. Jain, Effect of COVID-19 on the organs, Cureus 12 (2020) 9540, https://doi.org/10.7759/cureus.9540.
[90] E.V. Kryukov, et al., Association of low molecular weight plasma aminothiols with the severity of coronavirus disease 2019, Oxid. Med. Cell. Longev. 2021 (2021) 9221693, https://doi.org/10.1155/2021/9221693.
[91] L. Pei, et al., Plasma metabolomics reveals dysregulated metabolic signatures in HIV-associated immune reconstitution inflammatory syndrome, Front. Immunol. 12 (2021) 693074, https://doi.org/10.3389/fimmu.2021.693074.
[92] D.J. Wilkinson, et al., Untargeted metabolomics for uncovering biological markers of human skeletal muscle ageing, Aging (Albany. NY) 12 (2020) 12517–12533, https://doi.org/10.18632/aging.103513.
[93] H. Jiang, Y.F. Mei, SARS-CoV-2 spike impairs DNA damage repair and inhibits V(D)J recombination *in vitro*, Viruses 13 (2021), https://doi.org/10.3390/v13102056, 2056.