Detection of Coronavirus in Viral Transport Media using Ultraviolet and Near-Infrared Absorbance Spectra and Pattern Recognition Model

Muhammad Izzuddin Rumaling  
Universiti Malaysia Sabah

Fuei Pien Chee (✉ fpchee06@ums.edu.my)  
Universiti Malaysia Sabah

Abdullah Bade  
Universiti Malaysia Sabah

Jackson Hian Wui Chang  
Universiti Malaysia Sabah

Lucky Poh Wah Goh  
Universiti Malaysia Sabah

Floressy Juhim  
Universiti Malaysia Sabah

Research Article

Keywords: Absorbance spectra, pattern recognition model, coronavirus, principal component analysis (PCA), linear discriminant analysis (LDA)

Posted Date: December 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3684212/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Abstract

SARS-CoV-2 causes individuals to become infected with respiratory disease known as COVID-19. Rapid and robust identification ensures that the infected patients can be quarantined. In this paper, the detection of SARS-CoV-2 utilizes ultraviolet (UV) and near-infrared (NIR) absorbance spectra, along with principal component analysis and linear discriminant analysis (PCA-LDA). A total of 75 negative and 75 positive swab samples are separately placed in vials of viral transport media and transferred into cuvettes. The absorbance spectra are acquired and processed before they undergo dimensionality reduction using PCA. The dataset is divided into training set and testing set to develop and evaluate the PCA-LDA model. The scree plot analysis reveals that the two principal components are optimal for both UV and IR absorbance spectra. By utilizing the first two principal components, the performance indicators demonstrate higher accuracy (97.00%), sensitivity (94.84%), and specificity (99.31%) on IR absorbance spectra. This is attributed to the overall difference in IR absorbance, as well as two peaks centred at 558.5 nm and 972 nm respectively. Utilizing IR absorbance spectra with PCA-LDA model is cost-effective while showing performance comparable to conventional methods such as polymerase chain reaction. This method provides an alternative for rapid and effective SARS-CoV-2 detection.

1. Introduction

Back in December 2019, COVID-19 (coronavirus disease-19) outbreak started in Wuhan, China. The disease is caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2), which spread internationally and caused worldwide pandemic (Haynes et al., 2020). Since then, governments in many countries have enacted countermeasures such as border restriction, closure of public places and social distancing to reduce the spread of the disease (Ganasegaran et al., 2020). Public health enforcements have employed several methods in detecting SARS-CoV-2 in patients accurately to identify healthy and infected patients. Lateral flow immunoassay offers cost-effective method that gives result in short time (Di Nardo et al., 2021). Test kits over-the-counter employs lateral flow immunoassay, which can be conducted by the patients themselves. Despite the availability, results obtained from lateral flow immunoassay still require extensive validation and quality (Lingervelder et al., 2019). Polymerase chain reaction (PCR) is the most commonly used and reliable method used in many health facilities. PCR provides quantitative results with good accuracy and sensitivity. However, it requires a series of steps involving chemical and heat treatment which may take up to 48 hours. Furthermore, PCR cannot be used on recently recovered patients as this may lead to false positive result (Tan et al., 2022).

Optical spectroscopy is a non-invasive method that involves the interaction of light with matter to study the constituents of molecules inside a sample (Rumaling et al., 2022). There are several types of optical spectroscopy, including absorbance spectroscopy that makes use of difference in intensity from reference spectra, Raman spectroscopy that makes use of Raman scattering and fluorescence spectroscopy which involves excitation and emission of light at different wavelength (Auner et al., 2018; Bose et al., 2018). Due to its non-destructive nature, the usage of optical spectroscopy in recent years has become more common. In addition, optical spectroscopy has been proven to successfully identify healthy
samples from virus-infected samples. For example, a study conducted by Rumaling et al. (2023) showed that healthy and SARS-CoV-2 infected samples in viral transport media (VTM) can be differentiated based on a set of peaks detected using Raman spectroscopy.

Ultraviolet (UV) and infrared (IR) absorbance spectroscopy involves electromagnetic waves in UV and IR region. UV radiation is more energetic compared to visible light and is defined as wavelength below 400 nm. The absorption of electromagnetic radiation in the UV range is associated with excitation of electrons in atoms and molecules, from lower to higher energy levels (Atole and Rajput, 2018). Meanwhile, IR radiation is less energetic than visible light. Near-infrared (NIR) is a class of IR radiation with higher photon energy than other IR classes, having wavelengths range from 750–2500 nm. While NIR is energetic enough to cause electronic transition in molecules, it is sensitive to overtones and combinations of vibrations within a molecule (Pasquini, 2018; Veerasingam et al., 2021). It has been shown in previous studies that NIR absorbance spectra can be reliably used to identify viral infected samples, particularly Zika and influenza virus (Fernandes et al., 2018; Sakudo et al., 2012). Furthermore, the 260nm/280nm absorbance ratio has been used to identify the presence of viral proteins (Minamikawa et al., 2021; Piva et al., 2020; Porterfield and Zlotnick, 2010). These findings motivated this study to apply UV and IR absorbance spectra in identifying the presence of SARS-CoV-2 in samples.

Pattern recognition is typically employed in optical spectroscopy studies to classify healthy samples from infected samples. Pattern recognition concerns with detecting patterns in dataset using algorithm for the purposes such as grouping of data (Hasbi et al., 2022). Pattern recognition is mainly executed using classification algorithm that allows difference in spectra to be detected. The classification algorithm mainly involves multiple iterations of steps with stopping criterion that ensures optimization of specific parameters (Alfeilat et al., 2019; Boonamnuay et al., 2018). Exploratory data analysis (EDA) is executed on multidimensional data prior to classification algorithm. This is because the efficiency of classification algorithm decreases with higher dimensions of data (Hasbi et al., 2022). EDA transfers the data to lower dimensional space in a process known as dimensionality reduction (Sammut and Webb, 2010).

In recent years, many countries have entered endemic phase where public restrictions have been relaxed. Despite that, COVID-19 outbreak might still occur and the effects of COVID-19 are detrimental to high-risk individuals such as children, elderly, pregnant women and immunocompromised patients (Rumaling et al., 2023). Thus, there is still the need to detect COVID-19 in patients using rapid and robust method. Past studies involving optical spectroscopy have been utilizing Raman spectroscopy and Fourier transform infrared (FTIR) spectroscopy to study the presence of SARS-CoV-2 in samples. However, the UV and NIR absorbance spectroscopy, which are the lower-cost alternatives, has not been extensively studied for the same purpose. Thus, this study aims to identify the healthy and SARS-CoV-2 infected samples in VTM using UV and IR absorbance spectra. Principal component analysis (PCA) and linear discriminant analysis (LDA) are also used to aid identification of samples.

2. Data and Methods
2.1. Sample preparation

In this study, 150 swab samples were acquired from Kuala Lumpur International Healthcare Centre Kota Kinabalu (KLIHCKK) and stored in vials containing VTM at around −18°C. 75 of these swab samples were extracted from healthy patients, while the other 75 samples were taken from patients with confirmed COVID-19 infection. The samples were then prepared for acquisition of absorbance spectra. 1 mL of each sample was transferred into cuvettes using a micropipette. After that, the samples were analyzed for UV and NIR absorbance. Extraction and preparation of samples were conducted under safety precautions, preventing potential contamination between the samples and the environment. Additionally, the study was conducted according to ethical guidelines and regulations to ensure that the privacy of the patients is protected.

2.2. Acquisition and processing of spectral data

Figure 1 illustrates the setup for spectrometer system developed for this study. This setup is applicable for acquisition of both UV and NIR absorbance spectra. The setup consists of a light source, a diffraction grating and a charged couple device. The light source used in this setup depends on the spectra acquired. A 10-W halogen lamp is used to obtain NIR absorbance spectra, while a 30-W deuterium lamp is used for UV absorbance spectra. Light passes through cuvette with optical path of 10 mm before it moves along an optical fibre. A diffraction grating is used to separate the components of transmitted light based on wavelength. The separated light is captured by charge-coupled device (CCD) connected to a computer. UV absorbance spectra of samples are taken in wavelengths ranges between 220 nm and 400 nm. Meanwhile, NIR absorbance spectra are measured from 550 nm to 1050 nm.

Reference intensity $I_o$ and intensity of transmitted light $I$ must be known before absorbance $A$ can be obtained. Reference intensity spectra is obtained from an empty cuvette placed inside the holder while keeping the light source illuminated. The transmitted intensity spectra is taken for each sample and compared with reference intensity spectra. By obtaining both spectral data, the absorbance spectra can be obtained by calculating $A$ using Eq. (1) (Guo et al., 2020).

$$A = \log_{10} \left( \frac{I}{I_o} \right)$$

The absorbance spectra obtained using setup as illustrated in Fig. 1 have high signal-to-noise ratio. The absorbance spectra are smoothened using Savitzky-Golay filter with third-order polynomial, taken seven points at a time to remove the noise inside the data. This ensures that the background noise is removed, consequently reducing the processing steps required for the development of pattern recognition model (Khan et al., 2018).

2.3. Principal component analysis (PCA)
Principal component analysis (PCA) is a well-established and commonly used technique in achieving dimensionality reduction (Hasbi et al., 2022). PCA is a non-iterative technique that converts standardized original dataset \( X \) into linear combination of variables with the most significant variance, known as principal components, \( PC \). As shown in Eq. (2), the variables \( X \) and \( PC \) are related by factor loading \( l \), which indicates the contribution of a variable into certain PCs, with higher magnitude indicating higher contribution to that PCs (Kherif and Latypova, 2019). The factor loading is evaluated in such a way that the PCs are orthogonal to each other, to ensure reduced multicollinearity between each PC.

\[
PC_i = \sum_{j=1}^{n} l_{ji} X_i
\]

Each PC has corresponding eigenvalue which is obtained from covariance matrix of \( X \). Eigenvalues describe how much variance is explained by PCs in relation to the original dataset. Since each successive PC has lower eigenvalue, the first few PCs will be more significant to the original dataset (Kherif and Latypova, 2019). Other PCs can be discarded as they are not significant to the original dataset, based on their low eigenvalues. The selection of first few PCs can be made by plotting eigenvalues against corresponding PCs in scree plot. The number of PCs can be determined by point of inflexion, beyond which addition of principal components does not significantly contribute to dataset.

### 2.4. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a supervised classification method used to categorize positive and negative samples. The classification of LDA utilizes Bayes’ theorem that relates prior probability \( \pi_k \) (estimated from training set) and posterior probability \( P(X = x | Y = k) \), which gives the probability that the selected observation is in \( k \)-th class. Bayes’ theorem is described in Eq. (3) as (James et al., 2013):

\[
P(X = x | Y = k) = \frac{f_k(x) \pi_k}{\sum_{i=1}^{k} \pi_i f_i(x)}
\]

\( f_k(x) \) is the probability density function of \( X \) located in \( k \)-th class. \( f_k(x) \) can be assumed to be a normal distribution with mean vector \( \mu_k \) and covariance matrix \( \Sigma \) when only one predictor is involved. With such assumption, LDA assigns the observation \( X = x \) to the class which maximizes linear discriminant function \( \delta_k(x) \) (as shown in Eq. (4)), derived from optimizing posterior probability and maximizing the term \( f_k(x) \pi_k \).

\[
\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)
\]
Classification between healthy and SARS-CoV-2 infected samples is an example of binary classification, which divides dataset into only two classes. Both classes are divided by decision boundary, derived by equating linear discriminant functions of both classes $\delta_k$ and $\delta_l$. Thus, decision boundary is defined as in Eq. (5):

$$\log \left( \frac{\pi_k}{\pi_l} \right) - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) = 0$$

2.5. Development and evaluation of pattern recognition model

In this study, the development and evaluation of PCA-LDA model is executed in MATLAB. The eigenvalue produced in PCA is used in scree plot to determine the optimal number of PCs required for LDA. Then, the dataset is randomly divided into 80% of training set and 20% of testing set, ensuring that both positive and negative samples are present in both sets in approximately equal proportions. The training set is used to develop the LDA model, while the testing set is used for classification based on the developed model. The results of classification in each of the data in testing set are divided into four types, namely true positive (correctly predicted positive sample, abbreviated as TP), true negative (correctly predicted negative sample, abbreviated as TN), false positive (incorrectly predict negative sample as positive, abbreviated as FP) and false negative (incorrectly predict positive sample as negative, abbreviated as FN).

The developed model is evaluated using three performance indicators. Accuracy is calculated based on number of accurate predictions relative to total predictions, as expressed in Eq. (6). Sensitivity, as shown in Eq. (7), is defined as number of accurate predictions on positive samples relative to total positive samples. Meanwhile, specificity indicates the number of accurate prediction on negative samples relative to the total negative samples, calculated using Eq. (8) (Sammut and Webb, 2010).

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

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$
3. Results and Discussion

3.1. Principal component analysis (PCA)

Figure 2 shows the scree plot for the first 10 principal components. It is revealed that the point of inflexion occurs at second principal component for both spectra, indicating that additional PCs beyond PC$_2$ is not significant. The first two PCs account for 93.20% of data in UV absorbance spectra, and 98.16% of data in IR absorbance spectra.

The composition of PCs based on factor loading can be observed in Figs. 3 and 4. In the case of UV absorbance spectra (Fig. 3), it can be observed that the factor loading of PC$_1$ peaks at around 230 nm and decreases to around 270 nm. A slight peak can be observed at around 280 nm, just before the factor loading decreases to negative value. As for PC$_2$, the factor loading increases up until 235 nm. The factor loading of PC$_2$ is maximum at this point, but not as high as in PC$_1$.

Meanwhile, Fig. 4 reveals the loading factor for the first two PCs in IR absorbance spectra. Both factors peak at around 560 nm before they drop and reach to a steady value. The factor loading in PC$_1$ has smaller peak and reaches to a steady value not far from maximum. Meanwhile, the factor loading of PC$_2$ has more prominent peak just before it drops sharply and reaches to a negative steady value.

3.2. PCA-LDA model

The first two PCs generated from PCA are used as dataset for LDA model. 80% of dataset is used as training set to develop the classification model and define the decision boundary between positive and negative samples. Meanwhile, 20% of dataset is used as testing set to identify positive and negative sample based on model developed on training set. The result of identification is illustrated in scatter plot of two first PCs as shown in Fig. 5. In the case of UV absorbance spectra in Fig. 5(a), negative samples are generally located below the decision boundary, while positive samples are located below the decision boundary. However, it can be observed that three negative samples and one positive sample are incorrectly identified. As for IR absorbance spectra in Fig. 5(b), positive and negative samples are located at the right and left of the decision boundary respectively. Only one negative sample is incorrectly identified based on the decision boundary.

The performance of PCA-LDA model is quantified using the performance indicators as tabulated in Table 1. It can be observed that PCA-LDA model generally has better performance in IR absorbance spectra compared to UV absorbance spectra. With performance indicators above 90%, IR absorbance spectra reliably identifies both positive and negative samples. In contrast, the performance degradation in UV absorbance spectra is reflected by incorrect identification of few samples as observed in Fig. 5(a).
reliability of UV absorbance spectra in terms of identification of positive and negative samples could not surpass that of IR absorbance spectra.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Performance indicators of PCA-LDA model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absorbance spectra</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>84.67</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>83.23</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>86.20</td>
</tr>
</tbody>
</table>

### 3.3. Spectral data

Figure 6 reveals the average spectral data of positive and negative samples for UV absorbance spectra. It can be observed that the spectra showing sharp decrease in absorbance with a peak located between 260 and 280 nm. Protein and nucleic acid contents in samples can be qualitatively determined based on 260nm/280nm absorbance ratio (Minamikawa et al., 2021; Piva et al., 2020). Nucleic acid tends to absorb UV radiation at wavelength of around 260 nm, leading to higher 260nm/280nm absorbance ratio. Meanwhile, protein absorbs radiation with wavelength close to 280 nm, causing lower 260nm/280nm absorbance ratio (Porterfield and Zlotnick, 2010). It is found that the positive and negative samples have average (± standard deviation) of 260nm/280nm absorbance ratio of 1.1149 ± 0.1903 and 1.1878 ± 0.1208 respectively. While positive samples have slightly higher nucleic acid content based on lower 260nm/280nm absorbance ratio compared to negative sample, the difference in ratio between both samples is not significant. This might contribute to the lower performance for PCA-LDA model.

Figure 7 illustrates the average IR absorbance spectral data of positive and negative samples. It can be clearly observed that the absorbance value for positive samples is lower for the whole range of wavelengths compared to negative samples. This might contribute to the high performance of LDA-PCA model. While both samples share similar peaks, a peak at around 558.5 nm is very noticeable in the negative sample. There is also a peak at 972 nm present in both samples. This peak is close to water band in samples (Sakudo et al., 2012).

### 4. Conclusion

In this study, PCA-LDA model is applied on UV and IR absorbance spectra to extract information on spectral data and identify healthy and SARS-CoV-2 infected samples in VTM. The usage of factor loading in PCA helps in identifying spectral differences between positive and negative samples, which is utilized in LDA model. With the performance comparable to conventional methods, this technique allows the identification of SARS-CoV-2 in samples rapidly and cost-effectively, while reducing the occurrence of false positives and false negatives.
Declarations

Acknowledgement

The authors would like to express gratitude to Universiti Malaysia Sabah for supporting this research by providing research grant with code SDK0163-2020.

1. Ethical approval

Not applicable as this study doesn't involve human and/ or animal studies. This study was registered at National Medical Research Register with the reference number: NMRR ID-22-02382-ZNZ.

2. Competing interests

The authors declare no conflict of interests.

3. Authors’ contributions

Conceptualization, Fuei Pien Chee and Abdullah Bade; Methodology, Muhammad Izzuddin bin Rumaling, Fuei Pien Chee, Abdullah Bade and Lucky Poh Wah Goh; Validation, Fuei Pien Chee, Lucky Poh Wah Goh and Jackson Chang Hian Wui; Formal analysis, Muhammad Izzuddin bin Rumaling and Floressy Juhim; Resources, Fuei Pien Chee and Abdullah Bade; Data curation, Muhammad Izzuddin bin Rumaling and Floressy Juhim; Writing – original draft, Muhammad Izzuddin bin Rumaling; Fuei Pien Chee, Abdullah Bade and Lucky Poh Wah Goh; Visualization, Floressy Juhim, Jackson Chang Hian Wui; Supervision, Fuei Pien Chee and Abdullah Bade; Funding acquisition, Fuei Pien Chee.

All authors reviewed the manuscript.

4. Funding

Funding is supported by Universiti Malaysia Sabah under research grant with code SDK0163-2020.

5. Availability of data and materials

The datasets analysed during the current study are not publicly available due to confidentiality but are available from the corresponding author on reasonable request.

References


---

**Figures**

[Diagram of optical spectroscopy setup]
Figure 1

Simplified setup of absorbance spectrometer

Figure 2

Scree plot for both UV and IR absorbance spectra
Figure 3

Factor loading of first two PCs for UV absorbance spectra
Figure 4

Factor loading of first two PCs for IR absorbance spectra
Figure 5

Scatter plot of testing set of (a) UV, and (b) IR absorbance spectra predicted using PCA-LDA model
Figure 6

Average spectral data of positive and negative samples for UV absorbance spectra
Figure 7

Average spectral data of positive and negative samples for IR absorbance spectra