Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2024.00144

Research Article

Detection of COVID-19 infection by using Raman spectroscopy of serum samples and machine learning

Fatma UYSAL ÇİLOĞLU[1,*]

[1]*Department of Biomedical Engineering, Erciyes University, Kayseri, 38039, Türkiye*

ARTICLE INFO

Article history
Received: 03 August 2023
Revised: 28 September 2023
Accepted: 22 October 2023

Keywords:
COVID-19, Decision Tree; K-Nearest Neighbors; Raman Spectroscopy, Principal Component Analysis, Support Vector Machine

ABSTRACT

Rapid, simple, and accurate detection is important to slow down the spread of epidemics in the era of the pandemic. COVID-19 has been the biggest epidemic of the current century and continues. Therefore, it is extremely important to develop methods that allow the detection of COVID-19 by eliminating the disadvantages of the existing methods. The aim of this study is to perform rapid and reliable detection of COVID-19 using Raman spectroscopy and machine learning techniques. Here, Raman spectra of serum samples collected from COVID-19 patients, suspected cases, and healthy controls were utilized. Machine learning techniques were employed due to the absence of significant discernible variations between the Raman spectra of the three groups with the naked eye. Therefore, principal component analysis (PCA) was utilized to reveal discriminative features of the classes. Support vector machine (SVM), k-nearest neighbors (KNN), and decision tree (DT) classification models were utilized by using extracted features with PCA. SVM and KNN provide high accuracy ± standard deviation values of 86.5±0.7% and 87.3±0.6% respectively. Sensitivity (recall), precision, and area under the curve (AUC) which are important performance evaluation metrics were also calculated for comparison. Results show that Raman spectra combined with machine learning presents a promising tool for the accurate detection of COVID-19 in clinical use.

Cite this article as: Uysal Çiloğlu F. Detection of COVID-19 infection by using Raman spectroscopy of serum samples and machine learning. Sigma J Eng Nat Sci 2024;42(6):1892–1898.

INTRODUCTION

Over the last two decades, there have been three significant outbreaks caused by human coronaviruses. These outbreaks involved the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002, the Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012, and the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) in 2019. The global spread of coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 that emerged in Wuhan, China has rapidly spread all over the world and become a worldwide pandemic [1, 2]. More than 675 million people have infected and about 6.5 million deaths have occurred globally since the day COVID-19 appeared according to the Center for Systems Science and Engineering (CSSE) at John Hopkins University [3]. This epidemic not only damaged the health systems of countries

*Corresponding author.
*E-mail address: fatmauysal@erciyes.edu.tr

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

but also greatly affected economies. The most important thing in keeping the epidemic under control is to perform rapid, accurate and reliable tests. These tests are necessary for both diagnosis and tracking.

The gold standard to identification of SARS-CoV-2 is real-time reverse transcriptase-polymerase chain reaction (rRT-PCR) test [4]. Although being the most validated nucleic acid-based test for detecting COVID-19, this method exhibits certain limitations. The rRT-PCR test encompasses a multi-step approach that involves nucleic acid extraction, amplification, and purification. Moreover, it requires advanced equipment, high-cost reagents, trained personnel, suffers from false-negative results and being time-consuming. Immunoassay based techniques such as enzyme-linked immunosorbent assay (ELISA) are also used for rapid detection. However, this kind of immunoassays are not suitable for the identification of specific viral species or strains since they depend on specific antigen/antibody binding. Further, these techniques are destructive for the samples. Hence, conventional methodologies are unsuitable for rapid detection due to the aforementioned constraints. There exists an unaddressed requirement for novel technologies characterized by streamlined procedures and enhanced sensitivity.

Raman Spectroscopy is a robust scientific technique founded upon the principle of inelastic scattering of a laser beam. This method yields invaluable molecular fingerprint information regarding the specimen, whereby each molecule manifests its distinct and unique Raman spectral signal. Moreover, Raman spectroscopy requires a minimum sample preparation process and nondestructive [5]. However, this technique has some limitations due to the weak signal intensity of the Raman scattering [6]. This situation makes harder the interpretation and discrimination of similar spectra. To overcome this problem using machine learning methods for the detection or discrimination of biological samples is essential.

Raman spectroscopy applications span a broad range in microbial identification in the biomedical field [7–9]. However, there are few studies reporting the use of Raman spectroscopy for the detection of SARS-CoV-2. Carlomagno et al. reported a Raman-based approach to detect current and past infections of SARS-CoV-2 from saliva samples [10]. Zhang et al. used surface enhanced Raman spectroscopy array functionalized with ACE2 to capture SARS-CoV-2 spike protein to test environmental species [11]. Akdeniz et al. developed a virus-infected cellular model by transfecting mammalian cells with plasmids encoding the M, N, and E proteins of SARS-CoV-2 [12]. They isolated proteins from the cells and collected their spectra. Results show that there is a clear discrimination between proteins of SARS-CoV-2 and the control group according to the principal component analysis (PCA).

In this study, the discrimination of Raman spectra of serum samples collected from COVID-19 patients, suspected cases, and healthy individuals obtained from a publicly available dataset [13] was performed. To reveal the discriminative features, PCA was used. Afterward, some machine learning techniques such as support vector machine (SVM), k-nearest neighbors (KNN), and decision tree (DT) were applied. Until now, the conducted studies by using this dataset have primarily focused on categorizing spectra from COVID-19, suspected cases, and healthy subjects into two classes [13–15]. Yin et al. used the whole dataset, and they classified the COVID-19 patients, suspected cases, and healthy controls into the combination of two classes [13]. They found accuracy values of COVID-19 versus suspected, COVID-19 versus healthy control, and suspected versus healthy control as 0.87±0.05, 0.91±0.04, and 0.69±0.05 respectively. Deepaisarn et al. used the same dataset and discriminated COVID-19 patients and healthy controls by using different feature extraction methods and classifiers [14]. They obtained the accuracy of 98.38% by using Light Gradient Boosting Model. Wei et al. also used this dataset to classify COVID-19 and healthy subjects and found accuracy of 97.9±0.2% by using Multi-Scale Sequential feature Selection model (M3S) [15]. However, classifying COVID-19, suspected cases, and healthy spectra simultaneously into three distinct classes would be more clinically meaningful as a classification problem. Since the spectrum of the serum sample obtained from an individual will fall into one of these three categories. Consequently, after training the machine learning model, when the spectrum from an individual is provided as input, the model will be able to ascertain to which group (COVID-19, suspected, or healthy) the person belongs. Here, the Raman spectra of all three classes were analyzed simultaneously, resulting in the successful identification of the respective class to which the serum samples belonged.

MATERIALS AND METHODS

The Dataset

Here, a portion of publicly available dataset was used [13]. The original dataset includes 2655 Raman spectra of serum samples collected from 177 individuals including confirmed COVID-19 patients, suspected cases, and healthy controls. However, the accessible part of the dataset includes 465 spectra belong to 159 spectra from COVID-19 patients, 156 spectra from suspected cases, and 150 spectra from healthy controls. Every individual diagnosed with COVID-19 exhibited positive results in viral nucleic acid detection through real-time polymerase chain reaction (RT-PCR) utilizing respiratory tract samples. The suspected group indicated flu-like symptoms similar to COVID-19. Each spectrum was recorded by exciting with a laser 785 nm wavelength under the 70-mW laser power. All spectral data was used in the range of 600-1800 cm^{-1}.

Data Analysis

The dataset consists of 465 spectra. Firstly, the whole data was shuffled randomly and standardized by using

standard normal variate. To extract discriminative spectral features from Raman data PCA is a statistical technique often used [16, 17]. PCA is used for dimensionality reduction and data exploration. PCA is commonly employed as a technique to decrease the dimensionality of a dataset by generating a reduced set of variables that effectively capture the majority of the variability found in the original data. This enables the simplification of intricate datasets while still preserving crucial information [18]. PCA extracts linear features by detecting linear relationships from the data. It projects the data to a new coordinate space so that the variance within the data remains maximum. The first principal component has the highest variance and the variance that is carried by each principal component decreases through the last principal component. Here, the first 40 principal components that hold a large amount of variance (> 95%) in the data were used.

To discriminate COVID-19 patients, suspected cases, and healthy controls traditional classifiers were utilized. For this purpose, traditional machine learning techniques that are preferred widely for the classification of Raman spectra were used. SVM, KNN, and DT were utilized for the discrimination of COVID-19 patients, suspected cases, and healthy controls. The Support Vector Machine (SVM) is a supervised machine learning algorithm employed for tasks related to classification and regression. Its operation involves the identification of the optimal hyperplane within a high-dimensional feature space, effectively segregating data points belonging to distinct classes [19]. Multiclass SVM classifier polynomial kernel (degree: 3) with one-versus-all approach was selected. KNN classification is a supervised machine learning approach employed to categorize data points into distinct classes by determining the majority class among their nearest neighbors [19]. In KNN classifier, the most appropriate k parameter was determined as 1 and Euclidean distance was used. DT classification algorithm builds a tree-like structure to make decisions about the class labels of data points based on their features [19]. Random search method was used for the hyperparameter optimization. The evaluation of each classifier's performance was conducted by assessing the accuracy, sensitivity (recall), and precision which were derived from the respective confusion matrices. Furthermore, area under curve (AUC) value was also used for performance evaluation. K-fold cross-validation helps in preventing overfitting by providing a more robust evaluation of a model's performance, reducing the influence of a single data split, and ensuring that the model learns from a diverse set of training data. Consequently, this aids in the construction of models that exhibit improved generalization when confronted with novel, unseen data. The 10-fold cross-validation technique was employed to objectively assess the performance of the classifiers. In this method, the dataset is divided into ten groups of approximately equal sizes. Nine of these subsets are designated for training, while one subset is reserved for testing. The aforementioned process is repeated 10 times, ensuring that every subset has been utilized as the testing set. All these procedures were repeated 50 times to calculate mean ± standard deviation. All data processing procedures were performed using MATLAB software (The MathWorks, Natick, USA).

RESULTS AND DISCUSSION

Rapid and accurate diagnosis is extremely important to slow down the spread of epidemics. Therefore, there is a pressing demand for emerging technologies that enable rapid diagnosis. In this study, Raman spectra of serum samples collected from COVID-19 patients, suspected cases, and healthy controls were discriminated by using machine learning techniques.

Here a total of 465 spectra from COVID-19 patients, suspected cases, and healthy controls were used to discriminate these 3 groups. As seen in Figure 1, there is a high similarity between Raman spectra of all groups. SARS-CoV-2 could potentially manifest a distinct protein profile, setting it apart from both healthy samples and other diseases [20]. Furthermore, the pathological progression of COVID-19 triggers the initiation of a humoral response, resulting in the production of particular antibodies [21]. These processes make the COVID-19 serum spectra biochemically different from other spectra. The differences of the serum composition between groups can be detected by Raman spectroscopy. Thus, there are noticeable distinctions among spectral groups, but discerning these spectra with the naked eye appears to be exceedingly challenging. Therefore, the utilization of machine learning techniques is essential for addressing this issue.

PCA were firstly used for dimensionality reduction. This step is crucial because the optimum number of features are directly related to classifier performance. The use of all variables in spectral data will increase model complexity and training time. On the other hand, inadequate feature selection will lead to a decrease in classifier performance, resulting in poor classification accuracy. Therefore, choosing the optimal number of features is a key step for

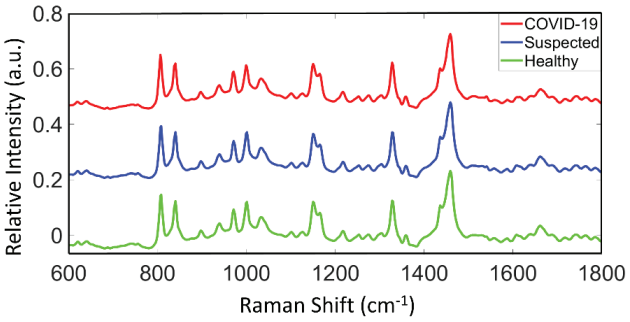

Figure 1. Mean serum Raman spectra of COVID-19 patients, suspected cases, and healthy controls.

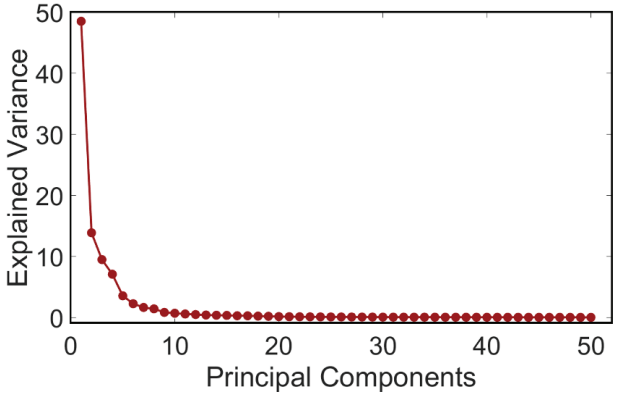

Figure 2. Explained variance of the first 50 principal components found by applying PCA to the dataset.

a classification problem. Here, PCA was performed for extracting linear features from the dataset. Figure 2 shows the variances holding by each principal component. First principal component has the maximum variance, and the variance decreases from the first principal component to last principal component. The first 20 principal components carry most of the variance in the dataset and variance does not change much after these 20 components as seen in Figure 2.

In this study, the first 40 principal components that carry the largest amount of variance (>95%) in the data set were used. These principal components were fed into the classifiers. Two and three dimensional PCA plots for COVID-19 patients, suspected cases, and healthy controls are given in Figure 3A and B, respectively. There is not a clear discrimination for two and three dimensional PCA space as seen in these figures. However, the data given to the classifiers are located in a 40-dimensional PCA space not in 2 or 3 dimensional spaces. PCA can effectively reveal patterns in the dataset for this reason first 40 principal components extracted with PCA were used as input variables for classification. Some classification algorithms (SVM, KNN, and DT) were performed to distinguish three groups of COVID-19 patients, suspected cases, and healthy controls. At this step, 40 principal components were used as features for all classifiers. The 10-fold cross-validation technique was employed to provide an objective evaluation of the classifier performances.

The mean accuracies with standard deviation were calculated by using 50 randomly distributed sets. Figure 4 demonstrates accuracies of each classifier through independent 50 runs. The SVM classifier exhibits mean accuracy value of 86.5%±0.7 while KNN classifier ensures the slightly better mean accuracy of 87.3%±0.6 than the SVM classifier. On the other hand, the DT classifier gives the worst mean accuracy of 69.2±1.9 among others. Generally, KNN and SVM classifiers provide better results for the classification of spectral dataset [17].

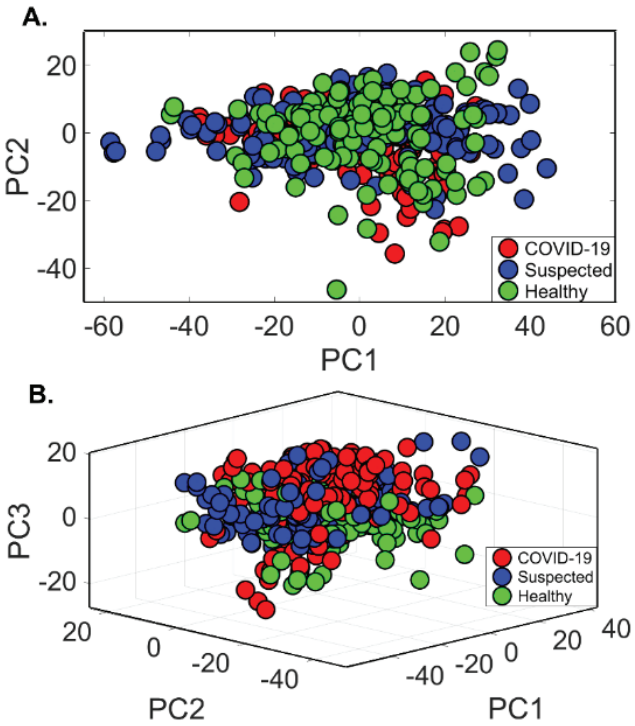

Figure 3. PCA scores of serum Raman spectra of COVID-19 patients, suspected cases, and healthy controls. Score plot of PC1 vs. PC2 (A). Score plot of PC1 vs. PC2 and PC3 (B).

Although determining the classifier performance calculation of mean accuracy is the frequently used method it is absolutely not enough. Other metrics are needed to objectively evaluate the classifier performance. The most used metrics are sensitivity, precision, and AUC. Accuracy, sensitivity, and precision values are derived from the confusion matrix through calculations. The confusion matrix serves as a vital performance evaluation tool within the domains of machine learning and statistics. It offers a more in-depth perspective on a classification model's performance when compared to accuracy alone. This is because it considers both false positives and false negatives, which can be pivotal in real-world applications. It is an essential tool for evaluating the effectiveness of classification models and comparing different models or tuning their parameters. Figure 5 a-c shows the confusion matrices of SVM, KNN, and DT classifiers respectively. Within this matrix, each row signifies the quantity of spectra in a true class, while each column signifies the number of spectra in a predicted class. The elements along the main diagonal indicate the number of correctly classified spectra, whereas the off-diagonal elements indicate the number of misclassified spectra. Here, SVM and KNN classifiers yield more successful results than DT in distinguishing classes. However, KNN provides slightly better results than SVM for the discrimination of COVID-19 and suspected classes. On the other hand, SVM exhibits more successful results in distinguishing the healthy class

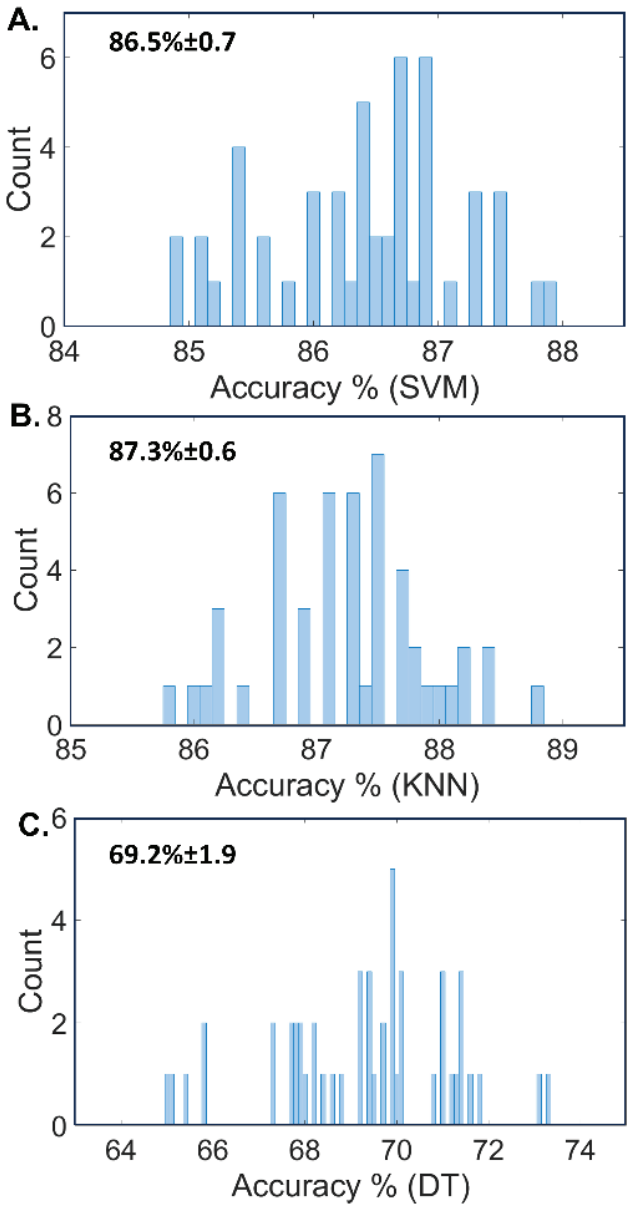

Figure 4. Accuracy values of **A.** SVM, **B.** KNN **C.** DT classifiers with 50 randomly distributed sets.

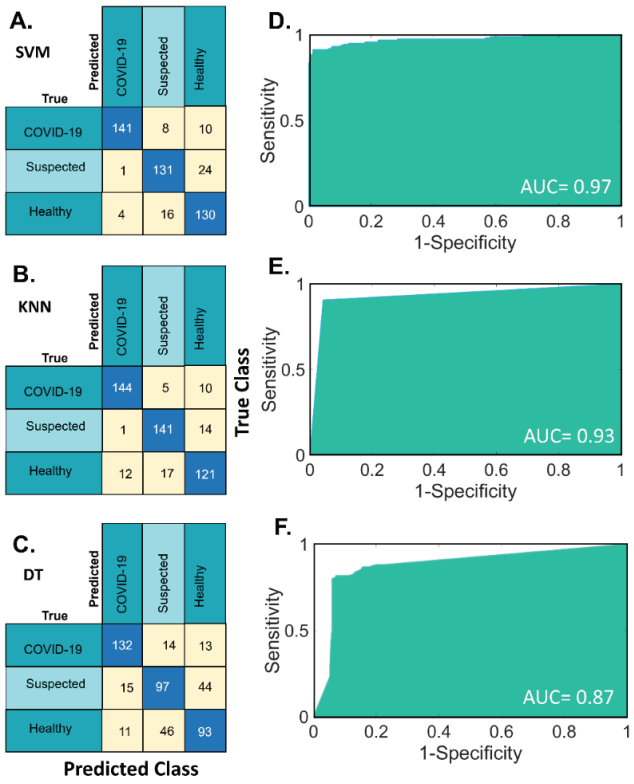

Figure 5. Confusion matrices and receiver operating curves of SVM, KNN, and DT classifiers. **A.** Confusion matrix of SVM **B.** Confusion matrix of KNN **C.** Confusion matrix of DT **D.** ROC of SVM **E.** ROC of KNN **F.** ROC of DT.

compared to KNN. The accuracy, sensitivity, and precision values of classifiers are provided in Table 1. Sensitivity (recall) quantifies the model's capacity to accurately recognize positive instances out of all the actual positive instances in the dataset. Especially in medical applications, the cost of missing a true positive (a disease) can be critical as it may delay treatment or lead to severe consequences for the patient. Therefore, high sensitivity is desirable for medical applications. As seen in Table 1, KNN classifier illustrates higher sensitivity values for COVID-19 and suspected cases. On the other hand, SVM shows better sensitivity for healthy group. However, it is important to keep in mind that sensitivity should be considered in conjunction with other performance metrics, such as precision, and accuracy. Precision is another fundamental performance metric used to evaluate the quality of a classification model. The precision value of SVM model for the COVID-19 class is found better with 0.96 and other classes give similar results for SVM and KNN models. By contrast with SVM and KNN models, DT classifier gives the worst results for sensitivity, precision, and accuracy.

The AUC value is derived from the ROC (Receiver Operating Characteristic) curve, which plots sensitivity (y-axis) against 1-specificity (x-axis). It serves as a significant metric for evaluating classifier performance as it represents the area under the ROC curve [22, 23]. The ROC curve provides a visual representation of the balance between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various classification thresholds. The curve plots sensitivity (TPR) on the y-axis against 1-specificity (FPR) on the x-axis, and each point on the curve corresponds to a different threshold value. An optimal ROC curve closely follows the top-left corner, signifying high sensitivity and a low false positive rate across different threshold values. Here, the AUC values are found as 0.97, 0.93, and 0.87 for SVM, KNN, and DT classifiers respectively. Although, KNN classifier provides the highest accuracy, the AUC value of it lower than SVM classifier. A high

Table 1. Sensitivity, precision, and accuracy ± standard deviation values of SVM, KNN, and DT classifiers belong to COVID-19, suspected cases, and healthy controls.

Classifier	Classes	Sensitivity	Precision	Accuracy (%) ± standard deviation
	COVID-19	0.89	0.96	
SVM	suspected	0.84	0.84	86.5±0.7
	healthy	0.87	0.79	
	COVID-19	0.91	0.91	
KNN	suspected	0.9	0.86	87.3±0.6
	healthy	0.81	0.83	
	COVID-19	0.83	0.83	
DT	suspected	0.62	0.62	69.2±1.9
	healthy	0.62	0.62	

SVM: Support vector machine; KNN: k-nearest neighbors; DT: Decision tree.

accuracy value indicates that the classifier is making correct predictions overall, regardless of class. It calculates the ratio of correctly classified samples to the total number of samples within the dataset. High accuracy suggests that the model is good at making correct predictions for both positive and negative instances. A high AUC value, on the other hand, indicates that the classifier is good at distinguishing between positive and negative instances. The AUC represents the area under the ROC curve and measures the classifier's ability to rank instances correctly. A high AUC means that the classifier is capable of achieving high true positive rates (sensitivity) and low false positive rates (1-specificity) across various classification thresholds. Ultimately, the choice between the two classifiers depends on the specific requirements of the problem and the associated costs or implications of misclassifications. If correctly distinguishing positive instances is more critical (e.g., medical diagnosis), the classifier with a high AUC might be preferred.

Consequently, the classification of Raman spectra belonging to COVID-19, suspected cases, and healthy individuals' blood serum samples was performed. In this context, SVM and KNN yielded quite similar results, while the DT model provided the worst outcome. In some classes, SVM demonstrates superior performance concerning accuracy, sensitivity, and precision while KNN outperforms in other scenarios. Despite the KNN classifier yielding higher accuracy, in cases such as disease detection-based datasets, opting for the classifier with a higher AUC value would be more appropriate. Since a high AUC value shows high discrimination power between positive and negative classes. On the other hand, more data is needed to generalize these results. The main limitation of this study is limited access to the data. Making a greater portion of the dataset publicly available will enable machine learning models to achieve a more robust generalization of the results. Moreover, as the data volume increases, deep learning techniques that have the potential to provide superior outcomes compared to traditional machine learning models will also become available. As a result, although the results obtained are promising, more data are needed for generalization.

CONCLUSION

Rapid COVID-19 detection is of utmost importance as it allows for swift identification and isolation of infected individuals, curbing the spread of the virus within communities and preventing outbreaks. Timely detection also enables prompt medical intervention, reducing the severity of cases and ultimately saving lives. Raman spectroscopy can be a powerful alternative to conventional methods for the rapid detection of COVID-19. Here, the discrimination of COVID-19, suspected cases, and healthy controls serum Raman spectra were performed. Data preprocessing involved feature extraction using PCA, and classification was carried out using SVM, KNN, and DT classifiers. The performance of the classifiers on the test data, evaluated through 10-fold cross-validation, was assessed using accuracy, sensitivity, precision, and AUC values. Thanks to the machine learning techniques similar spectra were discriminated successfully. SVM and KNN algorithms provide high accuracy, sensitivity, and precision results for the discrimination of COVID-19, suspected cases, and healthy controls. In summary, Raman Spectroscopy and machine learning emerge as a dependable and robust approach for distinguishing COVID-19 cases rendering it highly prospective for clinical implementations.

ACKNOWLEDGMENTS

I would like to thank Yin G. et al. to provide open-source dataset of COVID-19 SERS spectra of serum samples.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request. The COVID-19 data used in this study is publicly available at: https://doi.org/10.6084/m9.figshare.12159924.v1

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrob Agents 2020;55:105924. [CrossRef]

[2] Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: What we know. Int J Infect Dis 2020;94:44–48. [CrossRef]

[3] COVID-19 Map. Available at: https://coronavirus.jhu.edu/map.html Last Accessed Date: 02.08.2023

[4] Younes N, Al-Sadeq DW, Al-Jighefee H, Younes S, Al-Jamal O, Daas, HI, et al. Challenges in Laboratory Diagnosis of the Novel Coronavirus SARS-CoV-2. Viruses 2020;12:582. [CrossRef]

[5] Zhu X, Xu T, Lin Q, Duan Y. Technical development of Raman spectroscopy: From instrumental to advanced combined technologies. Appl Spectrosc Rev 2014;49:64–82. [CrossRef]

[6] Kahraman M, Yazici MM, Slahin F, Bayrak ÖF, Çulha M. Reproducible surface-enhanced raman scattering spectra of bacteria on aggregated silver nanoparticles. Appl Spectrosc 2007;61:479–485. [CrossRef]

[7] Al-Shaebi Z, Uysal Ciloglu F, Nasser M, Aydin O. Highly accurate identification of bacteria's antibiotic resistance based on Raman spectroscopy and u-net deep learning algorithms. ACS Omega 2022;7:29443–29451. [CrossRef]

[8] Ciloglu FU, Hora M, Gundogdu A, Kahraman M, Tokmakci M, Aydin O. SERS-based sensor with a machine learning based effective feature extraction technique for fast detection of colistin-resistant Klebsiella pneumoniae. Anal Chim Acta 2022;1221:340094. [CrossRef]

[9] Ye J, Yeh Y-T, Xue Y, Wang Z, Zhang N, Liu H, et al. Accurate virus identification with interpretable Raman signatures by machine learning. Proc Natl Acad Sci U S A 2022;119:e2118836119. [CrossRef]

[10] Carlomagno C, Bertazioli D, Gualerzi A, Picciolini S, Banfi PI, Lax A, et al. COVID-19 salivary Raman fingerprint: innovative approach for the detection of current and past SARS-CoV-2 infections. Sci Rep 2021;11:4943. [CrossRef]

[11] Zhang D, Zhang X, Ma R, Deng S, Wang X, Wang X, et al. Ultra-fast and onsite interrogation of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in waters via surface enhanced Raman scattering (SERS). Water Res 2021;200:117243. [CrossRef]

[12] Akdeniz M, Ciloglu FU, Tunc CU, Yilmaz U, Kanarya D, Atalay P, et al. Investigation of mammalian cells expressing SARS-CoV-2 proteins by surface-enhanced Raman scattering and multivariate analysis. Analyst 2022;147:1213–1221. [CrossRef]

[13] Yin G, Li L, Lu S, Yin Y, Su Y, Zeng Y, et al. An efficient primary screening of COVID-19 by serum Raman spectroscopy. J Raman Spectrosc 2021;52:949–958. [CrossRef]

[14] Deepaisarn S, Vong C, Perera M. Exploring Machine Learning Pipelines for Raman Spectral Classification of COVID-19 Samples. In: 2022 14th International Conference on Knowledge and Smart Technology (KST). pp. 51–56. [CrossRef]

[15] Wei Y, Chen H, Yu B, Jia C, Cong X, Cong, L. Multiscale sequential feature selection for disease classification using Raman spectroscopy data. Comput Biol Med 2023;162:10705. [CrossRef]

[16] Arslan AH, Ciloglu FU, Yilmaz U, Simsek E, Aydin O. Discrimination of waterborne pathogens, Cryptosporidium parvum oocysts and bacteria using surface-enhanced Raman spectroscopy coupled with principal component analysis and hierarchical clustering. Spectrochim Acta A Mol Biomol Spectrosc 2022;267:120475. [CrossRef]

[17] Ciloglu FU, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. Identification of methicillin-resistant Staphylococcus aureus bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. Analyst 2020;145:7559–7570. [CrossRef]

[18] Jollife IT. Principal Component Analysis. New York: Springer; 2002.

[19] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2013.

[20] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020;181:281–292.e6. [CrossRef]

[21] Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and metabolomic characterization of COVID-19 patient sera Cell 2020;182:59–72.e15. [CrossRef]

[22] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–1159. [CrossRef]

[23] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–874. [CrossRef]