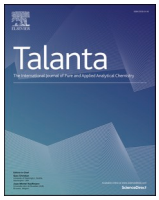
Talanta 279 (2024) 126652

Contents lists available at ScienceDirect

Talanta

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



Spectrochemical and explainable artificial intelligence approaches for molecular level identification of the status of critically ill patients with COVID-19

Gorkem Tokgoz [a], K. Kubra Kirboga [a], Faik Ozel [b], Serkan Yucepur [c], Isa Ardahanli [d], Rafig Gurbanov [a,e,*]

[a] Department of Bioengineering, Faculty of Engineering, Bilecik Şeyh Edebali University, Bilecik, 11100, Turkey
[b] Department of Internal Medicine, Faculty of Medicine, Bilecik Şeyh Edebali University Bilecik, 11100, Turkey
[c] Department of Anesthesiology and Reanimation, Faculty of Medicine, Bilecik Şeyh Edebali University Bilecik, 11100, Turkey
[d] Department of Cardiology, Faculty of Medicine, Bilecik Şeyh Edebali University Bilecik, 11100, Turkey
[e] Central Research Laboratory, Bilecik Şeyh Edebali University, Bilecik, 11100, Turkey

ARTICLE INFO

Handling editor: Kin-ichi Tsunoda

Keywords:
COVID-19
FTIR
Biomarker
Explainable artificial intelligence
Shapley explanations

ABSTRACT

This study explores the molecular alterations and disease progression in COVID-19 patients using ATR-FTIR spectroscopy combined with spectrochemical and explainable artificial intelligence (XAI) approaches. Blood serum samples from intubated patients (IC), those receiving hospital services (SC), and recovered patients (PC) were analyzed to identify potential spectrochemical serum biomarkers. Spectrochemical parameters such as lipid, protein, nucleic acid concentrations, and IgG glycosylation were quantified, revealing significant alterations indicative of disease severity. Notably, increased lipid content, altered protein concentrations, and enhanced protein phosphorylation were observed in IC patients compared to SC and PC groups. The serum AGR (Albumin/Globulin Ratio) index demonstrated a distinct shift among patient groups, suggesting its potential as a rapid biochemical marker for COVID-19 severity. Additionally, alterations in IgG glycosylation and glucose concentrations were associated with disease severity. Spectral analysis highlighted specific bands indicative of nucleic acid concentrations, with notable changes observed in IC patients. XAI techniques further elucidated the importance of various spectral features in predicting disease severity across patient categories, emphasizing the heterogeneity of COVID-19's impact. Overall, this comprehensive approach provides insights into the molecular mechanisms underlying COVID-19 pathogenesis and offers a transparent and interpretable prediction algorithm to aid decision-making and patient management.

1. Introduction

At the end of 2019, humanity faced the coronavirus disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus. Following the rapid and worldwide spread of SARS-CoV-2, the World Health Organization (WHO) declared COVID-19 a global pandemic on March 11, 2020 [1]. Coronaviruses are enveloped, non-segmented, positive-sense RNA viruses commonly found in humans, other mammals, and birds that cause respiratory, enteric, hepatic, and neurological diseases. All seven coronavirus strains are known to cause human disease and mainly cause cold symptoms in non-immune individuals. At the same time, SARS-CoV-2 presents many clinical signs, including severe pneumonia and cold symptoms [2]. The pathogenesis of this disease is a complex process in which the virus binds to human cells and subsequent infection begins. The spectrochemical methods are used as important tools for molecular-level prediction of disease pathogenesis. Spectroscopy has recently become a tool in biomedical applications, making significant strides in clinical evaluation. Techniques, like Fourier transform infrared (FTIR) spectroscopy, have been used on various biological tissues due to their simplicity, reproducibility, and non-destructive nature. These methods require only minimal sample preparation and small amounts of material, from micrograms to nanograms. They provide molecular-level information, allowing the study of functional groups, bonding types, and molecular

* Corresponding author. Department of Bioengineering, Faculty of Engineering, Bilecik Şeyh Edebali University, Bilecik, 11100, Turkey.
E-mail address: rafig.gurbanov@bilecik.edu.tr (R. Gurbanov).

https://doi.org/10.1016/j.talanta.2024.126652
Received 3 May 2024; Received in revised form 29 July 2024; Accepted 31 July 2024
Available online 31 July 2024

conformations. The spectral bands in vibrational spectra are molecule-specific, offering precise details about biochemical composition, molecular structure, conformation, and environment [3,4]. Hence, these optical techniques are potentially useful for gaining insights into the pathogenesis of COVID-19 at the molecular level [5].

Current research has associated the severity of COVID-19 and the associated risk of death with many disease factors, including demographical, biochemical, and molecular manifestations [6,7]. However, a complete and comprehensive understanding of immune responses against SARS-CoV-2 at the molecular level is still lacking. Therefore, it is crucial to identify and characterize the immune responses associated with disease pathogenesis and severity by enabling the identification of predictive biomarkers. On the other hand, the intensive admission of SARS-CoV-2-infected patients to hospitals during the pandemic indicates the need to fully understand the patient characteristics and laboratory findings related to disease severity and mortality. Within this framework, artificial intelligence (AI) technologies are applied to minimize human errors and increase diagnostic efficiency by using image inputs as decision-support tools [8]. However traditional AI models serve as black box models for most researchers and professionals, using them for various tasks, including medical prediction purposes. Such traditional AI methods lack the details and explanations to help doctors make accurate decisions and interpretations. Explainable AI (XAI) provides this opportunity that transfer AI-based black box models to more explainable and transparent white box models [9–11].

In this research, we assessed disease progression and molecular alterations in patients admitted to the hospital with COVID-19 diagnosis. Despite the availability of multiple strategies for preventing, diagnosing, and treating COVID-19, the significant hospitalization rates of SARS-CoV-2-infected individuals across a spectrum of disease severity, coupled with elevated mortality rates, underscore the urgency of identifying and validating disease biomarkers. In this context, blood serum emerges as an ideal candidate for early disease detection, as it is a highly accessible and highly informative biofluid [12,13]. When infrared (IR) spectroscopy is evaluated together with different analysis techniques, it becomes a valuable tool for predicting diseases and assessing treatment efficacy [14–16]. Thus, we employed the ATR-FTIR spectroscopy technique in conjunction with diverse statistical analyses and explainable artificial intelligence approaches. Our goal was to evaluate alterations in serum biomolecules among COVID-19 patients admitted to Bilecik Training and Research Hospital. Through this comprehensive approach, we aimed to identify and validate early disease-associated biomarkers in intubated patients (Intubated COVID-19/IC), patients receiving treatment at a service point of the hospital (Serviced COVID-19/SC), and patients recovered from the disease (Post-COVID-19/PC). XAI approaches are utilized to analyze the spectral datasets. The objective is to develop a transparent and interpretable prediction algorithm utilizing IR spectra as a source of molecular data to aid in the decision-making process in patient care. Identifying factors that predict the progression of the disease can potentially help professionals prioritize patients, personalize treatment plans, monitor clinical progress, and potentially reduce morbidity and mortality. In managing the pandemic, the ability to rapidly process molecular data and understand disease pathogenesis can accelerate decision-making and improve patient management.

2. Methods

2.1. Spectrochemical methods

2.1.1. Population sampled and medical evaluation

This study included serum samples taken from critically ill 132 patients hospitalized with a positive diagnosis of COVID-19 infection in the COVID service and intensive care units of Bilecik Training and Research Hospital between June and August 2021. The diagnosis was made by RT-PCR assay and CT scan. Disease levels of individuals were classified as Intubated COVID-19 (IC), Serviced COVID-19 (SC), and Recovered from COVID-19 (Post-COVID-19/PC). This study was conducted according to Bilecik Şeyh Edebali University Clinical Research Ethics Committee guidelines (ethics committee number: 050.04.01–210460). Invited volunteers were informed about this study's aims, recommendations, and conditions, and those who agreed to participate signed the free and informed consent form. A 2 mL of whole blood was collected by peripheral vein puncture from each patient and the samples were centrifuged at 3.000 rpm for 15 min at +4 °C to separate the serum from the cellular components. Then, sera were aliquoted in 500 μL volumes and stored in a deep freezer at −80 °C until the analysis. The study included 40 intubated (16 female, 24 male), 33 serviced (10 female, 23 male), and 59 recovered (26 female, 33 male) individuals, aged between 22 and 91 years. Overall, 39.4 % of the individuals were female and 60.6 % were male. Demographic information about the patients is presented in Supplementary Table S1.

Intubation is the process of inserting a tube into the trachea (windpipe) through the mouth or nose to keep the airway open and provide mechanical ventilation in patients with respiratory failure. The intubated patient group in our study was kept under invasive mechanical ventilation in the intensive care unit, and sedation was applied during this process.

Exclusion criteria: The patients intubated for other reasons (e.g., trauma, COPD exacerbation, heart failure, other infections such as bacterial pneumonia, poisonings) were excluded. The patients receiving immunosuppressive treatment (e.g., chemotherapy, high-dose steroids) were also excluded as the course of COVID-19 may be different in this group and these treatments may increase the risk of intubation. Descriptive information about the clinical conditions of the intubated patients is given in Supplementary Table S2.

2.1.2. ATR-FTIR spectroscopy measurements and data analysis

Frozen serum samples were thawed at room temperature and vortexed before the spectroscopic examination. Spectra were collected by the ATR-FTIR spectrometer (Frontier FTIR Spectrometer, PerkinElmer, USA) equipped with an ATR unit (MIRacle, PIKE, USA) containing Zn/Se crystal. Serum samples were placed in a volume of 1 μl on the Zn/Se crystal of the ATR unit without any pretreatment and dried with inert nitrogen gas (N_2) for 2 min to remove unbound water. Spectra were obtained with Spectrum One (PerkinElmer, USA) software with 4 cm^{-1} resolution and 32 scan counts in the spectral range of 4000-650 cm^{-1} [17,18]. After obtaining the background spectrum of the ambient air before each measurement, three separate measurements were made with the ATR-FTIR spectrometer for each serum sample. The averages of these spectra were used in all further analyses. The average spectra of each sample were baseline-corrected by the Rubberband correction method with 64 base points in OPUS 5.5 (Bruker, USA) software. Then, the qualitative and quantitative analyses of the spectral band parameters were assessed by calculating the area and position of bands specific to serum biomolecules using the band integration method B in OPUS 5.5 (Bruker, USA) software [19].

2.2. Biochemical studies

Serum levels of albumin and total protein were spectrophotometrically determined using the BCG (Bromocresol Green) and Biuret assays, respectively, according to the manufacturer's instructions. Globulin was calculated by subtracting albumin from total protein. Then, albumin concentrations were divided into globulins to get the albumin-to-globulin ratio (AGR) index (i.e., serum albumin/total protein - albumin).

2.2.1. Univariate statistics

To determine the statistical significance of the differences between the spectrochemical parameters, an unpaired *t*-test was employed in GraphPad Prism 10.2.1 (GraphPad Software, LLC, USA) statistical analysis program. The degree of significance was always set at a 95 % confidence interval and denoted as less than or equal to $P < 0.05$ *, $P <$

0.01 **, P < 0.001 ***, and P < 0.0001 ****. The results were expressed as means ± standard error of the mean.

The ROC (Receiver Operating Characteristic) curve analysis was applied to evaluate the overall prediction performance of each spectral parameter. Using GraphPad Prism 10.2.1 software, in which the area under the curve (AUC) and P values were computed at a 95 % confidence interval. The % points of sensitivity (True Positive Rate) and specificity (False Positive Rate) were plotted on an ROC curve graph as the threshold varies.

2.3. Explainable artificial intelligence methods

The research was conducted using Python version 3.11. The libraries employed included Matplotlib version 3.7.1, pip version 22.0.4, Sklearn version 1.2.2, Pandas version 2.0.1, RDkit version 2023.3.1, Shap version 0.41.0, eli5 version 0.13.0, scikit-plot version 0.3.7, and NumPy version 1.24.3. The computational operations were executed on a system equipped with an Intel® Core™ i5-8300H CPU at 2.30 GHz, a 64-bit operating system with an x64-based processor, and 32 GB of RAM. This configuration was meticulously selected to ensure high performance and reliability during the data processing and analysis phases of the research.

2.3.1. Data preprocessing

Data preprocessing is a critical phase in the model development process and has been meticulously applied in this study to enhance the quality and accuracy of the data used. Initially, missing data and outliers were identified and addressed using appropriate statistical methods. Missing data were imputed using multiple imputation techniques, and outliers were detected and managed using the IQR (Interquartile Range) method [20]. Additionally, data normalization and standardization were performed. Normalization was done using min-max scaling, and standardization was applied to ensure the data followed a Gaussian distribution. Considering the imbalance in our dataset, random undersampling was implemented to eliminate class imbalance. This method ensured equal representation of each class by randomly removing samples from overrepresented classes. These steps are crucial for improving the accuracy and generalizability of our model. We considered that gender differences could influence the serum molecular composition of the participants in our study. Furthermore, age and comorbidities such as diabetes, hypertension, and cardiovascular diseases were considered in our analysis. Balancing procedures were carried out among the classes to ensure equal representation for these factors. To ensure equal representation of males and females, balancing procedures were carried out among the classes. These balancing procedures involved organizing the dataset to provide equal representation for each class and were supported by bootstrap techniques. Thus, any biases in our model predictions due to gender differences were minimized.

2.3.2. Patient-based models

In the patient-based models, we utilized various demographic and clinical variables. These variables are listed in detail in Supplementary Table 1. Additionally, the specific 'S' variables used in the analysis are defined and described in the same table.

2.3.3. Model development

In the model development phase of the study, various algorithms were employed to manage individual variability effectively and to ensure high accuracy in classification tasks. The algorithms utilized included the Random Forest Classifier (RFC), XGBClassifier, Support Vector Machine (SVM) Classifier, and Decision Tree Classifier (DTC). These were specifically selected and fine-tuned to address the unique challenges and characteristics of the dataset. The RFC was particularly effective in classifying complex patterns due to its high accuracy and capability to handle large datasets with multiple variables [21]. The XGBClassifier, with its gradient boosting framework, offered a balance between speed and performance [22]. The SVM Classifier provided high precision, which is ideal for datasets with a clear margin of separation [23]. The DTC, meanwhile, offered a simple and interpretable model structure, beneficial for initial insights into the data structure [24,25]. The data was split into an 80 % training set and a 20 % test set to evaluate the performance of our models. The training set was used during the model learning phase, while the test set was used to assess the accuracy and generalizability of the models independently. The test set consisted of samples that the model had not seen before, selected randomly and balanced to reflect real-world performance better. This diverse array of models played a crucial role in developing a robust predictive framework that accommodates individual differences and delivers high accuracy in classification tasks. The performance metrics, such as the *F1-Score, Recall,* and *Precision* Scores, *Specifity* underscore the effectiveness of the models in both training and unseen test data, demonstrating the strong generalizability and reliability of the developed models.

2.3.4. Permutation Feature Importance

Permutation Feature Importance (PFI) is a technique employed to ascertain the impact of individual features on the predictive performance of a model [26]. This method involves the calculation of the importance score for each feature by permuting its values and observing the resultant variation in model accuracy. Mathematically, the importance (I) of a feature (f) can be expressed as *formula 1*:

$$I(f) = \frac{1}{N}\sum_{i=1}^{N}\left(Acc_{\text{original}} - Acc_{\text{permuted}_i}\right) \tag{1}$$

where (N) is the number of permutations, (Accoriginal) is the accuracy of the model with the original dataset, and (Acc permuted$_i$) is the accuracy of the model with the (i) permutation of feature (f). A higher value of (I(f)) signifies a greater impact of feature (f) on model performance.

In the context of this study, PFI was utilized to rank the features according to their importance in the best-performance model algorithm. The permutation process involved systematically shuffling the values of each feature across the data points and measuring the decrease in model accuracy. The relationship between the change in performance (Delta P) and the importance score (I) is given by *formula 2*:

$$\Delta P = -\sum_{f\in F} I(f) \tag{2}$$

where (F) is the set of all features. The more significant the decrease in performance (\Delta P), the higher the cumulative importance of the permuted features, indicating their critical role in the model's predictive ability.

2.3.5. Explainable artificial intelligence (XAI)

In the realm of machine learning, Explainable AI (XAI) has emerged as a pivotal approach to elucidate model predictions. Central to this approach is the SHAP (Shapley Additive Explanations) method, which quantifies the contribution of each feature to the model's output. Grounded in game theory, SHAP leverages the Shapley values to establish a feature importance ranking, thereby offering a transparent view on the model's decision-making process [27–30]. Mathematically, the contribution (\phi_i) of a feature (i) is determined by the Shapley value, which is the average marginal contribution of a feature across all possible coalitions. The formula is given by *formula 3*:

$$\phi_i(v) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}\left(v(S\cup\{i\}) - v(S)\right) \tag{3}$$

where (N) is the set of all features, (S) is a subset of features excluding (i), (v(S)) is the prediction model's output when only the features in (S)

are used, and (|S|) is the number of features in (S).

In this investigation, we employed TreeSHAP, a variant of SHAP designed for tree-based models. TreeSHAP capitalises on the inherent structure of decision trees and spectral wavenumbers to compute feature contributions with heightened efficiency. This method showed us the significance of sample-based and spectral wavenumber-based features in our classification tasks [31,32]. Consequently, we identified which samples and spectral wavenumbers—and by extension, band positions—hold paramount importance in the classification schema. The SHAP methodology endows researchers with an invaluable instrument to dissect their models' internal mechanics and rigorously assess individual features' influence on the predictions. This granular understanding is instrumental in refining model performance and ensuring the reliability of its outcomes.

3. Results

3.1. Spectrochemical findings in important molecular features

Supplementary Fig. S1 depicts the average and baseline-corrected spectra of serum from critically ill (intubated and serviced groups) and recovered patients in which the selected spectral bands for which the band area values and band positions were calculated are labelled, and the assignments of these bands are given in Supplementary Table S3. Fig. S1a demonstrates the changes at the 3800-2800 cm^{-1} spectral window. On the other hand, the provided fingerprint spectral window between 1800 and 650 cm^{-1} contains bands expressing changes in functional groups of proteins, carbohydrates, lipids, and nucleic acids (Fig. S1b, Table S3).

As per the Lambert-Beer law, the intensity or area of the changing absorption bands in infrared spectroscopy is directly proportional to the concentration of the relevant molecule. Consequently, several spectrochemical parameters were computed, including lipid, protein, and nucleic acid amounts and structures, phosphorylation of proteins, IgG glycosylation, and glucose concentrations. These parameters provide valuable insights into the composition and behavior of molecules, aiding researchers in understanding complex biological systems and chemical interactions. Supplementary Fig. S2 illustrates the quantitative changes in various lipids-associated spectrochemical parameters. Elevated serum lipids in intubated patients (IC group) are reflected in the increases in lipid/protein index (Fig. S2a). The bands at ~2920 cm^{-1} (CH_2 antisymmetric stretching) and ~2852 cm^{-1} (CH_2 symmetric stretching) belonging to aliphatic CH_2 groups originate from the long hydrocarbon chains in lipids [12]. The quantification of integral band areas of the CH_2 antisymmetric and symmetric bands indicates enhanced lipid content in IC patients' sera (Figs. S2b–c). Modulations in lipid metabolism are also reflected in the increases in saturated lipids and cholesterol esters (C=O groups) (Figs. S2d–e). Acyl chain length of fatty acids is an important parameter associated with the overall metabolic activity of the cell membrane. A higher acyl chain length value indicates the existence of longer-chain fats, while a lower value suggests the presence of shorter-chain and/or more branched lipids [12,33]. Increased acyl chain length is calculated for IC patients compared with other COVID-19 patient groups (Fig. S2f).

Different protein-related spectrochemical parameters are shown in Supplementary Fig. S3. The calculations of sub-structural entities of proteins emerging from various amide bands and CH_3 symmetric stretching band [16] demonstrated a significant diminish in proteins of the IC group, except for the amide A band (Figs. S3a–e). The amide I band arises from the C=O stretching of proteins, and the amide II and amide A bands emerge from N–H bending and N–H stretching, respectively. On the other hand, the amide III band reflects the both C–N stretching and N–H bending of proteins (Table S3). The ratio of well-known protein bands that is amide I/amide I + amide II index expresses the protein concentration in biological systems [12]. Moreover, the amide I/amide II ratio is an index for protein conformation [34]. The results demonstrated that concentrations and conformational changes of proteins are respectively higher and stronger in the IC group compared with the PC group, and vice versa when compared to the SC group (Figs. S3f–g). The enhanced quantity of phosphorylated proteins was depicted for the IC group compared with other patient groups (Fig. S3h).

Albumins and globulins are major components of the blood. The modulations in nonspecific immune responses, specifically a shift in the serum AGR index, have been identified as correlating with disease presence and outcomes [35,36]. Therefore, this study employed a serum AGR index (calculated as serum albumin/total protein - albumin), for the evaluation of the immune responses and COVID-19 progression among IC, SC, and PC patient groups. The results revealed a significant shift in the AGR index among the patient groups (Fig. 1a). This index decreased in the IC group compared with the PC group (IC group/AUC = 0.765). However, the lowest value was calculated for the SC group, and vice-versa for the PC group (SC group/AUC = 0.893). Hence, the AGR index can be considered a fast biochemical parameter during patient care and hospitalization of COVID-19 patients.

In their study of IgG structural features in serum samples from COVID-19 patients, Bandeira et al. emphasized the significance of the 1702-1785 cm^{-1} spectral range as a key indicator of IgG glycosylation levels. Furthermore, they proposed that this spectral window could be used to explore distinct subpopulations based on the severity of COVID-19 cases [37]. According to our findings, the integral area of the 1783 cm^{-1} band assigned to IgG glycosylation (C=O groups in IgG) decreased as disease severity increased in intubated patients (IC group/AUC = 0.715) compared with the PC group (Fig. 1b). Glucose is a vital molecule for sugar metabolism. In the IC group, glucose concentration was higher than in other patient groups with a 0.977 AUC value (Fig. 1c). Therefore, the indices for IgG glycosylation level and glucose can be considered in the early prediction of COVID-19 severity.

Vibrations in PO_2 functional groups are appeared by phosphodiester groups of nucleic acids or phospholipids in serum [12]. In addition, it has been reported that two main phospholipids (sphingolipids and lysolecithin) in human serum increase with the severity of COVID-19 [14]. The bands at 1235 cm^{-1} and 1074 cm^{-1} were defined as antisymmetric and symmetric stretching vibrations of the PO_2^- functional groups, respectively. Our results indicated higher nucleic acid concentrations in the IC group with 0.956 AUC value (Fig. 2a). Moreover, the C–H deformation band at 720 cm^{-1} emerges at high rates in the IC group with 0.898 AUC value (Fig. 2b). The critical changes were also seen in the yet unassigned bands. Especially, the distinctive and pronounced band at 659 cm^{-1} in the IC group (AUC = 0.884) suggests that this band warrants special attention as a potential biomarker (Fig. 2c).

3.2. Explainable artificial intelligence findings

3.2.1. Model development and validation

In the patient-based analyses, we initially applied Permutation Feature Importance (PFI) to identify the top 50 features. Subsequently, we used SHAP (Shapley Additive exPlanations) to evaluate the contribution of these individual patient features to the classification of post-COVID, in-service, or intubated conditions. This allowed us to understand how disease states differ based on individual characteristics. In wavelength-based analyses, SHAP was used to examine the contribution of specific wavelengths to the classification model, helping us determine which wavelengths were more effective in distinguishing disease states at the molecular level. These two analytical approaches enabled us to analyze the performance of our model comprehensively and demonstrate the impact of different types of data (demographic and spectral) on classification success.

During the model development phase, the scikit-learn package was utilized to facilitate the construction and evaluation of the classification models. The classification reports, derived from the package's metrics, provided a comprehensive assessment of model performance (Fig. 3 and

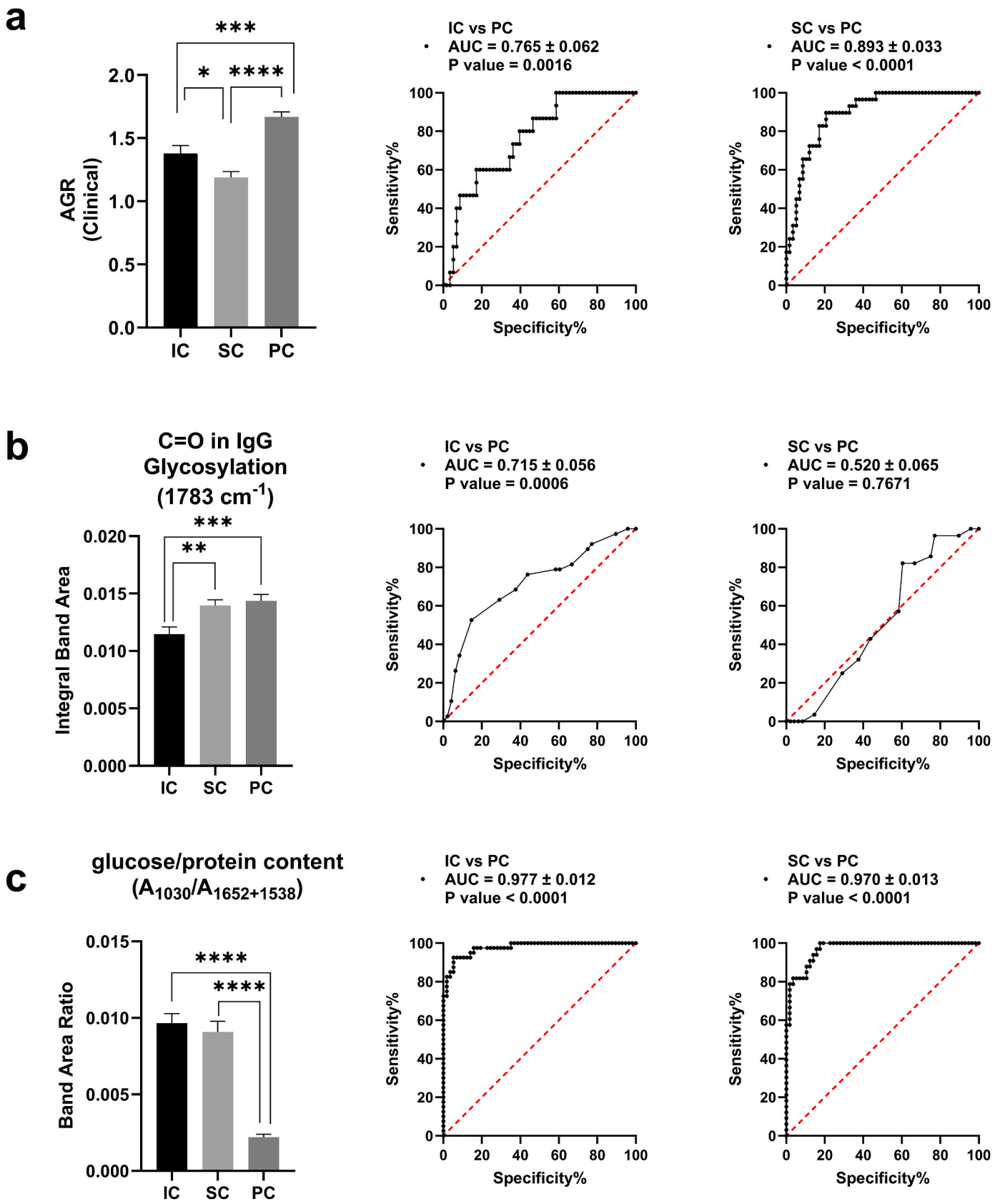


Fig. 1. The changes in molecular features for **(a)** Albumin to globulin (AGR) index, **(b)** IgG glycosylation (C=O/1783 cm^{-1}), and **(c)** glucose/protein index (A_{1030}/$A_{1652+1538}$) with corresponding ROC curves. The degree of significance is shown as * $p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

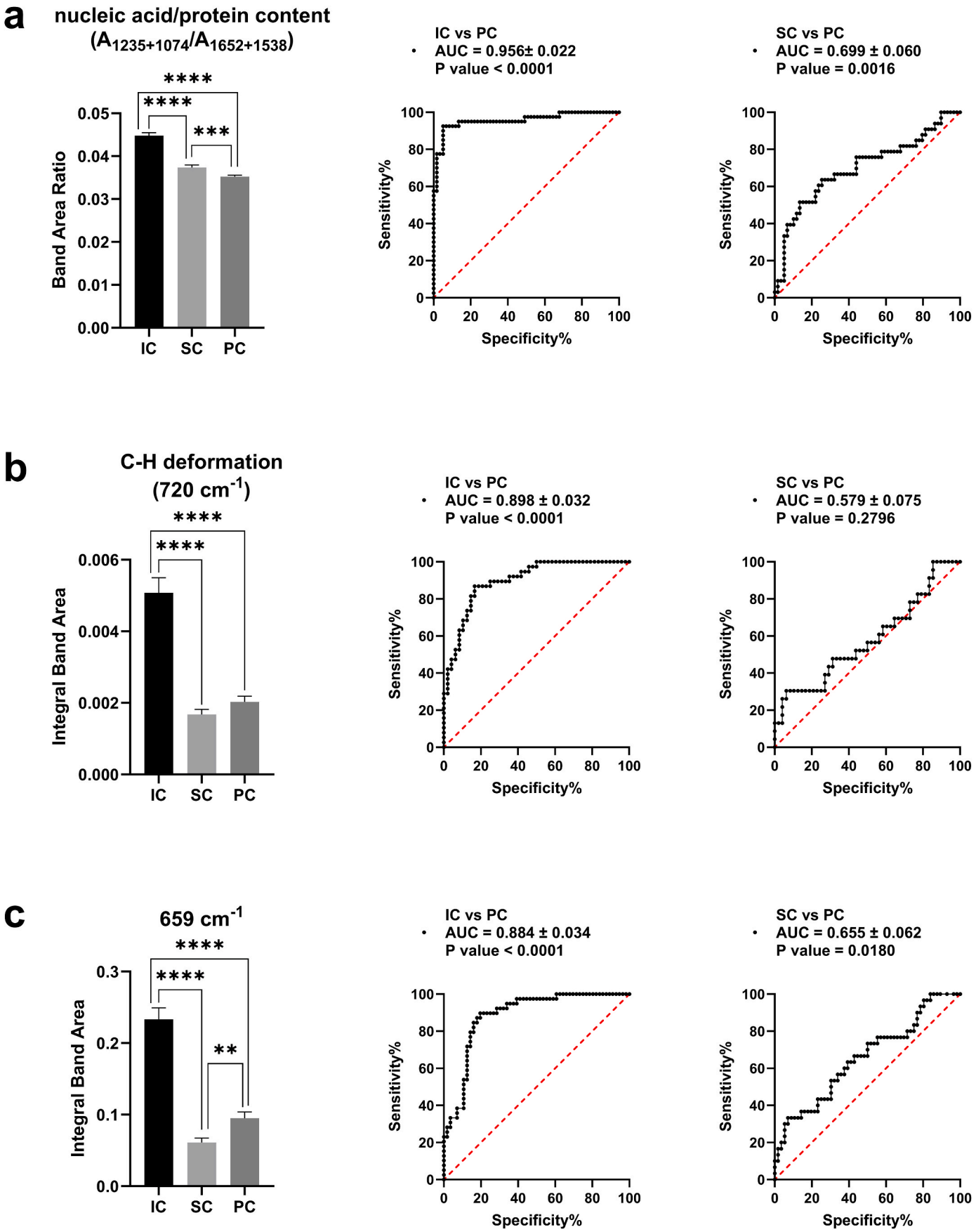


Fig. 2. Spectrochemical changes with corresponding ROC curves in molecular features for **(a)** nucleic acid/protein index ($A_{1235+1074}/A_{1652+1538}$), **(b)** C–H deformation (720 cm^{-1}), and **(c)** unassigned band (659 cm^{-1}). The degree of significance is shown as **p ≤ 0.01, ***p ≤ 0.001, ****p ≤ 0.0001.

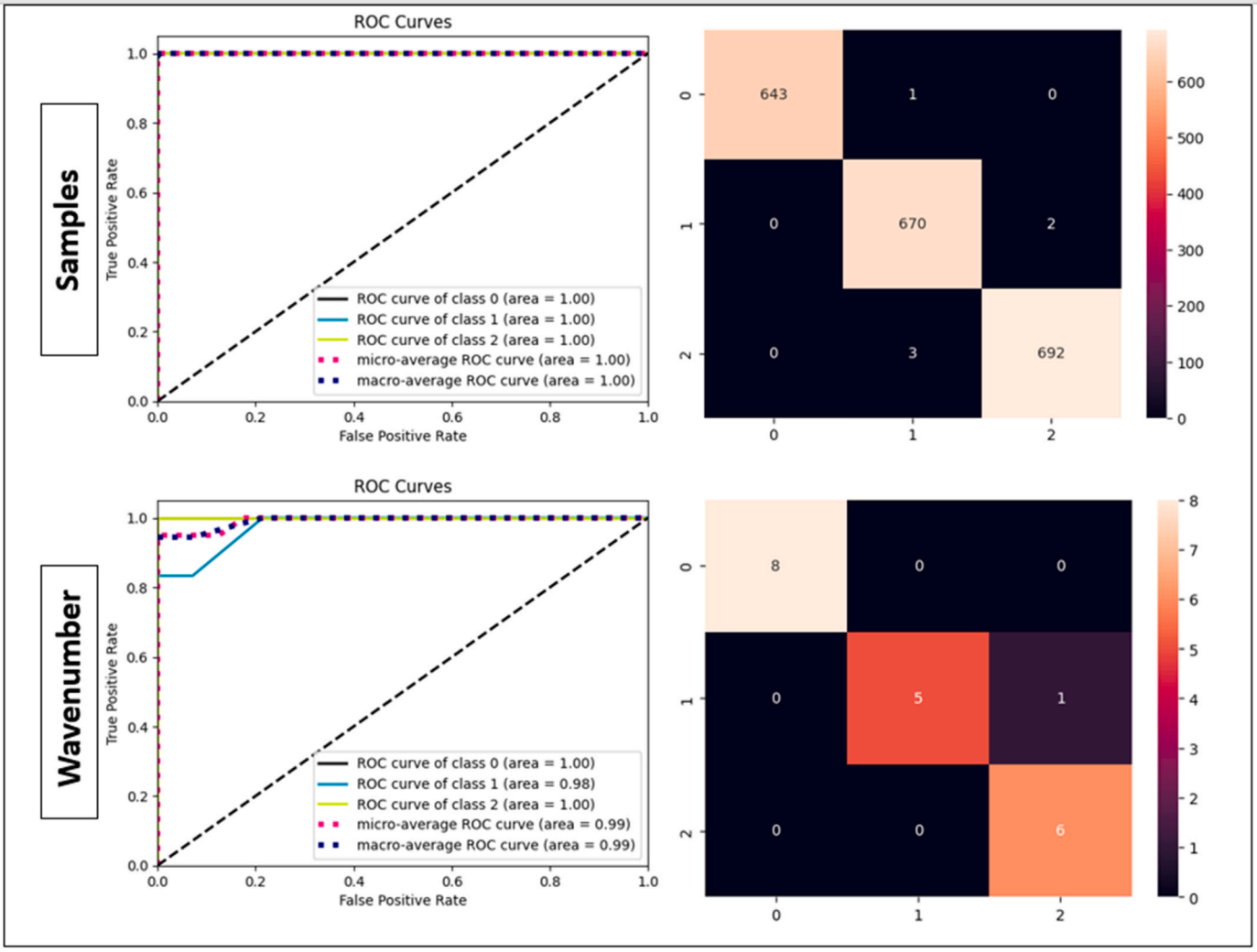


Fig. 3. ROC/AUC curve and confusion matrix of best performed Random Forest Classifier (RFC)model algorithm.

Table 1
Combined classification model performance.

Models	Train Accuracy	Test Accuracy	F1-Score	Recall Score	Precision Score	Specificity	Model Type
Random Forest Classifier (RFC)	1.0	0.996	0.996	0.996	0.996	0.996	Patient-Based
XGBClassifier	0.982	0.965	0.965	0.965	0.965	0.965	Patient-Based
SVM Classifier (SVM)	0.999	0.996	0.996	0.996	0.996	0.996	Patient-Based
Decision Tree Classifier (DTC)	1.0	0.980	0.980	0.980	0.980	0.98	Patient-Based
Random Forest Classifier (RFC)	1.0	0.95	0.95	0.95	0.95	0.95	Wavelength-Based
XGBClassifier	1	0.9	0.9	0.9	0.9	0.9	Wavelength-Based
SVM Classifier (SVM)	0.569	0.55	0.55	0.55	0.55	0.55	Wavelength-Based
Decision Tree Classifier (DTC)	1.0	0.9	0.95	0.95	0.95	0.95	Wavelength-Based

Table 1).

[In **patient-based** analyses, SHAP was used to evaluate the contribution of individual patient features to the classification of post-COVID, in-service, or intubated conditions. This allowed us to understand how disease states differ based on individual characteristics. In **wavelength-based** analyses, the contribution of specific wavelengths to the classification model was examined using SHAP. This helped us determine which wavelengths were more effective in distinguishing disease states at the molecular level. These two analytical approaches enabled us to analyze the performance of our model comprehensively and demonstrate the impact of different types of data (demographic and spectral) on classification success.]

3.2.2. Permutation Feature Importance (PFI)

In the evaluation of the RFC model's efficacy in classification, the PFI analysis was instrumental. The analysis illuminated that features S.13, S.16, S.9, S.11, S.6, S.12, and S.3 were paramount, with S.13 (PFI score: 0.120) exerting the most substantial influence on the model's output. Similarly, features S.16 and S.28 (PFI score: 0.117) were identified as critical determinants of performance. Conversely, feature S.17 demonstrated a minimal impact (PFI score: 0.009). Furthermore, the wavenumbers 1282 cm^{-1} and 1285 cm^{-1}, with PFI scores of 0.050, were recognized as significantly affecting the model's accuracy for COVID-19 prediction. Wavenumbers 2377 cm^{-1}, 1618 cm^{-1}, 2315 cm^{-1}, and 3710 cm^{-1}, despite bearing lower importance scores, were also highlighted as noteworthy contributors to the model's data analysis,

underscoring their potential utility in the identification and classification of COVID-19-related features (Fig. 4).

3.2.3. Shapley Additive exPlanations

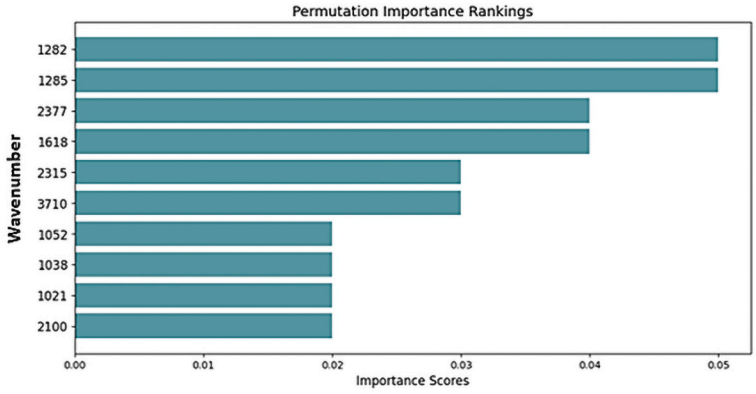
In the advanced analytical phase of our study, we applied SHAP to the top 50 features identified by PFI scores to gain deeper insights into the model's predictive behavior. The SHAP analysis, which explicates the contribution of each feature to the model's output, revealed distinct patterns of feature importance across different patient categories. In the post-COVID analysis based on individual data, features S.13, S.28, S.2, S.16, S.12, S.9, S.27, S.20, and S.3 were found to be highly influential. Features S.11, S.30, and S.6 stood out for patients receiving in-hospital services, while for those intubated, features S.31, S.24, S.17, and S.20 were prominent. In the wavelength-based analysis, the wavenumbers 2377 cm^{-1}, 3710 cm^{-1}, 2315 cm^{-1}, and 1282 cm^{-1} were significant in the post-COVID context, whereas 1285 cm^{-1} and 2315 cm^{-1} were crucial for in-hospital service, and 1021 cm^{-1}, 1038 cm^{-1}, and 1052 cm^{-1} were key for intubated patients. These findings underscore the heterogeneity of COVID-19's impact and highlight the potential of SHAP in unraveling complex feature interactions, thereby providing a clearer understanding of the factors driving model predictions in the classification of COVID-19 (Fig. 5).

4. Discussion

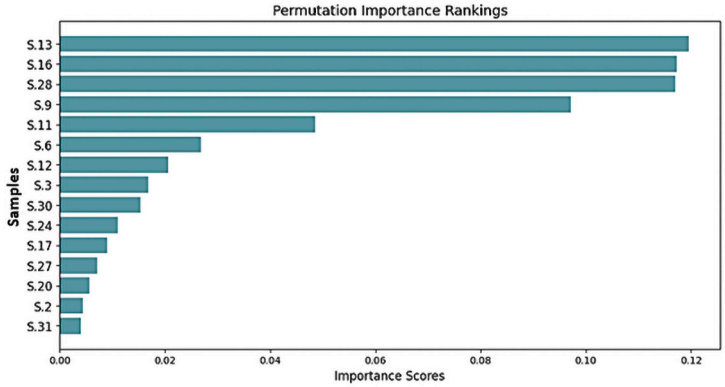
A recent COVID-19 study demonstrated a slight shift and decrease in spectral absorbances attributed to the amide I and amide II regions, indicating a potential decrease in protein production [38]. Additionally, a decrease in host protein levels was observed, indicating a consistent pattern of protein translation inhibition seen in other viral infections. In line with our results, there was a rise in phosphorylated proteins coupled with a reduction in protein abundance, alongside hyperphosphorylation observed in the CK2 and p38 MAPK pathways associated with cytokine production. These shifts in phosphorylation signify changes in the activities of crucial cellular processes co-opted during the infection [39].

Consistent with our findings, the studies dealing with COVID-19 infection demonstrate enhanced accumulation of nucleic acids in disease-positive biofluids such as serum and saliva [5,38,40], possibly associated with the generalized inflammatory response correlated with disease severity and an increase in cell-free DNA (cfDNA) in the bloodstream. Recent findings showed a rise in serum fasting glucose and blood HbA1c, both of which significantly decreased following complete remission, returning to levels observed before the onset of COVID-19 infection [41]. In the case of lipids, it has been documented that severe COVID-19 patients exhibit increased levels of ceramides in their serum. Moreover, the accumulation of ceramides featuring longer chains has been linked to adverse outcomes, albeit not correlated with conventional plasma lipid markers indicative of cardiovascular risk factors. The alterations in the structure of lipid molecules during the process of lipid biogenesis pathways may impact the entry of viruses through receptor-mediated mechanisms on the cell surface of endosomes and regulate the spread of viruses [42]. Variations in the CH_2/CH_3 ratio could indicate an increase in intermolecular chain disorder, with elevated concentrations of long CH_2 chains potentially weakening the integrity of the cell membrane-skeleton [43].

Serum proteins mainly albumin and globulins, play crucial roles in systemic inflammation events. Therefore, the albumin-to-globulin ratio (AGR) has been used to predict the severity and mortality of infectious diseases. A low AGR ratio has been associated with many diseases including COVID-19 [36,44]. Recent reports indicate that severe COVID-19 patients had lower AGR values than non-severe COVID-19 patients. Moreover, non-survivor patients with COVID-19 had lower AGR values than survivor patients pointing out the predictive value of the AGR for COVID-19 [36]. Albumin has been investigated as crucial in the clinical progression of COVID-19. It was reported that hypoalbuminemia, which is linked to an inflammatory response during



Wavelength	Importance Score
1282	0.050
1285	0.050
2377	0.040
1618	0.040
2315	0.030
3710	0.030
1052	0.020
1038	0.020
1021	0.020
2100	0.020



Samples	Importance Score
S.13	0.120
S.16	0.117
S.28	0.117
S.9	0.097
S.11	0.049
S.6	0.027
S.12	0.021
S.3	0.017
S.30	0.017
S.24	0.011
S.17	0.009
S.27	0.007
S.20	0.006
S.2	0.004
S.31	0.004

Fig. 4. Permutation Feature importance values and bar plot in samples and wavenumber classifications.

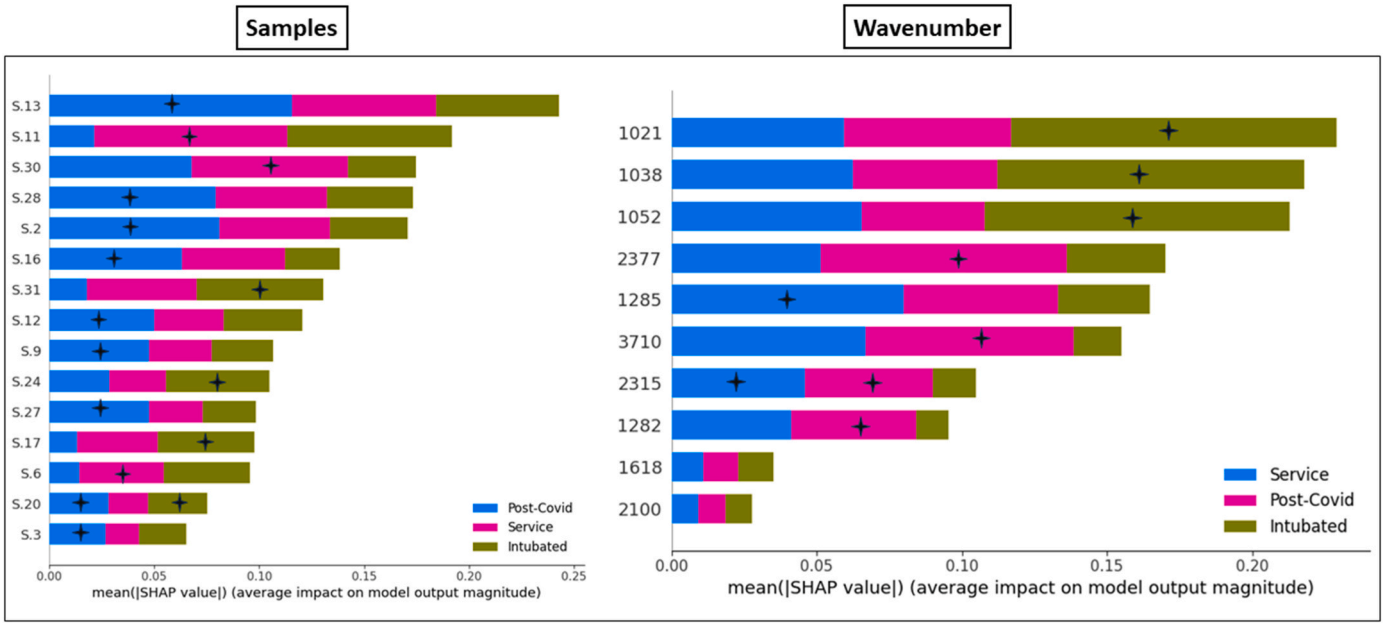


Fig. 5. The summary plot, the importance of which was determined according to the SHapley values formed as a result of explainable machine learning in the samples and wavenumber classification.

critical illness serves as a predictor of mortality. This is caused by the release of cytokines and chemokines, which induce an increase in capillary leakage, thereby altering the distribution of albumin between intravascular and extravascular compartments [38,45].

The attention on IgG glycosylation has become prominent, likely due to the rising use of antibodies for therapeutic purposes. Changes in galactosylation, fucosylation, and sialylation are acknowledged as pivotal factors affecting the diverse functions of IgG, ranging from inhibitory and anti-inflammatory effects to the activation of complement and the facilitation of antibody-dependent cellular cytotoxicity [46]. The Fragment, crystallizable (Fc) region of IgG in healthy individuals is predominantly core fucosylated, with over 90 % exhibiting this modification. It is increasingly recognized that antibody glycosylation also plays a significant role in responding to infectious diseases. Recent evidence from various viral and bacterial infections suggests notable alterations in antibody glycosylation, serving not only as disease markers but potentially influencing disease progression as well [47,48]. In healthy adults, about 35 % of circulating IgG Fc-glycans consist of agalactosylated structures, with monogalactosylated and digalactosylated glycans each making up roughly 35 % and 15 %, respectively. However, individuals with active autoimmune or inflammatory conditions tend to exhibit a shift towards increased agalactosylation, which is commonly associated with proinflammatory responses. Despite this association, the exact relationship between glycan alterations and disease causality remains unclear [47].

A study comparing IgG Fc glycosylation among healthy individuals, those with HBV-related cirrhosis, and individuals with chronic HBV infection revealed decreased Fc galactosylation in the HBV-exposed groups compared to controls, with lower galactosylation levels correlating positively with fibrosis severity [49]. Chronic infections such as HBV, tuberculosis, and HIV often present autoimmune-like glycan profiles, characterized by heightened agalactosylation and asialylation levels. These findings suggest that chronic inflammation across various diseases, infectious or otherwise, may lead to similar agalactosylated IgG Fc glycosylation patterns. Prolonged pathogen exposure and immune activation likely perpetuate this inflammation, resulting in elevated levels of inflamed agalactosylated IgG. Consequently, in both autoimmune conditions and certain infectious diseases, the skew towards agalactosylated and asialylated IgG likely reflects shared inflammatory pathways, thus serving as a potential marker of immune-mediated inflammation [49].

Recent research indicates that particular proinflammatory antibody types, characterized by IgG3 and IgG1 with F0N0 glycoform modification, are heightened in a larger proportion of patients experiencing severe COVID-19 symptoms compared to those with milder manifestations and seropositive children. Notably, there's a higher prevalence of IgG1 with afucosylated Fc glycans observed during severe cases of the disease. Moreover, it has been observed that males, in particular, exhibit elevated Fc afucosylation levels in severe COVID-19 cases. These findings suggest a potential link between immune-complex-mediated activation of inflammatory FcγR pathways, cytokine production, and the progression to severe COVID-19. The results also underscore the increased occurrence of proinflammatory IgG antibodies in severe COVID-19 cases and propose that future longitudinal investigations, incorporating pre-infection sample analyses, are necessary to ascertain whether these Fc structures could serve as pre-infection biomarkers indicative of the risk of severe COVID-19 progression [50]. In agreement with the findings of our study (Fig. 1b), Bandeira et al. also suggested a diminish in IgG glycosylation for severe COVID-19 cases using the FTIR spectroscopy approach enabling the exploration of distinct sub-populations among sera of COVID-19 patients [37]. Very recently, reduced IgG glycosylation in the saliva of COVID-19 patients was also reported, offering FTIR spectroscopy-based 2T2D-COS analysis as a valuable marker in monitoring the COVID-19 severity [51,52].

In this study, explainable machine learning algorithms were applied to the classification of critically ill (intubated and serviced groups) and recovered COVID-19 patient groups using IR spectral raw data. The transparent RFC algorithm was developed to classify COVID-19 disease and assess its performance. The model achieved a training accuracy of 100 % on the training dataset, which could suggest overfitting. However, the model's efficacy on the test dataset was evaluated using test accuracy, precision, sensitivity, and F1 score metrics. It accurately classified 99.6 % and 95 % of the samples in the test dataset. Sensitivity values were determined to be 1.00 and 0.96, and specificity values were found to be 1.00 and 0.95 (sample-based and wavenumber-based, respectively). F1 scores were also calculated as 1.00 and 0.95. The model demonstrated exceptional performance in sensitivity metrics. The selection of the RFC for this study stems from its intrinsic advantages

pertinent to the task. RFC is renowned for its robustness and capability to manage high-dimensional data, such as the IR raw data utilized here. In contrast to SVM or other algorithms that might necessitate meticulous parameter tuning and are prone to overfitting, RFC can yield effective results with minimal hyperparameter optimization. Furthermore, RFC offers feature importance scores, which are essential for explainability in medical diagnostics [53,54].

The concern of overfitting, indicated by perfect training accuracy, is alleviated by the model's substantial test dataset performance. The high test accuracy, precision, sensitivity, and F1 score suggest that the model generalizes well to new data. This is corroborated by the near-perfect sensitivity values for both sample-based and wavenumber-based analyses, affirming the model's aptitude in correctly identifying positive cases. The minor discrepancy between training and testing performance is an anticipated outcome due to the model's exposure to more variability in the test data, reflecting its application in real-world conditions. To ascertain that overfitting is not an issue, the model's performance underwent rigorous evaluation using diverse metrics, and the outcomes were consistent and reliable. The exemplary sensitivity metrics highlight the model's capacity to identify COVID-19 cases accurately, which is of paramount importance in medical settings.

Additionally, our study compared the performance of models based on demographic and clinical data with those based on spectral data. It was observed that demographic and clinical data-based models performed as well as or better than spectral data-based models in some cases. This finding is significant as it highlights the potential of using readily available demographic and clinical information for accurate COVID-19 patient classification. Demographic and clinical models provide practical advantages, such as ease of data collection and the potential for immediate application in clinical settings. However, the integration of spectral data can enhance model precision by incorporating molecular-level information, which is crucial for understanding disease mechanisms.

The SHAP results indicate a notable trend in the demographics of post-COVID patients, with a majority being male and over the age of 40. This pattern is consistent with the hospital service data, where patients are predominantly male and typically over the age of 45. Furthermore, the data on intubated patients reveals that individuals over the age of 50 are more likely to undergo this procedure (Table S1). These findings suggest a gender and age-related vulnerability, with older male patients exhibiting a higher risk profile for severe post-COVID complications. The reasons behind this trend could be multifaceted, involving biological, social, and behavioral factors contributing to the disease's susceptibility and progression. It is imperative to delve deeper into these factors to understand the underlying causes and develop targeted interventions to mitigate the risks associated with these demographic groups. The SHAP analysis provides a valuable framework for identifying and prioritizing such risk factors, thereby informing demographical decisions and public health strategies aimed at reducing the burden of COVID-19. Pointing to gender bias involving COVID-19 studies, previous studies have highlighted that gender differences in the prevalence and outcomes of infectious diseases occur at all ages, with men generally having a higher burden of bacterial, viral, fungal, and parasitic infections [55–57]. In addition, previous studies have reported that the male gender is a risk factor for several general adverse outcomes related to COVID-19 [56]. Genetic predispositions, sex hormones, immune system responses, and non-biological causes contribute to the disparity in COVID-19 responses between the sexes [58]. The reasons behind these differences still need to be fully understood. The Subject 11 (S.11) is an essential example of the service status. For this example, which is 68 years old and has male characteristics, it is seen that advanced age and male status significantly affect hospitalization status. Understanding how COVID-19 affects men and women differently may provide clues to understanding the disease pathophysiology that could lead to successful interventions.

XAI/SHAP analysis has been utilized to classify different health states of COVID-19 patients, and it has been observed that certain IR spectral wavelengths significantly contribute to the model predictions. Notably, the wavenumbers at 2377 cm $^{-1}$, 3710 $^{-1}$, 2315 cm $^{-1}$, and 1282 cm $^{-1}$ are significant in post-COVID conditions. For patients in hospital services, the wavenumbers at 1285 cm $^{-1}$ and 2315 cm $^{-1}$ have been determined to be important, while for intubated patients, the wavenumbers at 1021 cm^{-1}, 1038 cm$^{-1,}$ and 1052 cm^{-1} have been identified as significant (Fig. 5). The wavenumber of 3710 cm^{-1} is generally associated with hydroxyl groups (OH), which is particularly important in the spectroscopic analysis of water or other molecules containing hydroxyl groups [59]. The absorption at this wavelength can indicate the presence of molecular water and the different coordination states of hydroxyl groups. Additionally, it may reflect conditions such as the hydration state of cells or the presence of water in the cellular environment. The band at 2315 cm $^{-1}$ has been found significant in post-COVID and patient cases. This band is generally associated with Lewis acidity, which is particularly important in spectroscopic analyses that indicate the presence of carbonyl groups (C=O) [60]. Analyses such as C=O in IgG glycosylation (Figs. 1b and 5) and ester C=O stretching: cholesterol esters (Fig. S2e and Fig. 5), which contain carbonyl groups, can be associated with the wavenumber of 2315 cm^{-1}. The bands at 1282 cm^{-1} are associated with the vibrations of phosphate groups (P=O) and are typically crucial in analyses related to nucleic acid/-protein content [61]. Phosphate groups found in the backbone of nucleic acids and phosphorylated amino acids in proteins can exhibit absorption at this wavelength. Therefore, the 1282 cm^{-1} wavenumber band can be particularly associated with nucleic acid/protein content (Figs. 2a and 5) and protein phosphorylation (Fig. S3h and Fig. 5).

The carbohydrate group band, 1038 cm^{-1} (stretching vibrations of C–O), was the most important feature in the intubated group. In COVID-19 patients whose liver tissue is damaged, the ability of liver cells to use glucose to synthesize glycogen is reduced, which leads to worsening insulin resistance and elevation of blood glucose. A study by Wang et al. found that the influenza virus can induce the IFN regulatory factor 5 (IRF5) gene to bind to the O-GlcNAc transferase (OGT) enzyme [62]. The link shows that the influenza virus triggers K63-dependent ubiquitination of IRF5, leading to O-GlcNAsylation of IRF5. It has also been found that patients with high blood sugar levels are more vulnerable to virus attacks. This interaction suggests that patients with diabetes are at greater risk for viral infections. Cytokine storms are extreme inflammatory responses overactivated by the immune system and characterized by multiple cytokine releases. In this context, it is predicted that patients with diabetes may be more prone to cytokine storms and severe consequences from viral infections. Diabetes is a condition that can affect immune system responses and weaken defense mechanisms against infections. This link is an important step in triggering an extreme immune response known as a cytokine storm [62]. The patients with COVID-19 and diabetes were also reported to show significantly higher intubation rates [63]. Additionally, a study conducted on the FTIR spectra of saliva from COVID-19 patients found that those with low or undetectable viral loads showed a decrease in the saccharide (ribose) bands at 1038 cm^{-1} and 1074 cm^{-1} [64]. This decrease may indicate that the patients are recovering from a hyperglycemic state [65].

Our current study demonstrates the effectiveness of spectrochemical analyses and XAI results in identifying significant molecular markers for classifying different health states of COVID-19 patients. These results highlight the contribution of specific wavelengths to model predictions and demonstrate the potential of these methods. However, the scope of these findings is insufficient to showcase the broad potential of spectrochemical analyses and XAI methodology fully. Therefore, more comprehensive analyses and the application of various algorithms are required. Using different machine learning models and deep learning techniques can reveal complex relationships and patterns hidden in our datasets, allowing for a more detailed understanding of the disease's molecular mechanisms. Such an approach could play a critical role in combating the pandemic and understanding the disease better.

5. Conclusion

In conclusion, the findings of this study suggest that ATR-FTIR spectroscopy coupled with explainable AI approaches can be a valuable tool in assessing disease progression and identifying potential biomarkers in critically ill patients with COVID-19. The study revealed significant alterations in serum biomolecules among COVID-19 patients, particularly in intubated patients, compared to recovered patients. The observed changes included elevated serum lipids, altered protein structures, enhanced phosphorylation, decreased IgG glycosylation levels, and increased glucose and nucleic acid concentrations. The study also identified potential biomarkers for disease severity, including the serum AGR index and the 1783 cm^{-1} band assigned to IgG glycosylation. The bands at 720 cm^{-1} and 659 cm^{-1} also hold potential biomarker capacity for monitoring the disease progression. Furthermore, the PFI analysis and SHAP explanations highlighted the importance of specific spectral features and wavenumbers in differentiating between COVID-19 patient groups.

This study concluded that FTIR spectroscopy associated with XAI and serum fluid is an effective tool to distinguish COVID-19 intubated, serviced, and recovered patients. This explainable and interpretable model algorithm can potentially be used for molecular prediction bias and high-capacity and transparent screening in complex cases. The strengths of this study include rapid sample preparation, spectral data collection methodology, and the development of a multivariate explainable, interpretable, and transparent algorithm based on ATR-FTIR spectral data. Future studies should increase the sample size to improve prediction performance.

Funding

This work is supported by the Scientific Research Project Fund of Bilecik Şeyh Edebali University under project number 2023-01. BŞEÜ.25-01 to RG.

Compliance with ethical standards

The study was approved (or granted exemption) by the appropriate institutional and/or national research ethics committee (including the name of the ethics committee) and the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Informed consent was obtained from all patients for being included in the study.

CRediT authorship contribution statement

Gorkem Tokgoz: Methodology, Investigation, Formal analysis. **K. Kubra Kirboga:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Faik Ozel:** Resources, Methodology, Investigation, Data curation. **Serkan Yucepur:** Resources, Methodology, Investigation, Formal analysis, Data curation. **Isa Ardahanli:** Validation, Methodology, Investigation, Data curation. **Rafig Gurbanov:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing 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.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2024.126652.

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.

References

[1] WHO, WHO, 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen. (Accessed 1 March 2023).
[2] S. Durdagi, T. Avsar, M.D. Orhan, M. Serhatli, B.K. Balcioglu, H.U. Ozturk, A. Kayabolen, Y. Cetin, S. Aydinlik, T. Bagci-Onder, S. Tekin, H. Demirci, M. Guzel, A. Akdemir, S. Calis, L. Oktay, I. Tolu, Y.E. Butun, E. Erdemoglu, A. Olkan, N. Tokay, Ş. Işık, A. Ozcan, E. Acar, S. Buyukkilic, Y. Yumak, The neutralization effect of montelukast on SARS-CoV-2 is shown by multiscale in silico simulations and combined in vitro studies, Mol. Ther. 30 (2) (2022) 963–974, https://doi.org/10.1016/j.ymthe.2021.10.014.
[3] A. Talari, M. Garcia Martinez, Z. Movasaghi, S. Rehman, I. Rehman, Advances in fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 52 (2016), https://doi.org/10.1080/05704928.2016.1230863, 00-00.
[4] Z. Movasaghi, S. Rehman, D.I. ur Rehman, Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 43 (2) (2008) 134–179, https://doi.org/10.1080/05704920701829043.
[5] O. Calvo-Gomez, H. Calvo, L. Cedillo-Barrón, H. Vivanco-Cid, J.M. Alvarado-Orozco, D.A. Fernandez-Benavides, L. Arriaga-Pizano, E. Ferat-Osorio, J.C. Anda-Garay, C. López-Macias, M.G. López, Potential of ATR-FTIR-chemometrics in Covid-19: disease recognition, ACS Omega 7 (35) (2022) 30756–30767, https://doi.org/10.1021/acsomega.2c01374.
[6] B. Gallo Marin, G. Aghagoli, K. Lavine, L. Yang, E.J. Siff, S.S. Chiang, T.P. Salazar-Mather, L. Dumenco, M.C. Savaria, S.N. Aung, T. Flanigan, I.C. Michelow, Predictors of COVID-19 severity: a literature review, Rev. Med. Virol. 31 (1) (2021) 1–10, https://doi.org/10.1002/rmv.2146.
[7] Y.D. Gao, M. Ding, X. Dong, J.J. Zhang, A. Kursat Azkur, D. Azkur, H. Gan, Y. L. Sun, W. Fu, W. Li, H.L. Liang, Y.Y. Cao, Q. Yan, C. Cao, H.Y. Gao, M.C. Brüggen, W. van de Veen, M. Sokolowska, M. Akdis, C.A. Akdis, Risk factors for severe and critically ill COVID-19 patients: a review, Allergy 76 (2) (2021) 428–455, https://doi.org/10.1111/all.14657.
[8] L. Wang, Y. Zhang, D. Wang, X. Tong, T. Liu, S. Zhang, J. Huang, L. Zhang, L. Chen, H. Fan, M. Clarke, Artificial intelligence for COVID-19: a systematic review, Front. Med. 8 (2021) 704256, https://doi.org/10.3389/fmed.2021.704256.
[9] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[10] K.K. Kırboğa, S. Abbasi, E.U. Küçüksille, Explainability and white box in drug discovery, Chem. Biol. Drug Des. 102 (1) (2023) 217–233, https://doi.org/10.1111/cbdd.14262.
[11] K.K. Kırboğa, E.U. Küçüksille, M.E. Naldan, M. Işık, O. Gülcü, E. Aksakal, CVD22: explainable artificial intelligence determination of the relationship of troponin to D-Dimer, mortality, and CK-MB in COVID-19 patients, Comput. Methods Progr. Biomed. 233 (2023) 107492, https://doi.org/10.1016/j.cmpb.2023.107492.
[12] D. Yonar, M. Severcan, R. Gurbanov, A. Sandal, U. Yilmaz, S. Emri, F. Severcan, Rapid diagnosis of malignant pleural mesothelioma and its discrimination from lung cancer and benign exudative effusions using blood serum, Biochim. Biophys. Acta, Mol. Basis Dis. 1868 (10) (2022) 166473, https://doi.org/10.1016/j.bbadis.2022.166473.
[13] A. Koehler, M.L. Scroferneker, B.A.S. Pereira, N.M. Pereira de Souza, R. de Souza Cavalcante, R.P. Mendes, V.A. Corbellini, Using infrared spectroscopy of serum and chemometrics for diagnosis of paracoccidioidomycosis, J. Pharm. Biomed. Anal. 221 (2022) 115021, https://doi.org/10.1016/j.jpba.2022.115021.
[14] L. Zhang, M. Xiao, Y. Wang, S. Peng, Y. Chen, D. Zhang, D. Zhang, Y. Guo, X. Wang, H. Luo, Q. Zhou, Y. Xu, Fast screening and primary diagnosis of COVID-19 by ATR-FT-IR, Anal. Chem. 93 (4) (2021) 2191–2199, https://doi.org/10.1021/acs.analchem.0c04049.
[15] V.G. Barauna, M.N. Singh, L.L. Barbosa, W.D. Marcarini, P.F. Vassallo, J.G. Mill, R. Ribeiro-Rodrigues, L.C.G. Campos, P.H. Warnke, F.L. Martin, Ultrarapid on-site detection of SARS-CoV-2 infection using simple ATR-FTIR spectroscopy and an analysis algorithm: high sensitivity and specificity, Anal. Chem. 93 (5) (2021) 2950–2958, https://doi.org/10.1021/acs.analchem.0c04608.
[16] A. Dogan, R. Gurbanov, M. Severcan, F. Severcan, CoronaVac (Sinovac) COVID-19 vaccine-induced molecular changes in healthy human serum by infrared spectroscopy coupled with chemometrics, Turk. J. Biol. 45 (4) (2021) 549–558, https://doi.org/10.3906/biy-2105-65.
[17] H.T. Teker, T. Ceylani, S. Keskin, G. Samgane, B. Baba, E. Acıkgoz, R. Gurbanov, Reduced liver damage and fibrosis with combined SCD Probiotics and intermittent

fasting in aged rat, J. Cell Mol. Med. 28 (1) (2024) e18014, https://doi.org/10.1111/jcmm.18014.
[18] H.T. Teker, T. Ceylani, S. Keskin, G. Samgane, S. Mansuroglu, B. Baba, H. Allahverdi, E. Acıkgoz, R. Gurbanov, Age-related differences in response to plasma exchange in male rat liver tissues: insights from histopathological and machine-learning assisted spectrochemical analyses, Biogerontology 24 (4) (2023) 563–580, https://doi.org/10.1007/s10522-023-10032-3.
[19] H.T. Teker, T. Ceylani, S. Keskin, G. Samgane, H. Allahverdi, E. Acikgoz, R. Gurbanov, Supplementing probiotics during intermittent fasting proves more effective in restoring ileum and colon tissues in aged rats, J. Cell Mol. Med. 28 (6) (2024) e18203, https://doi.org/10.1111/jcmm.18203.
[20] D.L. Whaley, The interquartile range: theory and estimation, Journal Name (2005).
[21] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[22] D. Tarwidi, S.R. Pudjaprasetya, D. Adytia, M. Apri, An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach, MethodsX 10 (2023) 102119, https://doi.org/10.1016/j.mex.2023.102119.
[23] A. Testas, Support vector machine classification with Pandas, scikit-learn, and PySpark, in: A. Testas (Ed.), Distributed Machine Learning with PySpark: Migrating Effortlessly from Pandas and Scikit-Learn, Apress, Berkeley, CA, 2023, pp. 259–280, https://doi.org/10.1007/978-1-4842-9751-3_10.
[24] E. Gilmore, V. Estivill-Castro, R. Hexel, More interpretable decision trees, Journal Name (2021) 280–292.
[25] S.B. Kotsiantis, Decision trees: a recent overview, Artif. Intell. Rev. 39 (4) (2013) 261–283, https://doi.org/10.1007/s10462-011-9272-4.
[26] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347, https://doi.org/10.1093/bioinformatics/btq134.
[27] S. A, S. R, A systematic review of Explainable Artificial Intelligence models and applications: recent developments and future trends, Decision Analytics Journal 7 (2023) 100230, https://doi.org/10.1016/j.dajour.2023.100230.
[28] A. Salih, Z. Raisi-Estabragh, I.B. Galazzo, P. Radeva, S.E. Petersen, G. Menegaz, K. Lekadir, Commentary on Explainable Artificial Intelligence Methods: SHAP and LIME, 2023 arXiv preprint arXiv:2305.02012.
[29] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Appl. Sci. 3 (2021), https://doi.org/10.1007/s42452-021-04148-9.
[30] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions, J. Comput. Aided Mol. Des. 34 (10) (2020) 1013–1026, https://doi.org/10.1007/s10822-020-00314-0.
[31] J. Yang, Fast Treeshap: Accelerating Shap Value Computation for Trees, 2021 arXiv preprint arXiv:2109.09847.
[32] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.
[33] T. Ceylani, H.T. Teker, G. Samgane, R. Gurbanov, Intermittent fasting-induced biomolecular modifications in rat tissues detected by ATR-FTIR spectroscopy and machine learning algorithms, Anal. Biochem. 654 (2022) 114825, https://doi.org/10.1016/j.ab.2022.114825.
[34] İ. Ardahanlı, H. Özkan, F. Özel, R. Gurbanov, H.T. Teker, T. Ceylani, Infrared spectrochemical findings on intermittent fasting-associated gross molecular modifications in rat myocardium, Biophys. Chem. 289 (2022) 106873, https://doi.org/10.1016/j.bpc.2022.106873.
[35] B. Suh, S. Park, D.W. Shin, J.M. Yun, B. Keam, H.K. Yang, E. Ahn, H. Lee, J.H. Park, B. Cho, Low albumin-to-globulin ratio associated with cancer incidence and mortality in generally healthy adults, Ann. Oncol. 25 (11) (2014) 2260–2266, https://doi.org/10.1093/annonc/mdu274.
[36] J.R. Ulloque-Badaracco, M.D. Mosquera-Rojas, E.A. Hernandez-Bustamante, E. A. Alarcón-Braga, P. Herrera-Añazco, V.A. Benites-Zapata, Prognostic value of albumin-to-globulin ratio in COVID-19 patients: a systematic review and meta-analysis, Heliyon 8 (5) (2022) e09457, https://doi.org/10.1016/j.heliyon.2022.e09457.
[37] C.C.S. Bandeira, K.C.R. Madureira, M.B. Rossi, J.F. Gallo, A.P.M.A. da Silva, V. L. Torres, V.A. de Lima, N.K. Júnior, J.D. Almeida, R.M. Zerbinati, P.H. Braz-Silva, J.A.L. Lindoso, H. da Silva Martinho, Micro-Fourier-transform infrared reflectance spectroscopy as tool for probing IgG glycosylation in COVID-19 patients, Journal Name 12 (2022) 4269, https://doi.org/10.1038/s41598-022-08156-6.
[38] A. Martinez-Cuazitl, G.J. Vazquez-Zapien, M. Sanchez-Brito, J.H. Limon-Pacheco, M. Guerrero-Ruiz, F. Garibay-Gonzalez, R.J. Delgado-Macuil, M.G.G. de Jesus, M. A. Corona-Perezgrovas, A. Pereyra-Talamantes, M.M. Mata-Miranda, ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients, Sci. Rep. 11 (1) (2021) 19980, https://doi.org/10.1038/s41598-021-99529-w.
[39] M. Bouhaddou, D. Memon, B. Meyer, K.M. White, V.V. Rezelj, M. Correa Marrero, B.J. Polacco, J.E. Melnyk, S. Ulferts, R.M. Kaake, J. Batra, A.L. Richards, E. Stevenson, D.E. Gordon, A. Rojc, K. Obernier, J.M. Fabius, M. Soucheray, L. Miorin, E. Moreno, C. Koh, Q.D. Tran, A. Hardy, R. Robinot, T. Vallet, B. E. Nilsson-Payant, C. Hernandez-Armenta, A. Dunham, S. Weigang, J. Knerr, M. Modak, D. Quintero, Y. Zhou, A. Dugourd, A. Valdeolivas, T. Patil, Q. Li, R. Hüttenhain, M. Cakir, M. Muralidharan, M. Kim, G. Jang, B. Tutuncuoglu, J. Hiatt, J.Z. Guo, J. Xu, S. Bouhaddou, C.J.P. Mathy, A. Gaulton, E.J. Manners, E. Félix, Y. Shi, M. Goff, J.K. Lim, T. McBride, M.C. O'Neal, Y. Cai, J.C.J. Chang, D. J. Broadhurst, S. Klippsten, E. De Wit, A.R. Leach, T. Kortemme, B. Shoichet, M. Ott, J. Saez-Rodriguez, B.R. tenOever, R.D. Mullins, E.R. Fischer, G. Kochs, R. Grosse, A. García-Sastre, M. Vignuzzi, J.R. Johnson, K.M. Shokat, D.L. Swaney, P. Beltrao, N.J. Krogan, The global phosphorylation landscape of SARS-CoV-2 infection, Cell 182 (3) (2020) 685–712.e19, https://doi.org/10.1016/j.cell.2020.06.034.
[40] B.R. Wood, K. Kochan, D.E. Bedolla, N. Salazar-Quiroz, S.L. Grimley, D. Perez-Guaita, M.J. Baker, J. Vongsvivut, M.J. Tobin, K.R. Bambery, D. Christensen, S. Pasricha, A.K. Eden, A. McLean, S. Roy, J.A. Roberts, J. Druce, D.A. Williamson, J. McAuley, M. Catton, D.F.J. Purcell, D.I. Godfrey, P. Heraud, Infrared based saliva screening test for COVID-19, Angew Chem. Int. Ed. Engl. 60 (31) (2021) 17102–17107, https://doi.org/10.1002/anie.202104453.
[41] S. Alshammari, A.S. AlMasoudi, A.H. AlBuhayri, H.M. AlAtwi, S.S. AlHwiti, H. M. Alaidi, A.M. Alshehri, N.A. Alanazi, A. Aljabri, M.M. Al-Gayyar, Effect of COVID-19 on glycemic control, insulin resistance, and pH in elderly patients with type 2 diabetes, Cureus 15 (2) (2023) e35390, https://doi.org/10.7759/cureus.35390.
[42] M. Caterino, M. Gelzo, S. Sol, R. Fedele, A. Annunziata, C. Calabrese, G. Fiorentino, M. D'Abbraccio, C. Dell'Isola, F.M. Fusco, R. Parrella, G. Fabbrocini, I. Gentile, I. Andolfo, M. Capasso, M. Costanzo, A. Daniele, E. Marchese, R. Polito, R. Russo, C. Missero, M. Ruoppolo, G. Castaldo, Dysregulation of lipid metabolism and pathological inflammation in patients with COVID-19, Sci. Rep. 11 (1) (2021) 2941, https://doi.org/10.1038/s41598-021-82426-7.
[43] A. Martinez-Cuazitl, M.M. Mata-Miranda, M. Sanchez-Brito, D. Valencia-Trujillo, A.M. Avila-Trejo, R.J. Delgado-Macuil, C. Atriano-Colorado, F. Garibay-Gonzalez, V. Sanchez-Monroy, G.J. Vazquez-Zapien, Clinical, biochemical, and ATR-FTIR spectroscopic parameters associated with death or survival in patients with severe COVID-19, Journal of Spectroscopy 2023 (1) (2023) 3423183, https://doi.org/10.1155/2023/3423183.
[44] M. Fukuda, T. Aoyama, I. Hashimoto, Y. Maezawa, A.Y.A. Kato, K. Hara, K. Kazama, K. Komori, A. Tamagawa, H. Cho, T. Ishiguro, K. Segami, M. Nakazono, K. Otani, S.H.O. Sawazaki, M. Numata, S. Kawahara, T. Oshima, A.Y.A. Saito, N. Yukawa, Y. Rino, Albumin-globulin ratio is an independent prognostic factor for gastric cancer patients who received curative treatment, In Vivo 38 (2) (2024) 904, https://doi.org/10.21873/invivo.13517.
[45] M.C. Viana-Llamas, R. Arroyo-Espliguero, J.A. Silva-Obregón, G. Uribe-Heredia, I. Núñez-Gil, B. García-Magallón, C.G. Torán-Martínez, A. Castillo-Sandoval, E. Díaz-Caraballo, I. Rodríguez-Guinea, J. Domínguez-López, Hypoalbuminemia on admission in COVID-19 infection: an early predictor of mortality and adverse events. A retrospective observational study, Med. Clin. 156 (9) (2021) 428–436, https://doi.org/10.1016/j.medcli.2020.12.018.
[46] B.A. Cobb, The history of IgG glycosylation and where we are now, Glycobiology 30 (4) (2020) 202–213, https://doi.org/10.1093/glycob/cwz065.
[47] E.B. Irvine, G. Alter, Understanding the role of antibody glycosylation through the lens of severe viral and bacterial diseases, Glycobiology 30 (4) (2020) 241–253, https://doi.org/10.1093/glycob/cwaa018.
[48] S.K. Vadrevu, I. Trbojevic-Akmacic, A.V. Kossenkov, F. Colomb, L.B. Giron, A. Anzurez, K. Lynn, K. Mounzer, A.L. Landay, R.C. Kaplan, E. Papasavvas, L. J. Montaner, G. Lauc, M. Abdel-Mohsen, Frontline Science: plasma and immunoglobulin G galactosylation associate with HIV persistence during antiretroviral therapy, J. Leukoc. Biol. 104 (3) (2018) 461–471, https://doi.org/10.1002/JLB.3HI1217-500R.
[49] C.-H. Ho, R.-N. Chien, P.-N. Cheng, J.-H. Liu, C.-K. Liu, C.-S. Su, I.-C. Wu, I.-C. Li, H.-W. Tsai, S.-L. Wu, W.-C. Liu, S.-H. Chen, T.-T. Chang, Aberrant serum immunoglobulin G glycosylation in chronic hepatitis B is associated with histological liver damage and reversible by antiviral therapy, J. Infect. Dis. 211 (1) (2014) 115–124, https://doi.org/10.1093/infdis/jiu388.
[50] S. Chakraborty, J. Gonzalez, K. Edwards, V. Mallajosyula, A.S. Buzzanco, R. Sherwood, C. Buffone, N. Kathale, S. Providenza, M.M. Xie, J.R. Andrews, C. A. Blish, U. Singh, H. Dugan, P.C. Wilson, T.D. Pham, S.D. Boyd, K.C. Nadeau, B. A. Pinsky, S. Zhang, M.J. Memoli, J.K. Taubenberger, T. Morales, J.M. Schapiro, G. S. Tan, P. Jagannathan, T.T. Wang, Proinflammatory IgG Fc structures in patients with severe COVID-19, Nat. Immunol. 22 (1) (2021) 67–73, https://doi.org/10.1038/s41590-020-00828-7.
[51] S. Karthikeyan, G.J. Vazquez-Zapien, A. Martinez-Cuazitl, R.J. Delgado-Macuil, D. E. Rivera-Alatorre, F. Garibay-Gonzalez, J. Delgado-Gonzalez, D. Valencia-Trujillo, M. Guerrero-Ruiz, C. Atriano-Colorado, A. Lopez-Reyes, D. Lopez-Mezquita, M. M. Mata-Miranda, Two-trace two-dimensional correlation spectra (2T2D-COS) analysis using FTIR spectra to monitor the immune response by COVID-19, Journal Name (2023), 10.21203/rs.3.rs-2856060/v110.21203/rs.3.rs-2856060/v1.
[52] S. Karthikeyan, M.M. Mata-Miranda, A. Martinez-Cuazitl, R.J. Delgado-Macuil, F. Garibay-Gonzalez, V. Sanchez-Monroy, A. Lopez-Reyes, M. Rojas-Lopez, D. E. Rivera-Alatorre, G.J. Vazquez-Zapien, Dynamic response antibodies SARS-CoV-2 human saliva studied using two-dimensional correlation (2DCOS) infrared spectral analysis coupled with receiver operation characteristics analysis, Biochim. Biophys. Acta (BBA) - Mol. Basis Dis. 1869 (7) (2023) 166799, https://doi.org/10.1016/j.bbadis.2023.166799.
[53] A. Parmar, R. Katariya, V. Patel, A Review on Random Forest: An Ensemble Classifier (2019) 758–763, https://doi.org/10.1007/978-3-030-03146-6_86.
[54] L. Langsetmo, J.T. Schousboe, B.C. Taylor, J.A. Cauley, H.A. Fink, P.M. Cawthon, D.M. Kado, K.E. Ensrud, Advantages and disadvantages of random forest models for prediction of hip fracture risk versus mortality risk in the oldest old, JBMR Plus 7 (8) (2023) e10757, https://doi.org/10.1002/jbm4.10757.
[55] K.L. Flanagan, A.L. Fink, M. Plebanski, S.L. Klein, Sex and gender differences in the outcomes of vaccination over the life course, Annu. Rev. Cell Dev. Biol. 33 (33) (2017) 577–599, https://doi.org/10.1146/annurev-cellbio-100616-060718, 2017.
[56] C.E. Gebhard, C. Sütsch, S. Bengs, M. Deforth, K.P. Buehler, N. Hamouda, A. Meisel, R.A. Schuepbach, A.S. Zinkernagel, S.D. Brugger, C. Acevedo, D. Patriki, B. Wiggli, J.H. Beer, A. Friedl, R. Twerenbold, G.M. Kuster, H. Pargger, S. Tschudin-

Sutter, J.C. Schefold, T. Spinetti, A. Dussault-Cloutier, C. Henze, M. Pasqualini, D.F. Sager, L. Mayrhofer, M. Grieder, J. Tontsch, F. Franzeck, P.D. Wendel Garcia, D.A. Hofmaenner, T. Scheier, J. Bartussek, L. Chrobok, D. Stähli, N. Lott, A. Haider, M. Grämer, N. Mikail, A. Rossi, N. Zellweger, P. Opic, A. Portmann, A. Todorov, A.P. Pazhenkottil, M. Messerli, R.R. Buechel, P.A. Kaufmann, V. Treyer, M. Siegemund, U. Held, V. Regitz-Zagrosek, C. Gebhard, Sex- and gender-specific risk factors of post-COVID-19 Syndrome: a population-based cohort study in Switzerland, medRxiv 10.1101/2021.06.30.21259757 (2021) 2021.06.30.21259757, doi: 10.1101/2021.06.30.21259757. .

[57] C.C. Sawyer, Child mortality estimation: estimating sex differences in childhood mortality since the 1970s, PLoS Med. 9 (8) (2012) e1001287, https://doi.org/10.1371/journal.pmed.1001287.

[58] R. Chaturvedi, B. Lui, J.A. Aaronson, R.S. White, J.D. Samuels, COVID-19 complications in males and females: recent developments, J Comp Eff Res 11 (9) (2022) 689–698, https://doi.org/10.2217/cer-2022-0027.

[59] B. Ludwig, Infrared spectroscopy studies of aluminum oxide and metallic aluminum powders, Part II: adsorption reactions of organofunctional silanes, Powders 1 (2) (2022) 75–87.

[60] K.I. Hadjiivanov, D.A. Panayotov, M.Y. Mihaylov, E.Z. Ivanova, K.K. Chakarova, S. M. Andonova, N.L. Drenchev, Power of infrared and Raman spectroscopies to characterize metal-organic frameworks and investigate their interaction with guest molecules, Chem. Rev. 121 (3) (2021) 1286–1424, https://doi.org/10.1021/acs.chemrev.0c00487.

[61] A. Fadlelmoula, D. Pinho, V.H. Carvalho, S.O. Catarino, G. Minas, Fourier transform infrared (FTIR) spectroscopy to analyse human blood over the last 20 Years: a review towards lab-on-a-chip devices, Micromachines 13 (2) (2022) 187.

[62] Q. Wang, P. Fang, R. He, M. Li, H. Yu, L. Zhou, Y. Yi, F. Wang, Y. Rong, Y. Zhang, A. Chen, N. Peng, Y. Lin, M. Lu, Y. Zhu, G. Peng, L. Rao, S. Liu, O-GlcNAc transferase promotes influenza A virus–induced cytokine storm by targeting interferon regulatory factor–5, Sci. Adv. 6 (16) (2020) eaaz7086, https://doi.org/10.1126/sciadv.aaz7086.

[63] J.A. Critchley, I.M. Carey, T. Harris, S. DeWilde, F.J. Hosking, D.G. Cook, Glycemic control and risk of infections among people with type 1 or type 2 diabetes in a large primary care cohort study, Diabetes Care 41 (10) (2018) 2127–2135, https://doi.org/10.2337/dc18-0287.

[64] D.L. Kitane, S. Loukman, N. Marchoudi, A. Fernandez-Galiana, F.Z. El Ansari, F. Jouali, J. Badir, J.-L. Gala, D. Bertsimas, N. Azami, O. Lakbita, O. Moudam, R. Benhida, J. Fekkak, A simple and fast spectroscopy-based technique for Covid-19 diagnosis, Sci. Rep. 11 (1) (2021) 16740, https://doi.org/10.1038/s41598-021-95568-5.

[65] S.T. Kazmer, G. Hartel, H. Robinson, R.S. Richards, K. Yan, S.J. van Hal, R. Chan, A. Hind, D. Bradley, F. Zieschang, D.J. Rawle, T.T. Le, D.W. Reid, A. Suhrbier, M. M. Hill, Pathophysiological response to SARS-CoV-2 infection detected by infrared spectroscopy enables rapid and robust saliva screening for COVID-19, Biomedicines 10 (2) (2022) 351.