Potential of vibrational spectroscopy coupled with machine learning as a non-invasive diagnostic method for COVID-19

Bingqiang Zhao, Honglin Zhai*, Haiping Shao, Kexin Bi, Ling Zhu

College of Chemistry & Chemical Engineering, Lanzhou University; South Tianshui Road 222, Lanzhou, Gansu 730000, PR China

ARTICLE INFO

ABSTRACT

Background and Objective: Efforts to alleviate the ongoing coronavirus disease 2019 (COVID-19) crisis showed that rapid, sensitive, and large-scale screening is critical for controlling the current infection and that of ongoing pandemics.

Methods: Here, we explored the potential of vibrational spectroscopy coupled with machine learning to screen COVID-19 patients in its initial stage. Herein presented is a hybrid classification model called grey wolf optimized support vector machine (GWO-SVM). The proposed model was tested and comprehensively compared with other machine learning models via vibrational spectroscopic fingerprinting including saliva FTIR spectra dataset and serum Raman scattering spectra dataset.

Results: For the unknown vibrational spectra, the presented GWO-SVM model provided an accuracy, specificity and F1_score value of 0.9825, 0.9714 and 0.9778 for saliva FTIR spectra dataset, respectively, while an overall accuracy, specificity and F1_score value of 0.9085, 0.9552 and 0.9036 for serum Raman scattering spectra dataset, respectively, which showed superiority than those of state-of-the-art models, thereby suggesting the suitability of the GWO-SVM model to be adopted in a clinical setting for initial screening of COVID-19 patients.

Conclusions: Prospectively, the presented vibrational spectroscopy based GWO-SVM model can facilitate in screening of COVID-19 patients and alleviate the medical service burden. Therefore, herein proof-of-concept results showed the chance of vibrational spectroscopy coupled with GWO-SVM model to help COVID-19 diagnosis and have the potential be further used for early screening of other infectious diseases.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

The ongoing coronavirus disease 2019 (COVID-19) pandemic has led to over 632 million confirmed infections and almost 6.6 million deaths up to now (https://coronavirus.jhu.edu/map.html) with multiple emerging severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants have recently been identified [1]. Signifi-

cant obstacles for the diagnosis of SARS-CoV-2 infection continue to hamper large-scale population-based screening to control the COVID-19 pandemic in the absence of widely available antiviral therapeutics [2]. Nasopharyngeal swabs nucleic acid amplification test, namely real-time reverse transcription polymerase chain reaction (RT-PCR) [3], is the gold standard for SARS-CoV-2 detection, plays a critical role in detecting infection, determining infection rate, characterizing the disease progression, and guiding clinical decision-making. However, some technical shortcomings such as long processing time, laborious, involving specialized instruments and skilled personal encourage the further efforts to develop more reliable diagnosis strategy. Hence, some other testing methods including chest computed tomography (CT) [4,5], chest X-ray images [6–8], enzyme-linked immunosorbent assays [9], chemiluminescence immunoassays [10], lateral flow immunoassays [11], electrochemical biosensor [12], and mass spectrometry [13–15], were presented as an alternative method for SARS-CoV-2 detection. Though these assays have been effective in reducing transmission rate, it is imperative to develop individual-friendly biofluids such as saliva or serum for COVID-19 patients screening and diagnosis. Vibrational spectroscopy [16], based on Fourier transform infrared (FTIR) spec-

troscopy [17,18] or Raman scattering spectroscopy [19–21] provides a detailed fingerprinting of a biological sample from its chemical composition. Diagnostic tools based on these technologies reveal the potential to revolutionize clinical systems leading to evolutionary patient outcome, more efficient public services and significant economic savings. Moreover, vibrational spectroscopy possesses label-free disease detection and diagnosis in a single step [22,23]. However, obtaining clinical diagnosis speeds and accuracies remains challenging due to weak infrared absorption or Raman scattering signals from samples. Thus, developing rapid, sensitive, and large-scale screening methods could be advantageous to prevent its spread and mitigate the pandemic.

Fortunately, artificial intelligence (AI) has been deployed at various levels of the healthcare systems, including diagnosis, public health, clinical decision-making, and therapeutics. Fong et al. discussed how AI can help fight this deadly virus, from early warnings, prompt emergency responses, and critical decision-making to surveillance drones [24]. Meanwhile, many machine learning (ML) methods have been used for classification or disease diagnosis to obtain meaningful information from vibrational spectroscopy [25,26]. Linear discriminative analysis (LDA) [27], k-nearest neighbors (k-NN) [28,29], random forest (RF) [30], Naïve Bayes (NB) [31], and support vector machines (SVM) [32,33] are the widely used ML algorithms for classification in disease diagnosis purpose. The presence of ML could be said to accelerate the speed for the development of fast, accurate and sensitive methods which allowing for the detection and diagnosis of infectious diseases [34–36]. Interestingly, SVM can deal with high dimensional data comprising multiple features from vibrational spectroscopy, which has been verified to have outstanding performance in many fields and is often considered the best classifiers [37,38]. The flexible characteristic of employing different kernel functions (*e.g.*, linear, polynomial, RBF and sigmoid) to demarcate the hyperplane boundary enables to discriminate high dimensional data. Particularly, SVM based on the radical basis function (RBF) kernel function is widely used for pattern recognition [39]. The penalty coefficient C and the kernel coefficient γ are two significant parameters of the SVM with RBF kernel function. Recently, many swarm intelligence (SI) optimization algorithms derived from meta-heuristics have been proposed in recent years [40,41]. Grey wolf optimization (GWO) is one such algorithm proposed by Seyedali Mirjalili et al. [42,43], which is a population-based optimization algorithm inspired by the leadership hierarchy and hunting mechanism of grey wolf. A detailed mathematical description of GWO optimizer was included in the **Supplementary Material** section 4.

Here, we reported a vibrational spectroscopy fingerprinting-ML model for the discrimination of saliva and serum specimen, which could remarkably discriminate the signal of positive patients infected by SARS-CoV-2 from healthy individuals or suspected with symptoms like COVID-19. Specifically, we developed a flexible discriminatory tool by combining GWO optimizer and SVM classifier (GWO-SVM) for the diagnostic of COVID-19 patients based on their vibrational spectroscopy. To obtain a model with better generalization property, we selected the best hyperparameters (C and γ) combination by GWO to avoid overfitting or local minima problems. The fitness function used was classification accuracy, which is one of the most popular metrics in classification models and it was directly computed from the confusion matrix. The proposed architecture consists of two steps. Firstly, GWO optimizer was employed to adjust the hyperparameter of SVM classifier to predict the vibrational spectroscopy for COVID-19 diagnosis. Secondly, GWO-SVM model was carried out to discriminate unknown vibrational spectroscopy of saliva and serum specimens and calculate the classification accuracy based on the optimal hyperparameter combinations obtained from GWO optimizer. Simultaneously, five other ML methods including LDA, kNN, NB, RF and SVM,

were tested for their classification performances as the references. Within this context, present study showed that vibrational spectroscopy in coupled with GWO-SVM model for analysis of saliva or serum biofluids from suspected patients can become a novel rapid, cost-effective diagnostic tool for COVID-19 and have the potential be further used for early screening of other infectious diseases.

The remainder of the paper is organized as follows. In Section 2, we reviewed the current state-of-the-art literature related to the COVID-19 detection and vibrational spectroscopy applications in biomedical engineering. In Section 3, we described the proposed method and experimental analysis. In Section 4, we presented the performance metrics and experimental results. Section 5 contains a discussion and outlook on the provided issue. Finally, a brief summary was described in last section.

2. Related works

In this section, we briefly reviewed the current state-of-the-art literature about the application of various techniques to COVID-19 detection and vibrational spectroscopy for biomedical scenarios.

2.1. Diagnosis methods of COVID-19

The ongoing COVID-19 pandemic caused by SARS-CoV-2 infection has led to severe economic burdens worldwide. Multiple emerging SARS-CoV-2 variants have been identified and are now spreading internationally. Efficient outbreak control will then need cost-effective and easy-to-operate detection tools that can be easily deployed in low-resource situations [44]. Currently, three approaches, including RT-PCR, serological/immunological antigen-based test and chest CT are generally used for COVID-19 diagnosis. For instance, Ketan et al. reported a rapid RNA extraction-free lateral flow assay for molecular point-of-care (POC) detection of SARS-CoV-2 augmented by chemical probes. The assay uses highly specific 6-carboxyfluorescein and biotin labeled antisense oligonucleotides as probes designed to target N-gene sequence of SARS-CoV-2. Besides, they utilized cysteamine capped gold-nanoparticles to augment the signal [45]. Zhang et al. established an integrated system, which incorporates a ML-based FTIR for rapid COVID-19 screening and air-plasma-based disinfection modules to prevent potential secondary infectious. A partial least squares discrimination analysis and a convolutional neural network model were built using the collected infrared spectral dataset of serum samples. The sensitivity, specificity and accuracy all reach over 94% from the blind test samples [46]. Zhang et al. reported a rapid and sensitive magneto fluidic immune-PCR platform that can address the current gap in POC serological testing for COVID-19. They evaluated this magneto fluidic immune-PCR platform with 108 clinical serum samples and achieved 93.8% sensitivity and 98.3 specificity, demonstrating its potential as a rapid and sensitive POC serological test for COVID-19 [47]. However, the utility of most of these techniques are limited. While serological tests suffer from the cross-reactivity with other pathogens, such as other human coronaviruses, immunological ones is limited by a detectable antibody response at the early stages of infection. Of the above, chest CT is a key screening tool for patients with COVID-19 symptoms. Mei et al. used artificial intelligence algorithms to integrate chest CT findings with clinical symptoms, exposure history and laboratory testing to rapidly diagnose patients who are positive for COVID-19. In a test set of 279 patients, the AI systems achieved an area under the curve of 0.92 and had equal sensitivity as compared to a senior thoracic radiologist [48]. Despite be widely available in cities, CT facilities typically do not reliably detect COVID-19 infection in its early stage, making them unsuitable for intensive patient surveillance.

2.2. Vibrational spectroscopy for biomedical applications

The potential of vibrational spectroscopy for biomedical applications has been well established through many proofs of concept studies over the past decades [49,50]. Due to its unique fingerprinting capability, vibrational spectroscopy can play significant role in disease diagnosis and drug discovery. However, these vibrational spectroscopies are often possessed multivariate signatures that allow one to differentiate between patients with disease and healthy controls. Marcia et al. developed a noninvasive diagnostic for COVID-19 from saliva biofluid via FTIR spectroscopy and multivariate analysis. They evaluated a mid-infrared dataset of saliva samples obtained from symptomatic patients using unsupervised and supervised strategy. This method presents an important tool for a fast, noninvasive diagnostic technique, reducing costs and allowing for infection risk reduction [51]. Du et al. proposed a method based on Raman spectroscopy combined with generative adversarial network and multiclass SVM to classify foodborne pathogenic bacteria. Better classification results are obtained by optimizing the parameters of the multiclass SVM [52]. Shu et al. developed a deep learning guided fiberoptic Raman diagnostic platform to assess its ability of real-time *in vivo* nasopharyngeal carcinoma diagnosis and post-treatment follow-up patients. The robust Raman diagnostic platform was established using multi-layer Raman-specified convolutional neural networks together with simultaneous fingerprint and high-wavenumber spectra acquired within sub-seconds. The optimized model provides an overall diagnostic accuracy of 82.09% for identifying nasopharyngeal carcinoma from control and post-treatment patients [22]. Based on the evidence provided herein, vibrational spectroscopy in combination with ML provides the first glimmer of hope for the development of an accurate, inexpensive, rapid, and non-invasive method for universal biomedical applications.

3. Methods

3.1. Data preparation

The first dataset consists of total 171 saliva FTIR spectra that 87 spectra for positive with SARS-CoV-2 and 84 spectra for negative with SARS-CoV-2, derived from 29 patients testing positive and 28 individuals testing negative with COVID-19 like symptoms receiving treatment in the Royal Melbourne Hospital by RT-PCR [53]. According to previously described, saliva was collected by asking the patient to cough out saliva from their throat into a sterile container, and viral transport medium was added. 10 mL saliva was deposited onto an infrared reflective substrate with three deposits per slide from a patient cohort and FTIR spectra (4000-800 cm^{-1}) were recorded on a Perkin-Elmer Spectrum 2 spectrometer with a dedicated reflection accessory. Totally, 183 spectra were acquired from 61 patients and suspected for anomalies. However, spectra belonging to four patients (2 negative and 2 positive) were identified as outliers, due to a high content of viral transport medium, or the spectra had no contribution on the 2058 cm^{-1} band, typical of a saliva spectrum.

The second dataset comprised of total 465 serum Raman scattering spectra (1800-600 cm^{-1}) from the serum of 53 patients who were confirmed with COVID-19 by RT-PCR, 50 healthy individuals and 54 individuals with flu symptoms like COVID-19 patients. According to the Gang Yin et al.[34], the Raman spectra for each serum sample was implemented by three experimenters and repeated five times, respectively. Finally, the five spectra collected from each serum sample by three experimenters were respectively averaged. Among the total serum samples in the suspected group, only two Raman scattering spectra were obtained in subjects 16 to 21. The details of the sample preparation, instrumentation, and vibrational spectra analysis can be referred to previously work by Gang Yin et al. [54].

The original saliva FTIR spectra and serum Raman scattering spectra were illustrated in Fig. 1. The total datasets were randomly divided into a training phase (2/3 of the dataset) and a test phase (1/3 of the dataset) using the Kennard-Stone sample selection algorithm [55] to develop our model.

3.2. Spectral preprocessing

Data preprocessing is essential in vibrational spectra, as the FTIR and Raman scattering spectra are often affected by multiple noise sources, such as instrument, environmental condition, nature of the sample and other factors [56]. For the saliva FTIR spectra dataset, the spectra were firstly truncated in the bio-fingerprint region (1300-800 cm^{-1}), this is where many important RNA and glycoprotein marker bands are located and there is less interference from viral transport media, thus increase the robustness of the modeling, because the more variables fed into the model, such as the variables account for viral transport media in the sample, the more chance of finding spurious correlations. Secondly, following preprocessing procedure were employed: Savitzky-Golay (SG) smoothing [57] to remove unnecessary noise from the original spectra (window = 9 points, 2nd order polynomial function), multiplicative scatter correction (MSC) [58] to adjust the light scattering and adaptive iteratively reweighted penalized least squares (airPLS) [59] to remove baseline absorptions. The total and mean spectra of two groups after preprocessing were referred to **Fig. S2-4.** For the serum Raman spectra dataset, an outlier detection [60] was implemented to exclude those samples away from others (**Fig. S5**). In addition, the original spectra were truncated in the bio-fingerprint region (1800-600 cm^{-1}), SG smoothing (window = 15
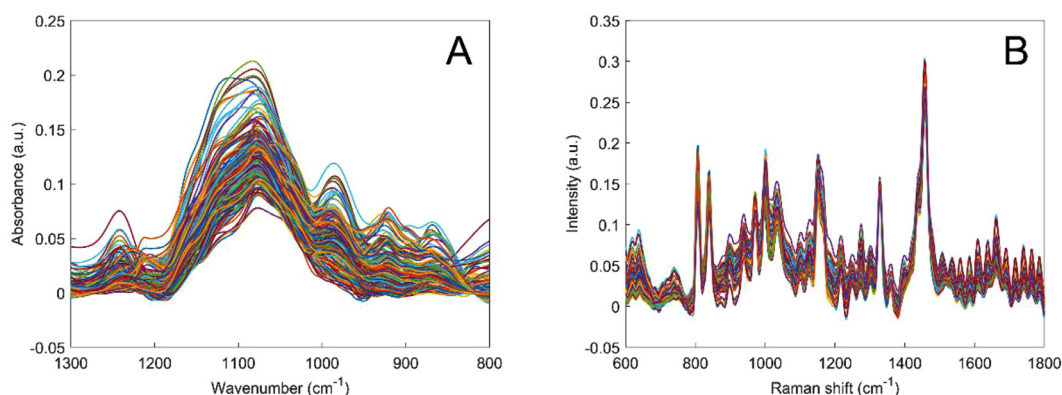

Fig. 1. (A) Raw saliva FTIR spectra dataset. (B) Raw serum Raman scattering spectra dataset.

points, 2nd order polynomial function), MSC and airPLS were also implemented for preprocessing. The total and mean spectra of three groups after preprocessing were referred to **Fig. S6-10**. It can be noted that the difference between the two vibrational spectra datasets is very tiny. Particularly, closely related spectra show almost identical mean FTIR or Raman scattering spectra. Therefore, vibrational spectroscopy needed to be analyzed using advanced statistical methods.

3.3. Feature extraction

In most cases, variables from spectroscopy, *e.g.*, Raman scattering spectroscopy and FTIR spectroscopy, are proportional to noise signals, which usually lead to a collinearity problem. Supervised models that often minimize bias (*i.e.*, error in the estimates) on training data tend to be overfitting when the number of object or spectra sample is less than the number of variables, so the model generalizability would be poor on test data. Feature extraction is an alternative strategy to deal with the highly correlated variables before modeling. Here the Tchebichef curve moments (TCM) [61] was employed for feature extraction. As one of the discrete orthogonal moments, TCM exhibit the characteristic of powerful curve description capability, multi-resolution property, invariance property and thus can be used to capture important features of an spectra curve [62,63]. A brief note on TCM instructions and their mathematical description can be referred in **Supplementary Material**

3.4. Discriminant analysis using ML algorithms

The proposed GWO-SVM model consists of four main procedures (Scheme 1). Firstly, a herd of grey wolves were randomly created. There were two hyperparameters (C and γ) to be optimized in the SVM classifier, thus, each grey wolf population was termed as a two-dimensional array. Secondly, the fitness function of the GWO-SVM was determined based on the cross-validation accuracy. Thirdly, the fitness of each population would be obtained after the initial population was established. These fitness values were ranked to find the three individuals with the highest fitness values and marked as the three grey wolves with the highest hunting ability, which were called as α, β and δ wolves, and then they were used to guide the position updating for other wolves. Hereby, a new population with the updated grey wolf positions were formed and the individual fitness was obtained and evaluated. The above process was repeated consistently until the maximum number of iterations accomplished. Finally, when the iteration process completed, the ideal solution would be fed into SVM classifier to make it the finest classifier. And the classification performance of the GWO-SVM model was measured by the test phase divided from the original data set.

To fully gauge the discrimination capabilities of GWO-SVM model for saliva FTIR spectra and serum Raman scattering spectra, several other supervised ML algorithms were considered in this work, including LDA, kNN, NB, RF and SVM. These were introduced succinctly in **Supplementary Material** with their main features.

3.5. Figures of merit

Cross-validation was employed to determine the validity of a model on test phase by evaluating if the model is overfitted to noise. Considering the small number of samples, hyperparameters of models were finetuned by a leave-one-spectra-out cross-validation (LOSOCV) and the determined one was chosen according to the global optimum accuracy. In this regard, the original cali-
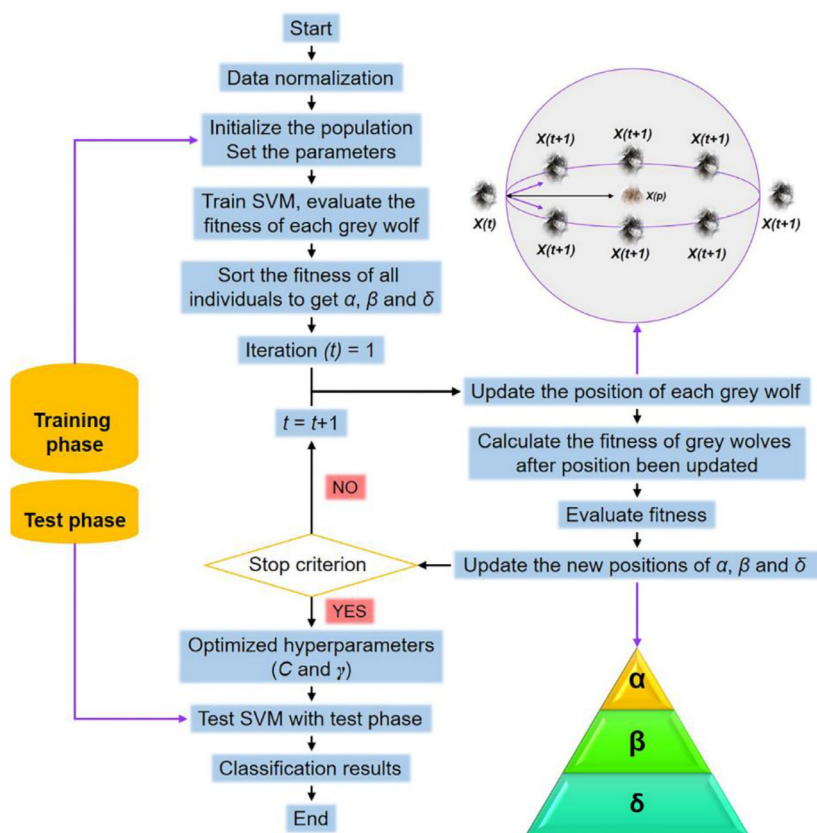
Scheme 1. Illustration of the whole procedure of grey wolves optimized support vector machine (GWO-SVM).

bration dataset was randomly divided into two phase, C and C\i, where C was the training phase and C\i was the cross-validation phase. Firstly, the training phase was used to train the model, then the cross-validation phase was used to verify the accuracy of the classification model. When the final models were achieved using different ML algorithms, the performance of each one would be back-evaluated according to the prediction accuracy for the training phase. Finally, the optimal model was validated by independent samples in the test phase with prediction accuracy.

The model validation of the ML models was performed based on seven figures of merit (FOM) [64,65], namely accuracy, precision, recall, F1-score, specificity, area under the receiver operating characteristics curve (AUROC) and area under the precision-recall curve (AUPRC). Accuracy indicates the proportion of correctly predicted events using the optimized classification model. Precision and recall are the two basic FOM for classification model, while F1-score is their complementary parameter. Precision is a FOM expressing the proportion of correctly classified positive samples among all samples classified as positive. Thus, this is a more robust FOM to evaluate the detection of positive samples. Recall is the proportion of true positives (TP) among all positives, while specificity is the rate of true negative (TN) predictions. F1-score is the weighted average value between precision and recall, which considers both false negatives (FN) and false positives (FP) into account, thus, which is particularly useful for unbalanced dataset. For the multi-class classification, the same previous parameters were calculated, but in macro-averaging level as mentioned in Marina Sokolova et al. [66]. The equations used to calculate these performance metrics were as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{Tp}{TP + FN}$$

$$F1\_score = 2 \times \frac{precision \times Recall}{precision + Recall}$$

$$Specificity = \frac{TN}{TN + FP}$$

The receiver operating characteristics (ROC) analysis [67] is based on statistical decision theory and has been applied extensively to the evaluation of classification methods. The ROC curve can manifest the relationship between the true positive rate (TPR) and false positive rate (FPR) with the variations of decision threshold [68]. Specifically, AUROC is one of the most widely used metrics for overall discrimination ability of a classification model. It ranges between 0.5 and 1.0. A model with an AUROC of 1.0 suggests perfect separation ability, while an AUROC of 0.5 suggests there is no class separation. Similar to the ROC curve, the precision-recall curve (PRC) is also a useful metrics for unbalanced dataset, which shows the tradeoff between precision and recall during different threshold. A higher AUPRC indicates both higher precision and higher recall, where higher precision relates to a slow false positive rate (FPR), and higher recall relates to a low false negative rate (FNR).

3.6. Implementation

TCM for feature extraction routines were a custom-written program in m-file. LIBSVM toolbox was downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. All algorithms of spectral preprocessing, feature extraction, optimal wavelength selection, model calibration, and visualization were performed in MATLAB 7.0 (The MathWorks, Inc., Natick, USA) via lab-made routines on a desktop with Intel(R) Core (TM) i7-4770K CPU @ 3.50GHz and 64 GB RAM, with Windows 10 operating system (professional version).

4. Results

According to the results, the classification performance of the SVM that its hyperparameter combination was optimized by grid search (GS-SVM) was not desired during the test phase, as the accuracy was lower than 0.9. For saliva FTIR spectra dataset, in the case of GS-SVM, the test accuracy was 0.8947 and the AUROC was 0.9662. Moreover, in the Wood's work, the model recall and specificity were 0.93 and 0.82, respectively [53]. However, the accuracy obtained by GWO-SVM during the test phase can reach over 0.98, which was almost close to 1.0, while other ML algorithms cannot achieve such values even after many tests, suggesting the GWO-SVM model exhibited superiority to classify saliva FTIR spectra dataset (Fig. 2). In addition, it can be noted that the GWO-SVM
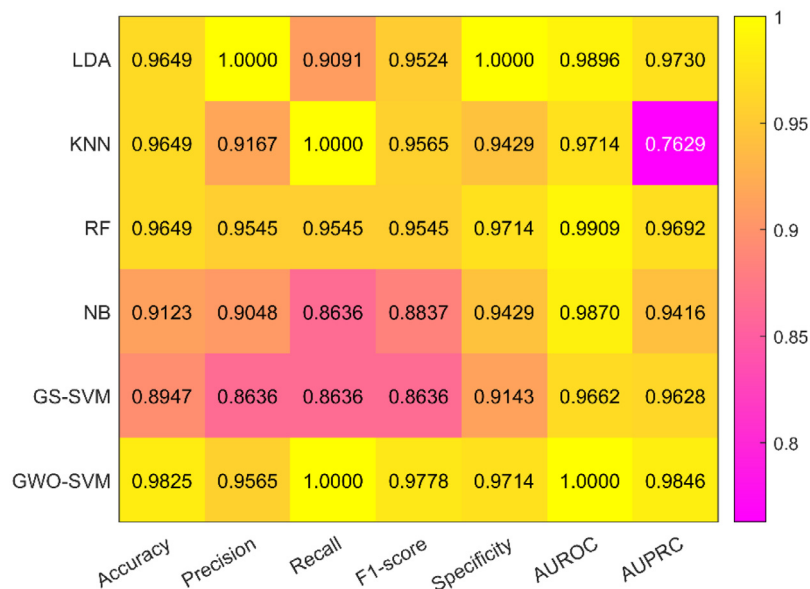

Fig. 2. Summary of the accuracy, precision, recall, F1-score, specificity, AUROC and AUPRC values obtained from six machine learning methods for unknown saliva FTIR spectra.
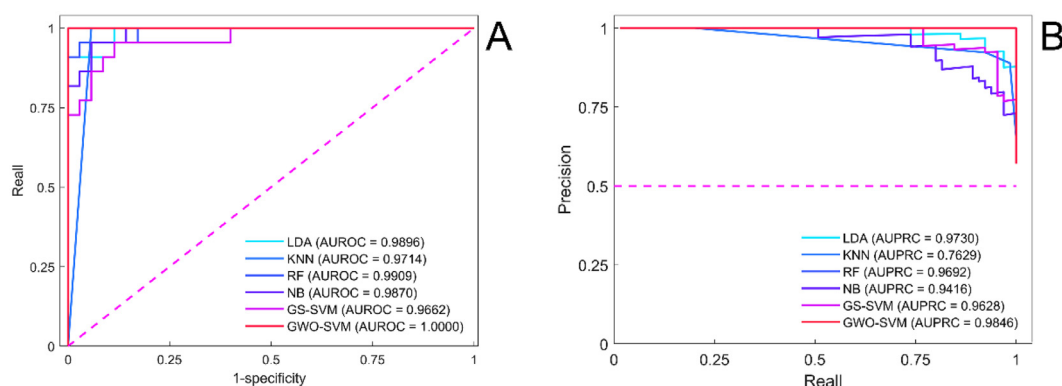
Fig. 3. (A) ROC curves of six machine learning methods for unknown saliva FTIR spectra. (B) PRC curves of six machine learning methods for unknown saliva FTIR spectra. (LDA: Linear discriminative analysis. *k*-NN: *k*-nearest neighbors. RF: Random Forest. NB: Naïve Bayes. GS-SVM: Grid search optimized support vector machines. GWO-SVM: Grey wolf optimized support vector machine).
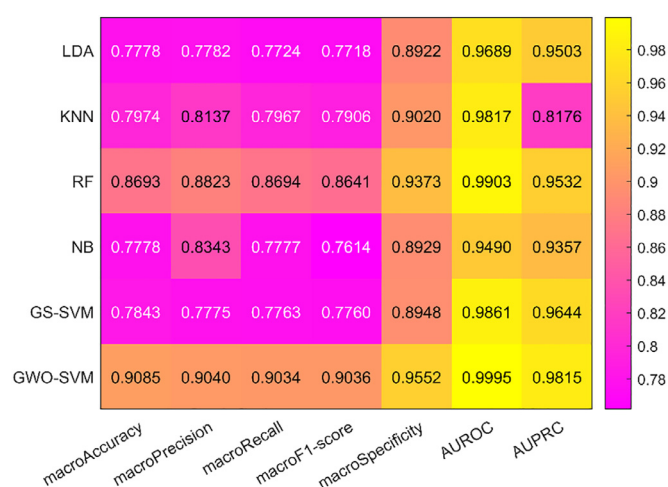

Fig. 4. Summary of the macroAccuracy, macroPrecision, macroRecall, macroF1-score, macroSpecificity, AUROC and AUPRC values obtained from six machine learning methods for unknown serum Raman scattering spectra.

got better results than other ML models. In particularly, the recall and AUROC were equals to 1.0, suggesting that a randomly positive sample spectra (*i.e.*, saliva FTIR spectra from patient with COVID-19) will be predicted more likely to be COVID-19 than a randomly negative sample (*i.e.*, saliva FTIR spectra from patient with non-COVID-19 with probability 1.0 (Fig. 3).

In addition, the overall accuracy of GS-SVM classifier for serum Raman scattering spectra dataset was only 0.7843. Moreover, the results presented significant variation when the vibrational spectra in the two datasets were classified several times. However, GWO-SVM classifier significantly outperformed other ML models to classify the serum Raman spectra dataset with an overall accuracy of 0.9085 and AUROC of 0.9995 in the test phase (Fig. 4). Moreover, Fig. 3A and Fig. 5A compared the ROC curves of the GWO-SVM and other ML models with the optimal hyperparameters. In the two vibrational spectra datasets, the highest AUROC both belonged to the GWO-SVM. As mentioned above, the higher the AUROC (the perfect value is 1.0) is, the better classification performance of the model will be. It was suggested that GWO-SVM possess a well-behaved classification capability for positive and negative samples based on saliva FTIR spectra dataset. Furthermore, the PRC of the classification model was considered. As presented in the Fig. 3B and Fig. 5B, the AUPRC for both vibrational spectra datasets were superior to GS-SVM as well as other machine learning models. In addition to the modeling procedure, the GWO-SVM converges faster than the

GS-SVM. Considering that the GWO-SVM had better classification performance than the GS-SVM in test phase for both vibrational spectra datasets, it deduced that a faster convergence rate can be achieved by GWO-SVM model based on these two vibrational spectroscopy datasets while ensuring the classification performance.

5. Discussion

Clinical decisions are often complex and include continuous against trade-offs between numerous and frequently clashing targets. The rapid ongoing spread of COVID-19 over the world results to force the healthcare service systems. Generally, infrared spectroscopy is a routine analytical technique for molecular functional groups identification in organic chemistry and material chemistry. When infrared light meets vibrational modes of molecules, a unique fingerprinting of the sample will be generated. Saliva is emerging as an attractive medium for POC diagnosis of COVID-19 in the current pandemic. The SARS-CoV-2 virus has a preferential tropism to human airway epithelial cells that express the cellular receptor angiotensin-converting enzyme 2 (ACE2) [69]. Besides, ACE2 was found to be higher in salivary glands compared to the lungs, suggesting that salivary glands could be a potential target for SARS-CoV-2 virus [70]. In addition, serum testing as a routine testing item in clinical situations, can provide low-cost and rapid screening of patients in hospitals. In clinical applications, Raman scattering testing can be performed in routine serum testing items for COVID-19 screening. Once high-risk patients are found, they are immediately quarantined and further confirmed with RT-PCR technique, thus reducing the risk of infection in medical institutions. In this paper, a hybrid classification model called GWO-SVM was proposed to explore the potential of vibrational spectroscopy coupled with ML to screen COVID-19 patients in its initial stage. The GWO-SVM model was tested and comprehensively compared with other ML models via vibrational spectroscopic fingerprinting including saliva FTIR spectra dataset and serum Raman scattering spectra dataset.

5.1. Feature extraction

The maximum order of TCM not only affect the extracted information from the original vibrational spectroscopy but also contribute to the total number of variables fed into ML models. For saliva FTIR spectra dataset, the reconstruction error decreases as the maximum order of TCM increases, and reconstruction error variation was gradually tended to be stable when mM = 61 (**Fig. S11**). The reconstructed FTIR spectra under different order mM were referred to **Fig. S12**. Hence, the maximum order of TCM for
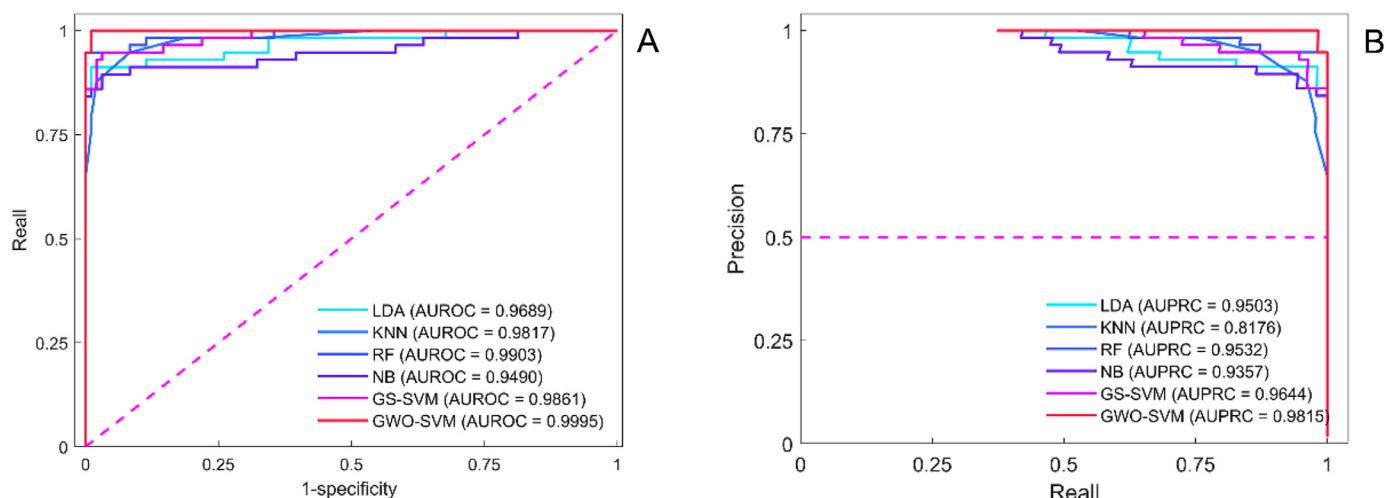
Fig. 5. (A) ROC curves of six machine learning methods for unknown serum Raman scattering spectra. (B) PRC curves of six machine learning methods for unknown serum Raman scattering spectra. (LDA: Linear discriminative analysis. k-NN: k-nearest neighbors. RF: Random Forest. NB: Naïve Bayes. GS-SVM: Grid search optimized support vector machines. GWO-SVM: Grey wolf optimized support vector machine).

saliva FTIR spectra dataset was determined and as a result a total of 62 TCMs were obtained. The calculation of the maximum order of TCM for serum Raman scattering spectra dataset was like that of saliva FTIR spectra dataset (**Fig. S13**). The reconstructed Raman scattering spectra under different order mM were referred to **Fig. S14**. Further, stepwise regression was used to select valid independent moments variables and develop classification models for the discrimination of vibrational spectroscopy dataset.

5.2. GWO parameter

GWO as a novel SI optimization algorithm, which has good performance in global search and convergence. The core idea of GWO algorithm is to simulate various behavior of grey wolves, including the hierarchy and hunting process within the wolf population, to find the optimal solution of the target problem. The grey wolf population in nature is divided into four grades, namely α, β, δ and ω, in order of social status from high to low. Define the current optimum solution in the wolf population as α wolf, the second-best solution as β wolf, the third-best solution as δ wolf, and other solutions as ω wolf to construct the hierarchy model of the grey wolf. The hunting behavior of the whole population consists of three steps: (1) Tracking the prey; (2) Encircling the prey; (3) Attacking the prey. In the GWO algorithm, the hunting task is performed by α, β and δ wolf. ω wolf follows the three wolves to carry out the prey tracking, encirclement, and suppression.

It is worth mentioning that the GWO algorithm has very few parameters to be finetuned compared to other meta-heuristics methods according to its mechanism detailed in the **Supplementary Materials** section 4. The adaptive values of GWO parameters allow a smooth transition between exploration and exploitation. This provides GWO a greater ability to avoid stagnation in local optimal solution and converge quickly. Therefore, the motivation of this study is to explore the feasibility of the hybrid classification model of GWO-SVM in predicting vibrational spectroscopy. Moreover, the original GWO algorithm is applicable for continuous single objective optimization problems. However, in this study, two SVM hyperparameter (penalty factor C and RBF kernel parameter γ) combination selection is inherently multi-objective. Herein, we proposed an improved strategy namely multi-objective binary GWO algorithm. Consequently, it has been modified and represented in a way that is suitable for parameter selection task. The proposed strategy can avoid stagnation in local optima by maintaining a balance between exploration and exploitation.

Based on the above, GWO algorithm will seek for the hyperparameter combinations in a more reasonable space under the restriction of the upper and lower bounds. Meanwhile, the optimal solution of GWO was affected by the maximum iteration numbers. Here, the lower bound was set as 1.0×10^{-4} for C and 1.0×10^{-2} for γ, while the upper bound was set as 1.0×10^{4} for C and 1.0×10^{2} for γ. The maximum number of iterations was 30.

5.3. Model performance

In this section, the GWO-SVM model performance was compared with other ML algorithms. Simultaneously, the original model performance results from relevant literatures for these two vibrational spectra datasets were used as a reference to evaluate the performance for current study. It should be noticed that a one *versus* rest strategy [71] was implemented for serum Raman scattering spectra dataset to split the three-class classification task into three binary classification problem. Thus, for COVID-19 diagnosis, decision-making can be obtained by COVID-19 individuals *versus* healthy and suspected individuals. When evaluating the classification performance, confusion matrix can be introduced to judge whether the predicted labels of the model are consistent with the real one. The higher value in the upper left (TP) and lower right (TN) in the confusion matrix, the higher the consistency between the predicted results and the real labels. According to the optimal hyperparameters of several ML models, the confusion matrix of the six classification models for independent test phase from two vibrational spectra datasets can be obtained (**Fig. S15-16**). It can be noted that the TP and TN values of GWO-SVM were both higher than other ML models, especially for GS-SVM, suggesting that the predicted labels of the GWO-SVM for test phase were almost consistent with the real labels. Specifically, for the unknown vibrational spectra, the presented GWO-SVM model provided an accuracy, specificity and F1_score value of 0.9825, 0.9714 and 0.9778 for saliva FTIR spectra dataset, respectively, while an overall accuracy, specificity and F1_score value of 0.9085, 0.9552 and 0.9036 for serum Raman scattering spectra dataset, respectively, which showed superiority than those of state-of-the-art models, thereby suggesting the suitability of the GWO-SVM model to be adopted in a clinical setting for initial screening of COVID-19 patients.

In summation, the focus of this work is to explore the feasibility and reliability of the newly proposed GWO-SVM model in classifying vibrational spectroscopy of COVID-19 patients with other ML models as a reference. According to the obtained results, the discussed vibrational spectroscopy in combination with GWO-SVM model strategy can be applied for COVID-19 detection to improve the accuracy of provisional and clinical diagnosis. The proposed technique demonstrates better or comparable results with respect to those other ML techniques. Further studies may include larger cohort or large-scale multicenter trials to prove its applicability in clinical settings and to demonstrate joint applicability of vibrational spectroscopy and ML models for biomedical scenarios.

6. Conclusion

The identification of COVID-19 patients in its early stages, where treatment could provide maximum therapeutic benefits, is not only likely to slow down disease progression but also to potentially provide a cure. In the current study, we have successfully developed a vibrational spectroscopy-based approach for COVID-19 diagnosis and introduced a new ML algorithm, namely GWO-SVM where GWO was used to finetune the hyperparameters of SVM. Simultaneously, two vibrational spectra datasets were selected to train and evaluate the performance of GWO-SVM and other ML models. Through the analysis of figures of merits, it can be noted that the GWO-SVM exhibited superiority to classify vibrational spectra dataset than others ML models. In addition, the proposed GWO-SVM model was applicable to both binary and multiclassification problems. As a result, the reported saliva FTIR spectra and serum Raman scattering spectra examination have the potential to complement clinical nucleic acid testing, make early COVID-19 detection quickly, accurate, and inexpensive. The results indicate that vibrational spectroscopy coupled with GWO-SVM can be employed as adjuvant or alternative approach in the clinical diagnosis of COVID-19 patients. While this study showed promise using a small sample set, further method validation on a large scale is required to indicate the true strength of the proposed strategy.

Data and Code availability

All study data were included in the article or **Supplementary Material**. Custom-built codes may be available by contacting the corresponding author on reasonable request.

Ethical approval statement

Not required for this study because no human or animals directly participated in this study.

Declaration of Competing Interest

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

Acknowledgements

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2022.107295.

References

[1] Z. Liao, Y. Song, S. Ren, et al., VOC-DL: Deep learning prediction model for COVID-19 based on VOC virus variants, Comput. Methods Programs Biomed. 224 (2022) 106981, doi:10.1016/j.cmpb.2022.106981.

[2] G.R. Shinde, A.B. Kalamkar, P.N. Mahalle, et al., Data Analytics for Pandemics: A COVID-19 Case Study, CRC Press, 2020, doi:10.1201/9781003095415.

[3] M. Shen, Y. Zhou, J. Ye, et al., Recent advances and perspectives of nucleic acid detection for coronavirus, J. Pharm. Anal. 10 (2020) 97–101, doi:10.1016/j.jpha.2020.02.010.

[4] H. Hassan, Z. Ren, C. Zhou, et al., Supervised and weakly supervised deep learning models for COVID-19 CT diagnosis: a systematic review, Comput. Methods Programs Biomed. 218 (2022) 106731, doi:10.1016/j.cmpb.2022.106731.

[5] Y. Chen, Y. Lin, X. Xu, et al., Classification of lungs infected COVID-19 images based on inception-ResNet, Comput. Methods Programs Biomed. 225 (2022) 107053, doi:10.1016/j.cmpb.2022.107053.

[6] A.I. Khan, J.L. Shah, M.M. Bhat, CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, Comput. Methods Programs Biomed. 196 (2020) 105581, doi:10.1016/j.cmpb.2020.105581.

[7] G. Wang, X. Liu, J. Shen, et al., A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images, Nat. Biomed. Eng. 5 (2021) 509–521, doi:10.1038/s41551-021-00704-1.

[8] L. Brunese, F. Mercaldo, A. Reginelli, et al., Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays, Comput. Methods Programs Biomed. 196 (2020) 105608, doi:10.1016/j.cmpb.2020.105608.

[9] M.A. MacMullan, A. Ibrayeva, K. Trettner, et al., ELISA detection of SARS-CoV-2 antibodies in saliva, Sci. Rep. 10 (2020) 20818, doi:10.1038/s41598-020-77555-4.

[10] Q. Yin, Y. Zhang, L. Lian, et al., Chemiluminescence immunoassay based serological immunoassays for detection of SARS-CoV-2 neutralizing antibodies in COVID-19 convalescent patients and vaccinated population, Viruses (2021) 13, doi:10.3390/v13081508.

[11] S.I. Owen, C.T. Williams, G. Garrod, et al., Twelve lateral flow immunoassays (LFAs) to detect SARS-CoV-2 antibodies, J. Infect. (2021), doi:10.1016/j.jinf.2021.12.007.

[12] Y. Peng, Y. Pan, Z. Sun, et al., An electrochemical biosensor for sensitive analysis of the SARS-CoV-2 RNA, Biosens. Bioelectron. 186 (2021) 113309, doi:10.1016/j.bios.2021.113309.

[13] F.M. Nachtigall, A. Pereira, O.S. Trofymchuk, et al., Detection of SARS-CoV-2 in nasal swabs using MALDI-MS, Nat. Biotechnol. 38 (2020) 1168–1173, doi:10.1038/s41587-020-0644-7.

[14] J. Delafiori, L.C. Navarro, R.F. Siciliano, et al., Covid-19 Automated diagnosis and risk assessment through metabolomics and machine learning, Anal. Chem. 93 (2021) 2471–2479, doi:10.1021/acs.analchem.0c04497.

[15] L. Yan, J. Yi, C. Huang, et al., Rapid detection of COVID-19 Using MALDI-TOF-based serum peptidome profiling, Anal. Chem. 93 (2021) 4782–4787, doi:10.1021/acs.analchem.0c04590.

[16] N.M. Ralbovsky, I.K. Lednev, Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning, Chem. Soc. Rev. 49 (2020) 7428–7453, doi:10.1039/d0cs01019g.

[17] N. AlMasoud, H. Muhamadali, M. Chisanga, et al., Discrimination of bacteria using whole organism fingerprinting: the utility of modern physicochemical techniques for bacterial typing, Analyst 146 (2021) 770–788, doi:10.1039/d0an01482f.

[18] J.M. Duarte, N.G.S. Sales, J.W.B. Braga, et al., Discrimination of white automotive paint samples using ATR-FTIR and PLS-DA for forensic purposes, Talanta 240 (2021) 123154, doi:10.1016/j.talanta.2021.123154.

[19] S. Yan, S. Wang, J. Qiu, et al., Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level, Talanta 226 (2021) 122195, doi:10.1016/j.talanta.2021.122195.

[20] K. Rebrošová, S. Bernatová, M. Šiler, et al., Raman spectroscopy—a tool for rapid differentiation among microbes causing urinary tract infections, Anal. Chim. Acta (2021), doi:10.1016/j.aca.2021.339292.

[21] I.A. Bratchenko, L.A. Bratchenko, Y.A. Khristoforova, et al., Classification of skin cancer using convolutional neural networks analysis of Raman spectra, Comput. Methods Programs Biomed. 219 (2022) 106755, doi:10.1016/j.cmpb.2022.106755.

[22] C. Shu, H. Yan, W. Zheng, et al., Deep learning-guided fiberoptic raman spectroscopy enables real-time in vivo diagnosis and assessment of nasopharyngeal carcinoma and post-treatment efficacy during endoscopy, Anal. Chem. 93 (2021) 10898–10906, doi:10.1021/acs.analchem.1c01559.

[23] D. Cialla-May, C. Krafft, P. Rosch, et al., Raman spectroscopy and imaging in bioanalytics, Anal. Chem. 94 (2022) 86–119, doi:10.1021/acs.analchem.1c03235.

[24] S.J. Fong, N. Dey, J. Chaki, Artificial Intelligence for Coronavirus Outbreak, Springer, 2021, doi:10.1007/978-981-15-5936-5.

[25] K. Mehta, A. Atak, A. Sahu, et al., An early investigative serum Raman spectroscopy study of meningioma, Analyst 143 (2018) 1916–1923, doi:10.1039/c8an00224j.

[26] D. Ami, A. Duse, P. Mereghetti, et al., Tear-based vibrational spectroscopy applied to amyotrophic lateral sclerosis, Anal. Chem. 93 (2021) 16995–17002, doi:10.1021/acs.analchem.1c02546.

[27] M. Paraskevaidi, C.L.M. Morais, D.L.D. Freitas, et al., Blood-based near-infrared spectroscopy for the rapid low-cost detection of Alzheimer's disease, Analyst 143 (2018) 5959–5964, doi:10.1039/c8an01205a.

[28] Q. Li, W. Li, J. Zhang, et al., An improved k-nearest neighbour method to diagnose breast cancer, Analyst 143 (2018) 2807–2811, doi:10.1039/c8an00189h.

[29] H. Hartatik, M.B. Tamam, A. Setyanto, in: 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1–5.

[30] S.F. Abdoh, M. Abo Rizka, F.A. Maghraby, Cervical cancer diagnosis using random forest classifier With SMOTE and feature reduction techniques, IEEE Access 6 (2018) 59475–59485, doi:10.1109/access.2018.2874063.

[31] B.H. Do, C. Langlotz, C.F. Beaulieu, Bone tumor diagnosis using a naive Bayesian model of demographic and radiographic features, J. Digit. Imaging 30 (2017) 640–647, doi:10.1007/s10278-017-0001-7.

[32] Y. Yang, Y. Yang, Z. Liu, et al., Microcalcification-based tumor malignancy evaluation in fresh breast biopsies with hyperspectral stimulated raman scattering, Anal. Chem. 93 (2021) 6223–6231, doi:10.1021/acs.analchem.1c00522.

[33] Y.T. Lin, C.Y. Chu, K.S. Hung, et al., Can machine learning predict pharmacotherapy outcomes? An application study in osteoporosis, Comput. Methods Programs Biomed. 225 (2022) 107028, doi:10.1016/j.cmpb.2022.107028.

[34] H. Yang, X. Li, H. Cao, et al., Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data, Comput. Methods Programs Biomed. 211 (2021) 106420, doi:10.1016/j.cmpb.2021.106420.

[35] Z. Guleken, P. Jakubczyk, P. Wieslaw, et al., Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy, and machine learning classifications, Talanta 237 (2022) 122916, doi:10.1016/j.talanta.2021.122916.

[36] Z. Guleken, Y. Tuyji Tok, P. Jakubczyk, et al., Development of novel spectroscopic and machine learning methods for the measurement of periodic changes in COVID-19 antibody level, Measurement 196 (2022) 111258, doi:10.1016/j.measurement.2022.111258.

[37] C. Mazo, E. Alegre, M. Trujillo, Classification of cardiovascular tissues using LBP based descriptors and a cascade SVM, Comput. Methods Programs Biomed. 147 (2017) 1–10, doi:10.1016/j.cmpb.2017.06.003.

[38] Y. Wang, J. Xu, D. Cui, et al., Classification and identification of archaea using single-cell Raman ejection and artificial intelligence: implications for investigating uncultivated microorganisms, Anal. Chem. 93 (2021) 17012–17019, doi:10.1021/acs.analchem.1c03495.

[39] E. Ryzhikova, N.M. Ralbovsky, V. Sikirzhytski, et al., Raman spectroscopy and machine learning for biomedical applications: Alzheimer's disease diagnosis based on the analysis of cerebrospinal fluid, Spectrochim. Acta A 248 (2021) 119188, doi:10.1016/j.saa.2020.119188.

[40] N. Dey, Advancements in Applied Metaheuristic Computing, IGI Global, 2017.

[41] Tang, R.; Fong, S.; Dey, N., Metaheuristics and chaos theory. 2018, 182-196. doi:10.5772/intechopen.72103.

[42] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey Wolf optimizer, Adv. Eng. Softw. 69 (2014) 46–61, doi:10.1016/j.advengsoft.2013.12.007.

[43] S. Mirjalili, How effective is the Grey Wolf optimizer in training multi-layer perceptrons, Appl. Intell. 43 (2015) 150–161, doi:10.1007/s10489-014-0645-7.

[44] O. Vandenberg, D. Martiny, O. Rochas, et al., Considerations for diagnostic COVID-19 tests, Nat. Rev. Microbiol. 19 (2021) 171–183, doi:10.1038/s41579-020-00461-z.

[45] K. Dighe, P. Moitra, M. Alafeef, et al., A rapid RNA extraction-free lateral flow assay for molecular point-of-care detection of SARS-CoV-2 augmented by chemical probes, Biosens. Bioelectron. 200 (2022) 113900, doi:10.1016/j.bios.2021.113900.

[46] D. Zhang, Y. Guo, L. Zhang, et al., Integrated system for on-site rapid and safe screening of COVID-19, Anal. Chem. 94 (2022) 13810–13819, doi:10.1021/acs.analchem.2c02337.

[47] P. Zhang, L. Chen, J. Hu, et al., Magnetofluidic immuno-PCR for point-of-care COVID-19 serological testing, Biosens. Bioelectron. 195 (2022) 113656, doi:10.1016/j.bios.2021.113656.

[48] X. Mei, H.C. Lee, K.Y. Diao, et al., Artificial intelligence-enabled rapid diagnosis of patients with COVID-19, Nat. Med. 26 (2020) 1224–1228, doi:10.1038/s41591-020-0931-3.

[49] M.J. Baker, H.J. Byrne, J. Chalmers, et al., Clinical applications of infrared and Raman spectroscopy: state of play and future challenges, Analyst 143 (2018) 1735–1757, doi:10.1039/c7an01871a.

[50] C.L.M. Morais, K.M.G. Lima, M. Singh, et al., Tutorial: multivariate classification for vibrational spectroscopy in biological samples, Nat. Protoc. 15 (2020) 2143–2162, doi:10.1038/s41596-020-0322-8.

[51] M.H.C. Nascimento, W.D. Marcarini, G.S. Folli, et al., Noninvasive diagnostic for COVID-19 from Saliva biofluid via FTIR spectroscopy and multivariate analysis, Anal. Chem. 94 (2022) 2425–2433, doi:10.1021/acs.analchem.1c04162.

[52] Y. Du, D. Han, S. Liu, et al., Raman spectroscopy-based adversarial network combined with SVM for detection of foodborne pathogenic bacteria, Talanta 237 (2022) 122901, doi:10.1016/j.talanta.2021.122901.

[53] B.R. Wood, K. Kochan, D.E. Bedolla, et al., Infrared Based Saliva Screening Test for COVID-19, Angew. Chem. Int. Ed Engl. 60 (2021) 17102–17107, doi:10.1002/anie.202104453.

[54] G. Yin, L. Li, S. Lu, et al., An efficient primary screening of COVID-19 by serum Raman spectroscopy, J. Raman Spectrosc. (2021), doi:10.1002/jrs.6080.

[55] R.W.K.A.J. Stone, Comput. Aided Des. Exp. 11 (1969) 137–148.

[56] M.J. Baker, S.R. Hussain, L. Lovergne, et al., Developing and understanding biofluid vibrational spectroscopy: a critical review, Chem. Soc. Rev. 45 (2016) 1803–1818, doi:10.1039/c5cs00585j.

[57] A. Savitzky, M.J.J.A.c. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639, doi:10.1021/ac60214a047.

[58] P. Geladi, D. MacDougall, H. Martens, Linearization and Scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491–500, doi:10.1366/0003702854248656.

[59] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemom. Intell. Lab. Syst. 82 (2006) 59–65, doi:10.1016/j.chemolab.2005.08.009.

[60] C.L.M. Morais, M. Paraskevaidi, L. Cui, et al., Standardization of complex biologically derived spectrochemical datasets, Nat. Protoc. 14 (2019) 1546–1577, doi:10.1038/s41596-019-0150-x.

[61] S.S. Li, B. Yin, H.L. Zhai, et al., An effective approach to the quantitative analysis of skin-whitening agents in cosmetics with different substrates based on conventional UV-Vis determination, Anal. Methods 11 (2019) 1500–1507, doi:10.1039/c9ay00007k.

[62] B. Yin, H.L. Zhai, B.Q. Zhao, et al., Chemometrics-assisted simultaneous voltammetric determination of multiple neurotransmitters in human serum, Bioelectrochemistry 139 (2021) 107739, doi:10.1016/j.bioelechem.2021.107739.

[63] S.H. Lu, H.L. Zhai, B.Q. Zhao, et al., Novel approach to the analysis of chemical third-order data, J. Chem. Inf. Model. 60 (2020) 4750–4756, doi:10.1021/acs.jcim.0c00554.

[64] E.M. El-Kenawy, A. Ibrahim, S. Mirjalili, et al., Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images, IEEE Access 8 (2020) 179317–179335, doi:10.1109/ACCESS.2020.3028012.

[65] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (2009) 427–437, doi:10.1016/j.ipm.2009.03.002.

[66] T.Y. Huang, J.C.C. Yu, Development of crime scene intelligence using a handheld raman spectrometer and transfer learning, Anal. Chem. 93 (2021) 8889–8896, doi:10.1021/acs.analchem.1c01099.

[67] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874, doi:10.1016/j.patrec.2005.10.010.

[68] J. Li, J.P. Fine, ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies, Biostatistics 9 (2008) 566–576, doi:10.1093/biostatistics/kxm050.

[69] S.S. Chandrasekaran, S. Agrawal, A. Fanton, et al., Rapid detection of SARS-CoV-2 RNA in saliva via Cas13, Nat. Biomed. Eng. (2022), doi:10.1038/s41551-022-00917-y.

[70] J. Xu, Y. Li, F. Gan, et al., Salivary glands: potential reservoirs for COVID-19 asymptomatic infection, J. Dent. Res. 99 (2020) 989, doi:10.1177/0022034520918518.

[71] M.A. Kumar, M. Gopal, Reduced one-against-all method for multiclass SVM classification, Expert Syst. Appl. 38 (2011) 14238–14248, doi:10.1016/j.eswa.2011.04.237.