TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES
College of Engineering
Electronics Engineering Department

SARS-CoV 2 Detection Using FTIR Spectroscopy by Comparison of Chemometric Analysis Through Oropharyngeal Swab Samples' Absorbance Levels

Members:

DE JOSE, Jhobelle A.

FABABEIR, Ralph Vincent F.

GRENCIO, John Dave S.

JARABEJO, Kianna Angela A.

NEMO, Czriss Paulimer C.

BSECE 4B

Chapter 1

Introduction

From business to the biomedical field, big data is evident. This pertains to the large volume of available data, either structured or unstructured, that cannot be simply analyzed by simple statistical means (Rai, 2020). A tantamount of information lies within these numbers. The extracted information from big data is key for informed decision making among organizations. For instance, recent advances of the biomedical field in handling personal information of patients is brought by the analyses of the continuous growth of biomedical information (Costa, 2014).

Data analytics is a necessity in order to obtain the underlying information from big data. chaiThis includes predictive models, statistical algorithms and analytic systems that propose what-if analyses (Chai, Labbe & Stedman, 2021). Acquiring such inferences from big data is a great help in improving the quality of life. In healthcare, big data analytics serves as a stepping stone in improving patient healthcare, ease of patient diagnosis, quick prediction of risks on patients, and so on (TestingXperts, 2020).

Big data can also be used in biostatistics, which is the analysis and design of medical and public health research studies (Leaverton and Zhu, 2017). In a sense, biostatistics utilize data in order to determine the effects of a particular treatment on a patient or even the factors that influence the progression or regression of a certain disease (Perry, 2013). In retrospect, biostatistics is a major contributor in deriving useful information from public clinical data.

Chemometric analysis is among the vast tools used in biostatistics. By definition, it pertains to the manipulation of data from chemical processes by mathematical and statistical approaches (Ferreira, 2019). This is commonly used in treating data derived from infrared spectroscopy; a technique that is gaining popularity nowadays in the biomedical field. What hinders the inclusion of infrared spectroscopy as an instrument in clinical laboratories is the lack of studies that testifies its capacity for patient diagnosis. Hence, the researchers would like to conduct the study that will provide further evidence of the potential of infrared spectroscopy as a clinical laboratory equipment through chemometric analysis of the available big data in the biomedical field.

1.1 Background of the Study

SARS-CoV-2, commonly known COVID-19, is the most prevalent disease to date. With a death toll of 19,763 deaths and 1,171,403 confirmed cases in the Philippines, it is truly alarming (WHO, 2021). To date, there are two types of tests for SARS-CoV-2: the molecular tests that use nasal and oropharyngeal swabs for Reverse Transcription Polymerase Chain Reaction (RT-PCR), and antibody tests that employ blood samples to look for antigens present in the patient (FDA, 2020).

With a seemingly exponential growth in cases, clinical laboratories face a challenge in meeting the demands of testing capacity while not compensating for the quality of laboratory operation (Binnicker, 2020). Clinical diagnosis is the first step in curbing the spread of the disease, thus this field must never be compromised.

Even though RT-PCR tests are the common method for SARS-CoV-2 testing, its accuracy and sensitivity is questionable at times. Huerta et al. (2020) stated that in different clinical stages of SARS-CoV-2 corresponds to different sensitivity and specificity of the RT-PCR test. Hence, the most feasible solution with such a problem is the use of more than one test for further validation.

Infrared spectroscopy is a method used for studying small molecular structures. It makes use of the vibrational transitions of molecules caused by infrared radiation (Barth, 2007). These vibrational frequencies, together with absorption probabilities, are dependent on the polarity and strength of the vibrating bonds. In general, these vibrations translate to the specific nature of a protein. On the other hand, the infrared absorption is derived from the proteins' polar bonds.

Knowing the potential of infrared spectroscopy, it is the most favored method of protein characterization among professionals. According to Barth (2007), with either the absorbance or transmittance levels plotted against the wavenumber, which is the inverse of wavelength, information such as the chemical structure of the vibrating group as well its neighbouring molecules, bond parameters, bond angles and conformation, hydrogen bonding and electric fields

can be determined. All of these parameters will eventually lead to identification of the exact protein present in a given substance.

Knowing all the shortcomings of the current tests administered in SARS-CoV-2 diagnosis as well as the problems that clinical diagnostic laboratories face, the researchers thought of another approach that can help in further improving clinical implications which will in turn improve the rate of mitigation of this prevalent disease. The study will be a website application whose input will be from a Fourier Transform Infrared (FTIR) spectrometer, a non-invasive technique in sample inspection which employs the concept of infrared spectroscopy, that will output the diagnosis for the patient.

1.2 Statement of the Problem

Clinical diagnosis plays a major role in determining the appropriate treatment in patients affected with COVID-19. Such diagnosis is administered by using the Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test. However, this process takes about 24 to 48 hours before getting the result. As discussed earlier, such misdiagnosis can take lives and because of the overwhelming number of samples subject for testing, laboratories cannot keep up. Hence, RT-PCR can be unreliable in some cases. Most laboratories use the antibody test as a parallel testing procedure to verify the accuracy of the RT-PCR diagnosis.

Aside from this, clinical diagnostic testing is also currently facing numerous technical and financial challenges during these trying times since there is an increase in the demand for testing capacities. One of these is the cost that is brought by administering inappropriate laboratory tests and the lack of automation. Medical technologists also need a less complex, and speedy procedure in analyzing the samples.

1.3 Objectives

1.3.1 General Objectives

The research aims to create a website application that will help in diagnosing if an oropharyngeal swab sample (OPS) is COVID-19 positive or negative with the help of the output absorbance spectrum from a Fourier Transform Infrared (FTIR) spectrometer.

1.3.2 Specific Objectives

1. Develop and compare predictive models with the use of chemometric packages of R Programming which will analyze the CSV file from the FTIR spectrometer.

2. Design a user-friendly website application that will output:

    2.1 Diagnostic report of the OPS and can be generated in a Portable Document Format (PDF) file

    2.2 Graphical presentation of the number of registered data of the diagnosed patients

3. Test the website application with the FTIR analysis of OPS in csv file format prior to project deployment.
4. Evaluate the accuracy and validity of results gathered from this engineering research in accordance with the ISO 25010:2020 software evaluation standard.

1.4 Significance of the Study

The findings of the study will be beneficial to clinical diagnostic laboratories since professionals currently struggle with regards to the efficiency and financial aspects in clinical diagnosis. Since most laboratories are still semi-automated, there are instances that it takes a substantial amount of time to release the results since the standard process of manual examination is very extensive.

The Fourier Transform Infrared (FTIR) spectrometer provides professionals with a non-invasive technique in studying the samples. The process in sample preparation is relatively easier in comparison with flow cytometry. The proposed website application will also be user-friendly since the main problem that limits scientists in using the Fourier Transform Infrared (FTIR) spectrometer is the required statistical analysis in quantifying proteins in general. The website application will also be easy to access since it is in the form of a web application.

Having knowledge of all this, contributing to existing research about Fourier Transform Infrared (FTIR) spectrometer as a potential diagnostic tool that can help in easing clinical diagnostic workload since all processes involved here are simple. Aside from this, knowing that infrared spectroscopy is a non-invasive process, preservation of samples is plausible thus, the same sample can also be used in other tests as well. This helps in cost reduction and lessen the time spent in examination since the laboratory test is made easier with the website application.

1.5 Scope and Limitations

The study will focus on developing a website application that helps in analyzing oropharyngeal swabs with the help of the output absorbance spectrum from a Fourier Transform Infrared (FTIR) spectrometer. The website application will utilize the CSV file from the Fourier Transform Infrared (FTIR) spectrometer and is limited only to diagnosing if the patient is positive or negative from SARS-CoV-2. The researchers will focus solely on the analysis of the dataset obtained from the Centers for Disease Control and Prevention. Moreover, this dataset will be also used for the evaluation of the predictive models. For this study, only three predictive models will be studied; the Principal Component Analysis-Quadratic Discriminant Analysis (PCA-QDA), Partial Least Square Regression (PLSR) and Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA).

1.6 Definition of Terms

CSV File – is the raw data from the FTIR spectrometer which will be the input to the website application.

Fourier Transform Infrared (FTIR) Spectrometer – is an electronic device that is capable of quantifying protein levels in the form of an infrared spectrum.

Oropharyngeal Swabs - is the samples that will be subjected to the FTIR spectrometer.

Pre-processing - refers to the treatment of the raw spectral data to reduce unwanted peaks for easier analyzation.

Server - is defined as the platform wherein the website application will be deployed for the end users to access.

Spectral Data - is the general term for the supposed input in the website application.

Remote Database - pertains to the storage of the results outputted by the website application.

R Programming – is the tool used in developing the predictive model as well as a mode of implementation of the model as a web application.

Chapter 2

Review of Related Literature and Studies

2.1 Conceptual Literature

2.1 Spectroscopy

2.1.1 Infrared Spectroscopy

An analytical technique between the infrared spectrum and matter. It uses vibrational transitions in microscopic molecules to examine the quality of materials and identify unknown materials present in the sample. The idea of infrared spectroscopy is to measure absorption or transmittance of light rays, the wavelength of infrared can be obtained.

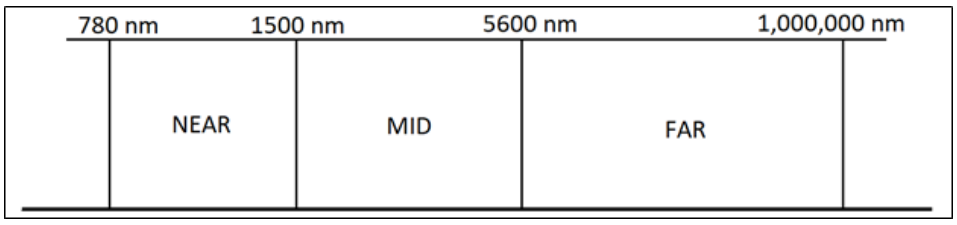


Figure 1: Infrared Spectrum

Infrared spectroscopy promises the capability to analyze samples in the form of the three (3) primary phases of matter (solid, liquid, and gas). It has two (2) primary variants which are Dispersive Infrared Spectrometer and Fourier Transform Infrared Spectroscopy, the researchers used the latter for this study (Libretexts, 2020).

2.1.2 Fourier Transform Infrared Spectroscopy

Fourier transform infrared spectroscopy (FTIR) is one of the most used types of spectroscopy due to its capability to simultaneously analyze different frequencies without compromising the accuracy of the testing. FTIR promises faster analysis of samples over any other type of spectroscopy. It uses an Interferometer, a device which is used to create an interference pattern by merging two light sources (LIGO, n.d.). After the process of recombining the splitted light of the interferometer, it will generate an interferogram which is converted into absorbance or transmittance spectrum versus the wavenumber or frequency of the spectrometer (Libretexts, 2020).

2.1.3 Application of Infrared Spectroscopy

In a vague point of view, Infrared Spectrometers can be used for Qualitative and Quantitative Analysis of samples. Due to the fact that each substance radiates their own unique spectra, the qualitative analysis of substances is possible in Infrared Spectroscopy. Generally, the output of an infrared spectrometer for qualitative analysis is transmittance versus the wavenumber. On the other hand, quantitative analysis is possible due to the direct proportionality between the intensity of the peak of the output of the spectrometer and the amount of substance present in the sample. Absorbance, by the virtue of Beer-Lambert Law, is used in quantitative analysis of spectrometers due to its linear dependence on concentration.

$$A=\epsilon cl$$

Beer-Lambert Law

Where $\boldsymbol{A}$ is the Absorbance, $\boldsymbol{\epsilon}$ is the molar absorptivity, $\boldsymbol{c}$ is the concentration, and $\boldsymbol{l}$ is the path length of sample.

For quantitative analysis of liquid samples such as blood, the peak at which the molar absorptivity is high is used to quantify the substance.

The absorbance at the chosen frequency versus the concentration of the substance is acquired after measuring the absorbance of a series of compounds with known concentrations (LIGO, n.d.).

2.2 Chemometric Analysis

2.2.1 Spectral Pre-processing

Pre-processing is commonly used first in a chemometric analysis of spectral data for reduction of noises and of unwanted variables that can affect the quantitative analysis of the data. Since biological components are analysed in most of the spectroscopy techniques, factors such as environment, experiment, and technical conditions can create different variances that affect the datasets (Ollesch et al., 2013). Spectral pre-processing mainly improves the robustness and accuracy of subsequent quantitative or classification analyses and interpretability, detects and removes outliers and trends, and reduces redundant information by feature selection (Lasch, 2012). It usually performs data binning, data smoothing, normalisation, and baseline correction (Ollesch et al., 2013).

2.2.1.1 Savitzky-Golay

Savitzky-Golay is a common filtering technique used in data smoothing in FTIR Spectroscopy data analysis. It can minimize the high frequency noise while the peak morphology is maintained (Ollesch et al., 2013). This uses linear least squares to fit successive sub-sets of adjacent data points with a low-degree polynomial. A single set of “convolution coefficients” is thus acquired through optimization (Whittaker and Robinson, 1924).

2.2.2 Principal Component Analysis

In multivariate analysis. principal component analysis, or PCA is the most basic of all analyses. It is a method that reduces a large set of data into

smaller sets of variables to better observe trends, jumps, clusters and outliers (Sartorius, 2020).

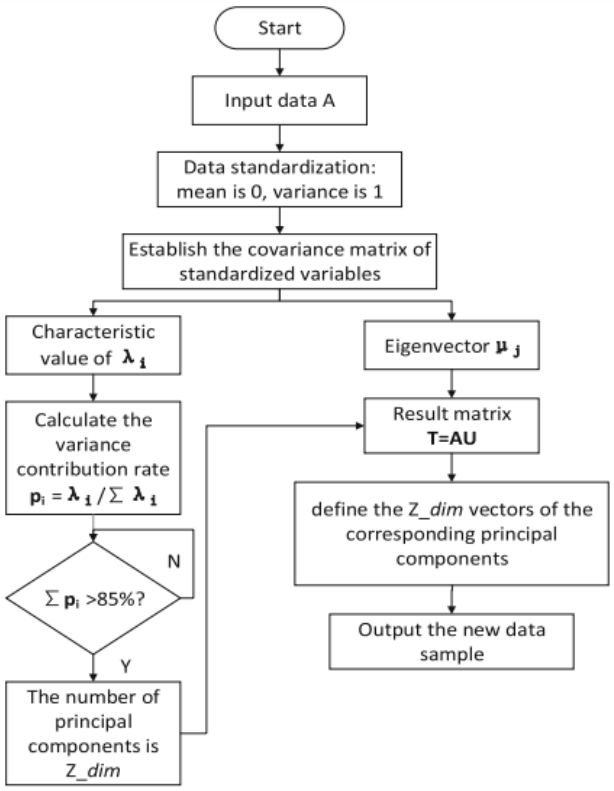


Figure 2: PCA flowchart from Wang et al. (2020)

The flowchart above shows the PCA. The input data is normalized first with mean is 0 and variance is 1. Afterwards, eigenvalues and eigenvectors, and the eigenvectors corresponding to the largest zdim eigenvalues are selected by establishing covariance. In the final process, the zdim eigenvectors mapped the data into the new space thus the data is compressed (Wang et al., 2020).

2.2.3 Linear Discriminant Analysis

Gaussian Discriminant Analysis (GDA) is a generative learning algorithm which solves a classification problem in analyzing probability distributions of the data (Mortuza, 2020). One type of this model is the

Linear Discriminant Analysis (LDA) that assumes observations from each class are normally distributed or in Gaussian distribution and that these observations share the same variance (Zach, 2020). The Gaussian distribution is obtained by following the formula:

$$f_k(x) = |2\pi\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

where:

Σk is the covariance matrix for the samples from class k

|.| is the determinant

μk is the mean of all training observations from the kth class.

The LDA makes its prediction based on Bayes' rule:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

where:

Pr is the prior probability

f(x) is the estimated probability that x belongs to that particular class

πk is the proportion of the training observations that belong to the kth class. The discriminant function for class k of this model with more than one predictor is

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

With this, the classification function can be obtained by the following equation:

$$G(x) = \arg\max_k \delta_k(x)$$

2.2.4 Quadratic Discriminant Analysis

Another type of a GDA is the Quadratic Discriminant Analysis (QDA), which also assumes that observations from each class are also normally distributed but not in the same covariance matrix. It estimates the mean μk and covariance Σk for each class k∈{1,…,K}. The discriminant function of QDA is as follows:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k .$$

Since the covariance is estimated separately for each class, QDA generates a higher number of effective parameters than LDA (Rasmussen, 2021).

2.2.5 Partial Least Squares Regression

Where the parameters are many and strongly collinear, partial least squares (PLS) is a tool for building predictive models. It's worth noting that the focus is on forecasting responses rather than attempting to comprehend the underlying relationship between the variables. PLS, for example, is not often used to filter out variables that have a minor impact on the response. On the other hand, it can be helpful where the target is prediction and there is no need to restrict the number of calculated variables (Tobias, n.d.).

According to Wold, Sjostro & Eriksson (2001), cross-validation is the pre-process used in determining which points from the given dataset are of significance to the supposed goal of the analysis. This is done by dividing the dataset into groups G, and then develop parallel models N from these groups.

Prior to obtaining needed information from the training set, the Partial Least Squares Regression (PLSR) first aim to determine the X-scores t in relation to weight w, which can be normalized, through the formula:

$$t = Xw$$

Not only the X-scores, but also the Y-weights are considered in the analysis. This can be determined through the obtained X-scores earlier. Mathematically, it is expressed as:

$$c = \frac{Y't}{t't}$$

The Y-scores will now be modified with respect to the weights with the formula:

$$u = \frac{Yc}{c'c}$$

After obtaining the needed information for training the model, convergence tests will be administered with respect to the formula:

$$\frac{\|t_{old} - t_{new}\|}{\|t_{new}\|} < \varepsilon$$

where ε is the criteria of convergence ranging from 10^{-6} to 10^{-8}. If this equation is not satisfied, then the X-scores must be recomputed.

If the obtained scores adhere to the aforementioned inequality, deflation of the X- and Y- matrices will take place. In this case, Wold, Sjostro & Eriksson (2001) stated that the results showed no significant difference whether the Y-matrix is deflated or not. Therefore, for simplicity, the X-matrix will be the only one deflated. Mathematically, the process of deflating the X-matrix are as follows:

$$p = \frac{X't}{t't}$$
$$X = X - tp'$$
$$Y = Y - tc'$$

After completing all these steps, the machine will now then determine if there are still any significant information in X about Y. If there is, then the process will start again from the top until cross validation. Otherwise, the process will end here.

The discussed flow is strictly the Partial Least Squares (PLS) approach. With regards to the regression applied, it employs the concept of Multiple Linear Regression (MLR). Mathematically, it is defined as:

$$Y = XC + F$$

where Y is the derived Y-matrix earlier, X is the X-scores, C is the weights and F are the residuals from the Y-matrix (Mevik and Wehrens, 2020).

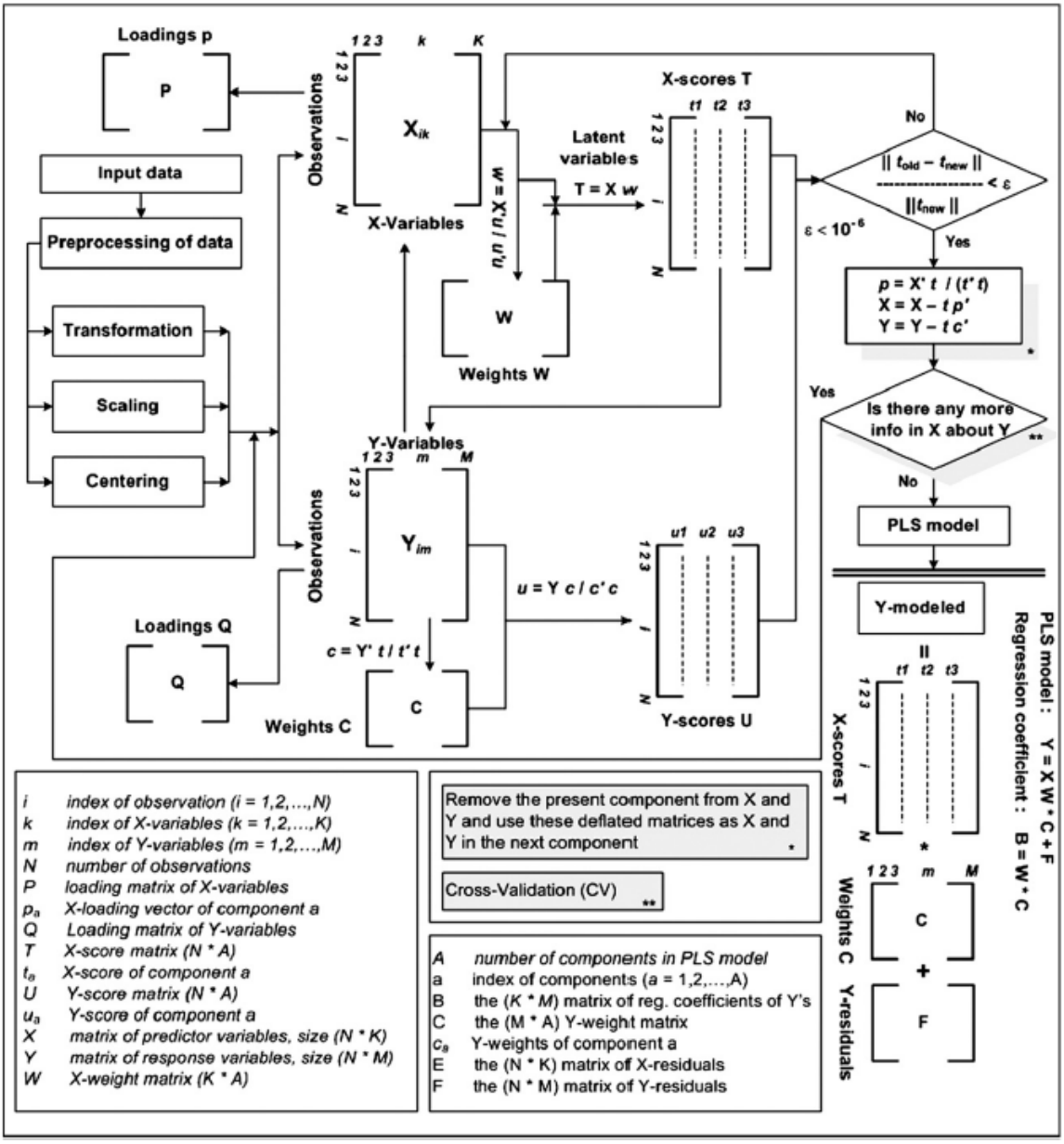


Figure 3: Flowchart of PLSR Algorithm from Farifteh et al. (2006)

2.2.5. Partial Least Squares Discriminant Analysis

The Partial Least Squares Discriminant Analysis is a procedure of classification that employs the partial least squares regression (Kucheryavskiy, 2021).

In this method of discrimination, the process of dimension-reduction for class distinction follows the among-groups variability. Hence, in a case wherein dimension-reduction is needed and the differences among the groups are not that distinct, the PLS-DA is much preferred over the PCA as a classification method (Barker & Rayens, 2002).

As stated by Barker and Rayens (2002), discriminant analysis using PLS is dependent on the "between-groups sums-of-squares and cross-products matrix" in dimension reduction thus, proving the superiority of the PLS over the PCA as a classification model in groups with significant variance. Specifically, the PLSDA uses the LDA construct.

Nocairi, et al. (2003) further discuss the PLS-DA model. This type of analysis tends to shrink the matrix T^{-1} derived from the data into an identity matrix. Afterward, eigenstructures will be formulated, which in turn will determine the correlation or co-variance coefficients that will be used in classifying such data.

2.2 Related Studies

2.2.1 FTIR Spectroscopy for Biomarker Identification

2.2.1.1 ATR-FTIR spectroscopy for virus identification: A powerful alternative

Virus-borne diseases are among the most serious public health issues. There are an uncountable number of viruses circulating in our world, several of which are now known to the science community and many of

which are unknown. Human Immunodeficiency Virus (HIV) and arboviruses such as Dengue, Zika, Chikungunya, and Yellow Fever are examples of well-known viruses that cause significant harm to humanity, either due to their severity or their capacity to adapt, giving birth to new serotypes. The technique of attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR) is a well-known spectroscopic technique that operates in the mid-infrared range. The 4000 to 400 cm-1 range of the electromagnetic spectrum is covered by this area. The biofingerprint area in biological samples is defined as the range between 1800 and 900 cm-1 because it contains a high density of information about essential biomolecules. The ATR-FTIR infrared spectra are an example of multivariate data. This is due to the fact that each continuum contains many variables (wavenumbers) that influence the absorbance of a sample. Multivariate analysis methods, which provide computational and mathematical instruments capable of interpreting data and providing accurate quantification or classification responses, can be used to help analyze this category of data. Since no reagents are used, ATR-FTIR spectroscopy can be a cost-effective option in terms of both time and money. Virus infections include complex pathways that result in modifications in the configurations of biomolecules and, as a result, spectral variations. These modifications, of course, can make ATR-FTIR detection of viral infections more difficult; however, if adequately examined, they may provide useful information on infection level. For instance, we can investigate what happens to viral RNA during the infection process (Santos et al., 2021).

2.2.1.2 Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity

A new strain of coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused the COVID-19 virus in

early2020, sparking a global pandemic. Some countries, like South Korea, battled the COVID-19 epidemic successfully at first. This is based on the following critical elements: (a) prevention through excellent cleaning procedures and isolation of possible cases; (b) testing to identify infected individuals and accurately isolate risk cases; and (c) antiviral therapy and, in the future, a vaccine. Testing is critical for identifying sick persons and high-risk areas. This allows for rational isolation of areas without damaging a whole country's economy and allows for more strategic deployment of resources to battle the sickness, with more ventilators, medicines, and medical staff assigned to places with more diagnosed cases. The cost and, in particular, the time required for each test result are the key hurdles for testing. Even in industrialized nations, gold-standard diagnosis by RT-qPCR is expensive, with a scarcity of testing facilities, and can take less than two days to obtain a result since specimens must be transferred for processing to frequently distant laboratories. This is not appropriate for large-scale testing.

Vibrational spectroscopy, especially attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy, has been widely utilized to distinguish and categorize normal and abnormal populations utilizing a variety of cell types, tissues, and biofluids. Because of their ease of collection and little sample preparation, readily available biofluids such as blood plasma/serum, saliva, or urine are thought to be excellent for clinical use. Interrogation of samples with infrared (IR) spectroscopic methods results in the formation of a "spectral fingerprint," which aids in the differentiation of various populations and the identification of possible biomarkers. In recent years, biofluid-based ATR-FTIR spectroscopy has been employed for diagnosing, screening, and monitoring disease progression/regression in a number of disorders. Spectroscopic procedures are quick, inexpensive, and non-destructive, making them ideal for

translation to the clinic, even as a complement to more established approaches (Barauna et al., 2021).

2.2.1.3 Saliva analysis using FTIR spectroscopy to detect possible SARS-CoV-2 (COVID-19) virus carriers

Despite the approximate percentage of accuracy that it permits (72 percent), the use of saliva in the diagnosis of COVID-19 using the Reverse Transcription Polymerase Chain Reaction (RT-PCR) methodology has been employed primarily because it is a less invasive procedure. Fourier transform infrared spectroscopy (FTIR) is a technique for examining the molecular structure of a sample using a signal or FTIR spectrum created by the vibrations of the chemical bonds that make up said sample when touched by infrared radiation frequencies. Because viral infections may be identified by FTIR spectroscopy by causing molecular changes, this approach may be used to do exams more quickly than RT-PCR (72 hours on average), since the spectrum capture takes around 15 minutes, including the drying period of the sample. It has been feasible to detect typical alterations of several forms of cancer, Parkinson's disease, and diabetes, among other diseases, using different parts of the FTIR spectrum; nevertheless, viral infections mostly influence the protein area, notably the region assigned to amide I. (1700-1600 cm-1). The growth or decay in a section of the FTIR spectrum reflects spectral changes, hence modifying the spectrum using derivatives helps to emphasize such changes (Sánchez-Brito et al., 2021).

2.2.1.4 Infrared Spectroscopy of Proteins

Proteins can be subjected under Infrared Spectroscopy. The researchers utilized reaction-induced infrared difference spectroscopy to prompt reactions to various protein molecules which in turn provides numerous characteristics of the molecules specifically, chemical structure of the vibrating group, chemical properties of neighboring groups in a molecule, Redox state, bond parameters, bond angles, conformation,

hydrogen bonding, electric field and conformational freedom. In the paper, the researcher provided a table encompassing numerous amino acid side chains alongside some remarks made with respect to the properties obtained from the used infrared spectrometer (Barth, 2007).

2.2.1.5 Enabling quantification of protein concentration in human serum biopsies using attenuated total reflectance – Fourier transform infrared (ATR-FTIR) spectroscopy

The serum samples were obtained at the Biochemical laboratory at the University Hospital CHU Bretonneau de Tours. The researchers prepared the samples in two ways: Whole Dilution Study, which makes the samples into 2-fold dilutions, and Spiked Human Serum Models which will both be the standards used in determining the reliability of the ATR-FTIR (Attenuated Total Reflectance – Fourier Transform Infrared) in quantifying. The protein levels were then measured by a tool known as the Cobas 6000 analyzer series. The ATR-FTIR spectra needed to quantify the Human Albumin Serum and Immunoglobulin were obtained by a machine known as the Bruker Vector 22. The samples to be used in the ATR-FTIR were prepared by different approaches: liquid, air dried and diluting them by 10% using deionized water. The process of measurement was repeated five times per type of prepared sample. Matlab was used in analysis of the obtained data. A software from the University of Strathclyde was used to remove other non-biochemical components to obtain a clearer spectrum. Partial Least Squares Regression (PLSR) Analysis was used in quantifying the protein concentrations and providing an estimate in the serum protein levels. Validating the analysis is done by using the Root Mean Square Error (RMSE). Aside from this, the dataset was then further validated using the blind testing model. It is discerned that air dried samples produced the best results since the IR light has a very strong water absorbance. Low concentrations also showed that integrated absorbance is directly proportional with the concentration but above 30% dilution, variations

among these variables are non-linear. By assessing the data via the PLSR, it is also concluded that dilution is an important process in ensuring the protein absorbance's range of validity in the Beer-Lambert Law (Liu, Shi & Mantsch, 2005).

2.2.1.6 Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning

Due to the fact that it is straightforward, label-free, and cost-effective, Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy in conjunction with machine learning approaches might be a viable alternative method for diagnosing COVID-19. This approach has shown potential as a diagnostic or screening tool for a variety of disorders including cancer, diabetes, hypertension, and physiological stress. It was reinforced by a research published in 2018 by Leal et al. and Baker et al., which found that tiny samples of bio fluids (saliva, blood, and urine) might be helpful for the detection of a wide spectrum of disorders, including infectious diseases. In this proof-of-concept investigation, we determined that FT-IR spectroscopy combined with artificial intelligence in nasopharyngeal swab suspension fluid was successful for distinguishing between COVID-19 positive and negative individuals (Nogueira et al., 2021).

2.2.2 Chemometric Analysis in Spectroscopy

2.2.2.1 Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy

Due to the wide variety of pre-processing methods available for chemometric analysis, FTIR spectroscopy is not commonly used in the biological field. The population of professionals seem to be unable to come to a consensus as to what is the best approach. Hence, they came to the

conclusion that choosing the most appropriate pre-processing method is highly reliant on the priori knowledge of the sample and spectral response.

Despite hindering the unification of these two fields, pre-processing methods are a necessity in spectral analysis. It helps in the reduction of unwanted variance derived from light scattering and noise which will in turn improve the resulting analysis, classification and interpretability of the input spectrum.

Aside from the sample to be tested, pre-processing methods are at times technique-specifiic. For instance, Raman spectroscopy employs cosmic ray removal as its pre-processing method. In general, pre-processing methods include binning, smoothing, normalization and baseline correction. The study will determine the most viable pre-processing methods for samples obtained from cancer patients through an algorithm known as the Random Forest Classification.

Binning is defined as the process of determining the average adjacent data points which helps in reducing the dataset's dimensions. Filtering, on the other hand, is the method of noise reduction in the dataset. The most common filter used is the Savitzky-Golay filter. Normalization is administered in order to reduce the intrinsic differences among samples. Min-max scaling and vector normalization are the most used in FTIR spectroscopy. Lastly, baseline correction is a fundamental technique in reducing light scattering. In FTIR spectroscopy, rubberband baseline correction is used most of the time.

Upon administering every possible pre-processing method to the obtained dataset, the researchers concluded that it is best that the pre-process is kept to the minimum. This conclusion is drawn upon by comparing the output of each pre-processing method in terms of its prediction value, Matthew's correlation coefficient, specificity, sensitivity, positive and negative predictive values. The best method is the vector

normalization followed by a second derivative filter. They also concluded that the pre-processing method truly is vital in spectral analysis since differences in the analyses between those datasets that underwent pre-processing and those that did not are noticeable (Butler et al., 2018).

CHAPTER 3

Methodology

3.1 Theoretical Framework

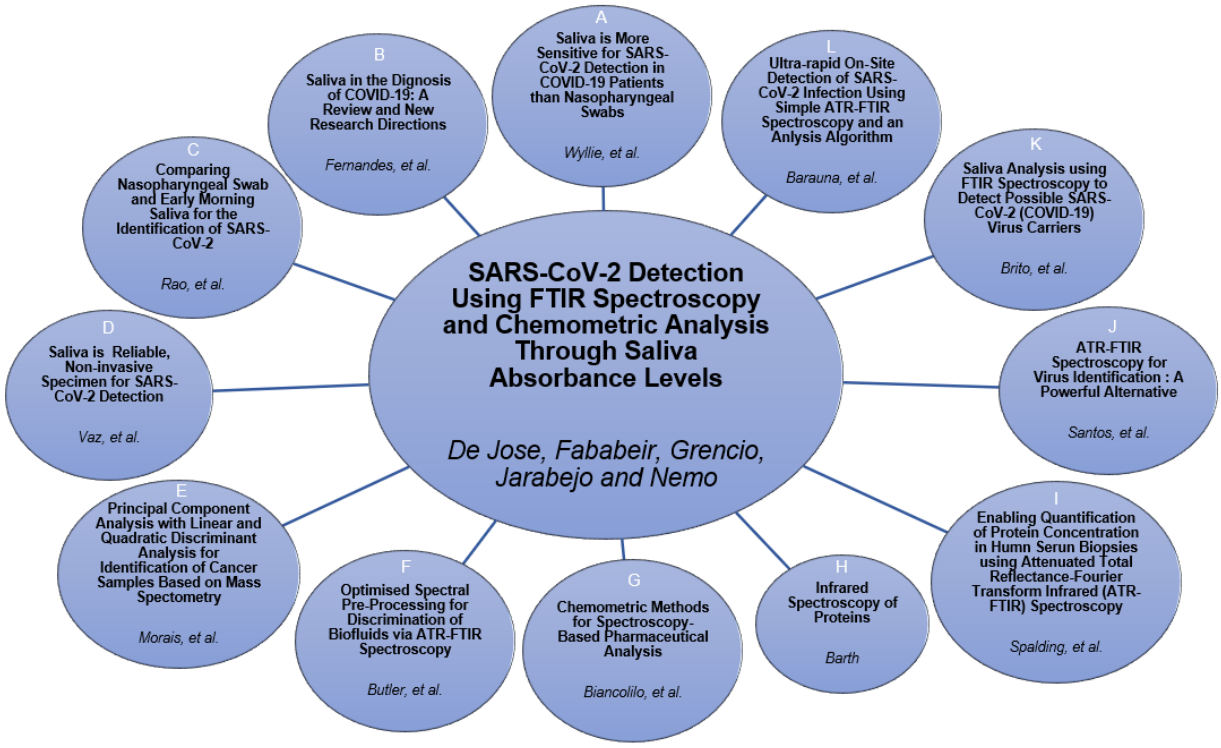


Figure 4: Block diagram of related studies

Figure 4 shows the studies that are correlated to make the research study possible. Each research study from Figure 3 gave the researchers ideas and concepts to come up with such as using Saliva to detect SAR-CoV-2, the use of Fourier transform infrared spectroscopy to get the data, and the use of different chemometrics analysis techniques.

Research studies from Figure 3 A-D proves that the virus SAR-CoV-2 can be detected using saliva. This will be a better way to collect COVID-19 specimens without risking the health of healthcare workers and it is less costly. Studies from Figure 3 E-G shows that the sample

gathered from the FTIR spectroscopy can be used in a different Chemometric Model while Figure 3 H-L shows that the SAR-CoV-2 can be determined using a FTIR spectroscopy.

3.2 Conceptual Framework

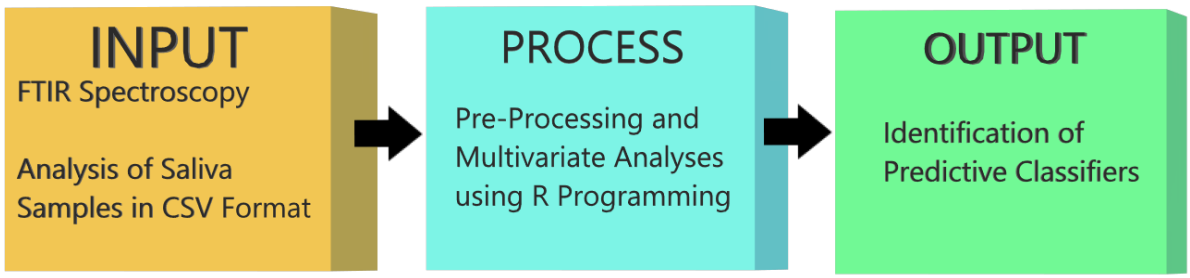


Figure 5: IPO For Chemometric Model

The input-process-output or IPO of Chemometric Analysis of the study is shown above. The input will be a CSV format of an FTIR spectroscopy analysis of a saliva sample. The input will be the data to be analysed by the predictive model obtained from the chemometric analyses which first performs a pre-processing to filter noise in the spectrum and then a multivariate model was optimized until classifiers are identified. The predictive model has certain thresholds with detection of the virus. When these thresholds are met, the diagnosis is positive.

3.2.1 Block Diagram

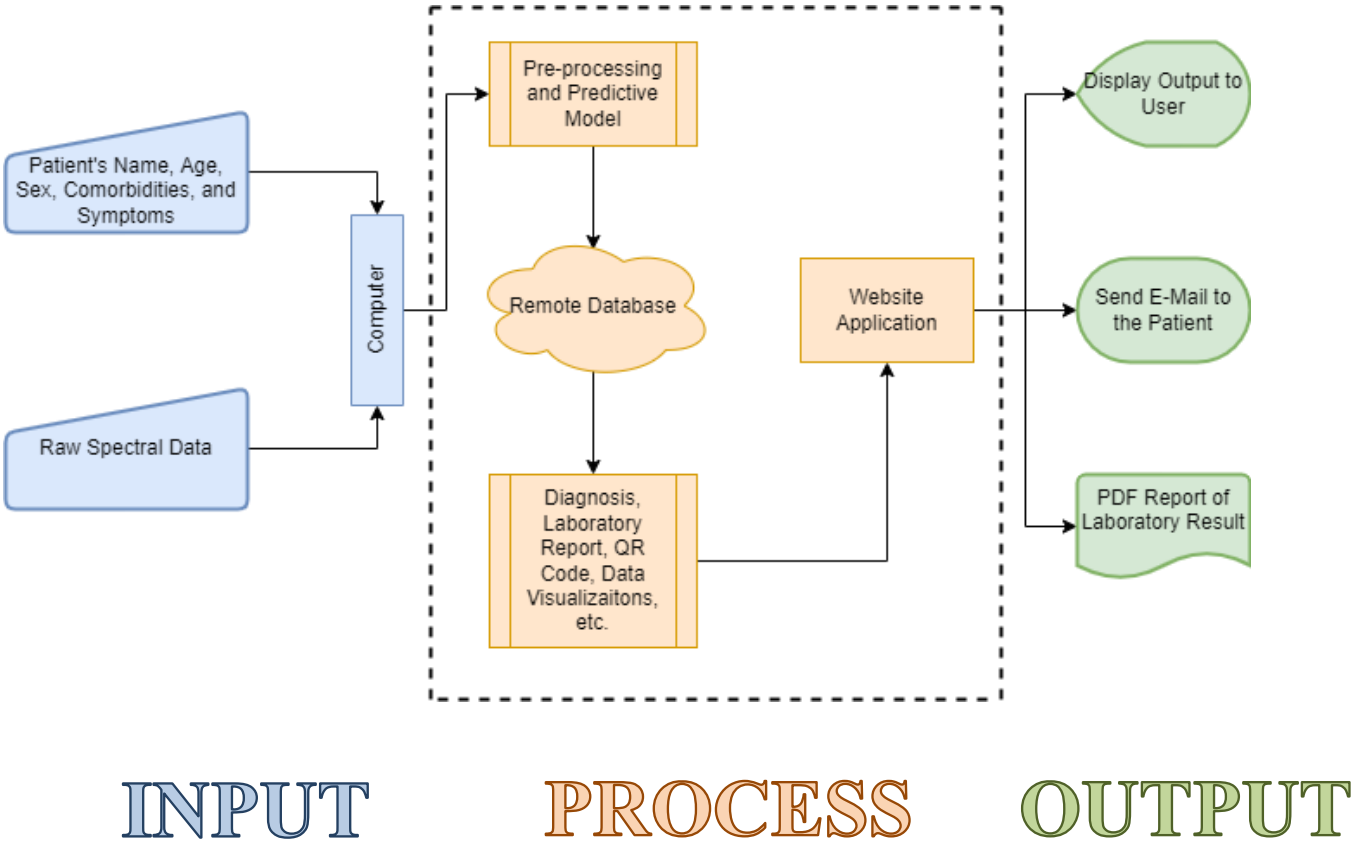


Figure 6: Simplified Block Diagram of the Present Study

The proposed system will start with the user's manual input of the patient's basic information such as their name, age, sex, comorbidities, and symptoms. The raw spectral data generated by the FTIR will also be manually inputted by the user. After inputting all the needed data, the input spectral data will be undergoing pre-processing. This pre-processed spectrum will be analyzed by the predictive model in order to determine the diagnosis for the patient i.e., if the patient is positive or negative from SARS-CoV 2. A remote database will then store all the acquired results. The aforementioned results encompass the patient's basic information, date when the laboratory report is generated, diagnosis on the patient, and other relevant information. Not only will these results be

displayed real-time through the website application's main panel, but it will also be available in a Portable Document File that can be downloaded by the user into their computer. This PDF laboratory report can then be sent to the patient's e-mail.

3.2.2 Network Architecture

Figure 7: Network Diagram

The figure seen above encompasses the network diagram for the project. With the help of shinyapps.io, the Shiny Server can be used by the proponents for deploying their website application COVID AppTect. ShinyApps.io does not register a dedicated address to the website application, instead it assigns it with a unique ID similar to a Virtual Private Cloud. The users can access this website application as long as their device has a browser installed in it.

The report of the patients can be stored with the help of GoogleSheets which will serve as a remote file storage of these results. This will help in retaining the past results

obtained by the user upon using the website application. This also enables the persistent data storage function in the website which help employs nonvolatile storage as long as the user stores their past data into the database.

3.3 Research Locale

The website application will be evaluated by the medical technologists, health practitioners, and research specialists working in the National Capital Region (NCR). They will be provided with CSV files obtained from the dataset of the researchers which will be used for the evaluation of the web application.

3.4 Research Design

3.4.1 Software Design

3.4.1.1 Algorithm Selection

For the predictive model of the website application, three types of algorithms will be used in order to compare the sensitivity, specificity and accuracy of each. Among these proposed algorithms are the Principal Component Analysis-Quadratic Discriminant Analysis (PCA-QDA) and the Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA) wherein both makes use of the principal components derived from PCA but uses different Gaussian Discriminant Analysis (GDA) for these components as discussed earlier. Another algorithm to be used will be the Partial Least Squares Discriminant Analysis (PLSDA) Algorithm wherein the derived Partial Least Squares (PLS) Model will be regressed with respect to the obtained scores and loadings.

After numerous retests, it is discerned that the Principal Component Analysis-Quadratic Discriminant Analysis (PCA-QDA) is the most optimal predictive model to use for the website application.

3.4.1.1.1. Pre-process Algorithm Implementation

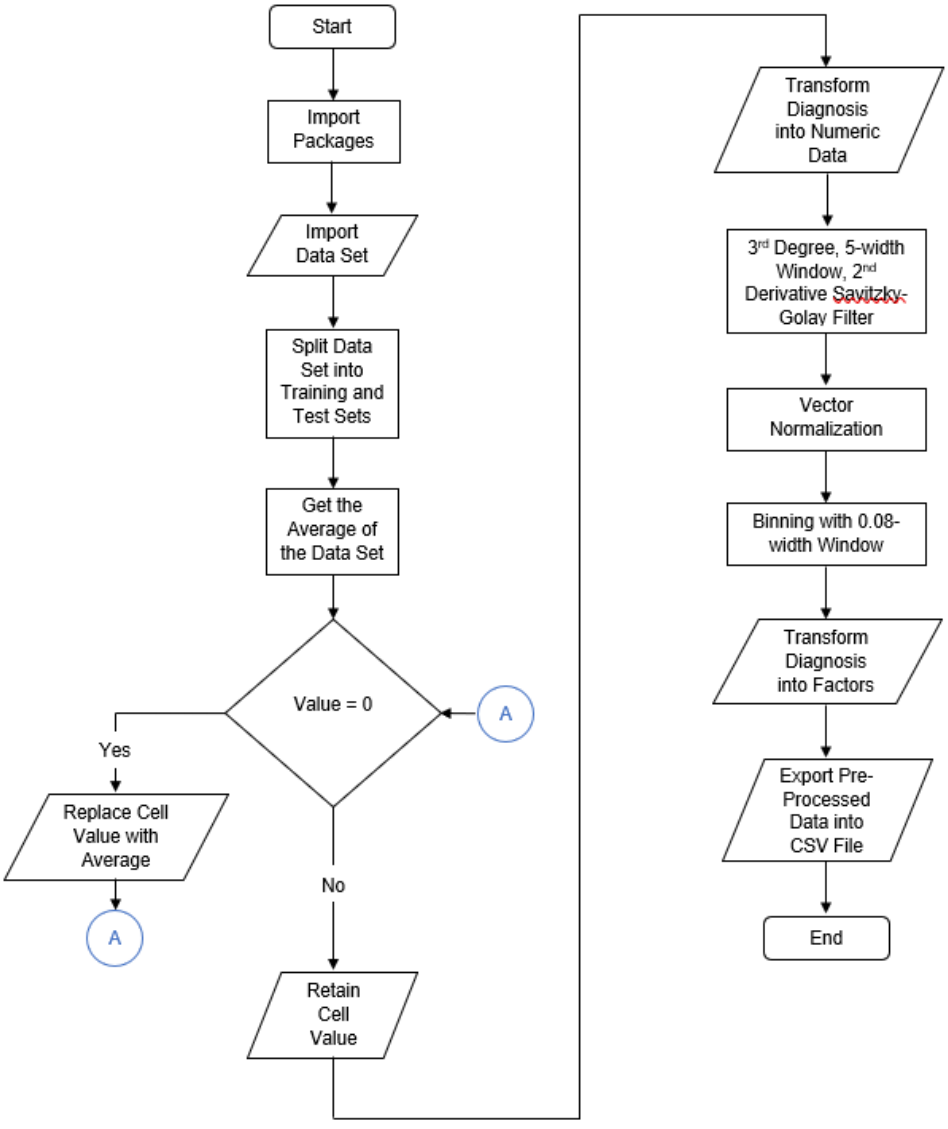


Figure 8: Pre-Process Flowchart

The flowchart presented above shows the process of the created syntax for the pre-processing model used by the website application. After importing the packages and the data set to be used, the average of this data set will be computed. Afterward, the model will proceed in determining the N/A or zero values within the data set. If it is equal to zero, then the model will proceed in replacing this value with the computed average from earlier. If it is not, then the model will retain this value. Afterward, the diagnoses N and P will be transformed into numeric data 1 and 2, respectively. Next, a 3rd Degree, 5-width Window, 2nd Derivative Savitzky-Golay Filter will be applied to the data set. After the filter, vector normalization and a 0.08-width bin will be used to further clean the data. Next, the model will now revert the diagnoses back into factors. Lastly, the pre-processed data will be exported into a CSV file format.

3.4.1.1.2. PCA-LDA Algorithm Implementation

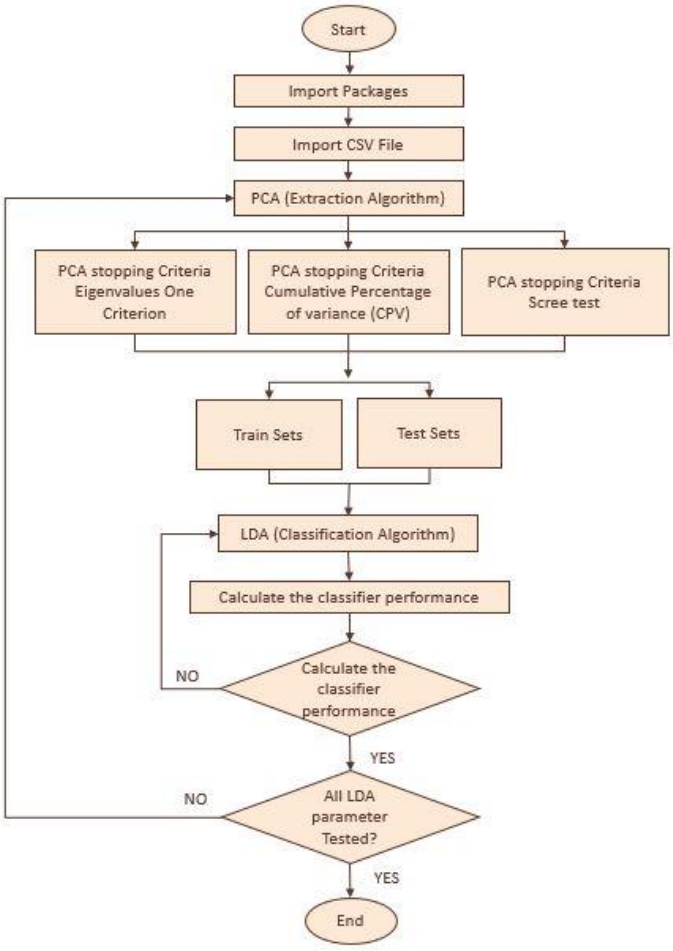


Figure 9: PCA-LDA Flowchart

3.4.1.1.3. PCA-QDA Algorithm Implementation

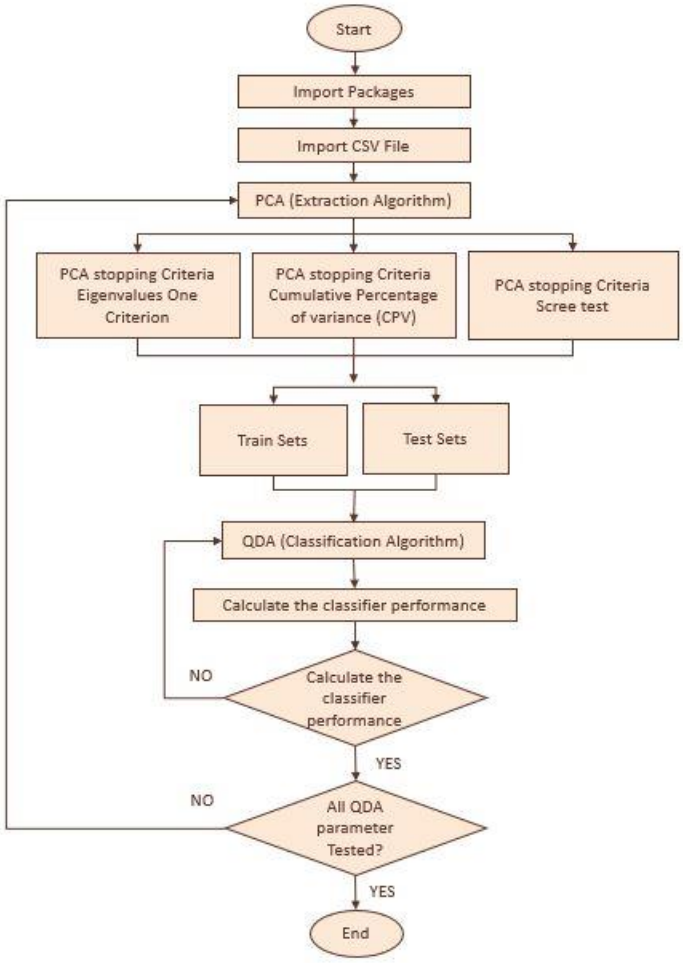


Figure 10: PCA-QDA Flowchart

The PCA-QDA and PCA-LDA both require selection of principal components before subjecting the QDA and LDA functions to the program code. The overall algorithm is adopted from the Predicting breast cancer using PCA + LDA in R by Shravan Kuchkula in kaggle.com. Based on Figures 8 and 9, necessary packages and the csv dataset file are imported first. After that, principal components are obtained by implementing the specific function for it. The PC scores will then be analyzed using the stopping criteria in which eigenvalues, percentage variance,

cumulative percent and their scree plots are observed and evaluated based on the retainment rules. The retained PC scores will be divided into test and train sets. These sets will proceed to running the LDA and QDA function. When the LDA and QDA are obtained, ROC curves can be plotted already. Also, quality performance can be run with the designated functions in R for them.

3.4.1.1.4. PLS-DA Algorithm Implementation

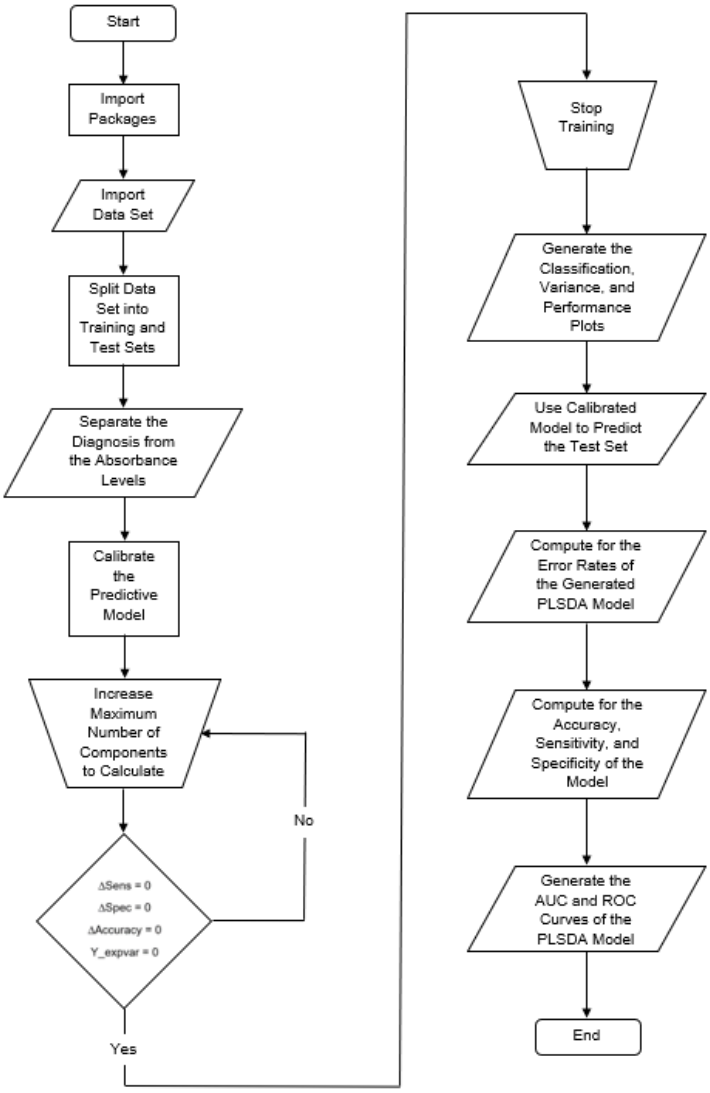


Figure 11: PLS-DA Flowchart

The figure above encompasses the flowchart of the implemented PLS-DA model on the web application. The proponents first determined the optimal number of components that the model will compute to ensure the best results. When there are no more observed variations between the sensitivity, specificity, and accuracy of two consecutive components, the incrementation will stop. Another factor of this process is the explained variance, which must be equal to zero. For the web application, it is discerned that using 20 components for the predictive model is the most optimal solution. The training will stop, then the classification, variance, and performance plots will be generated. The implemented model will then be fed with the test set. The resulting predictions created by the model will be used in computing the error rate, accuracy, sensitivity, and specificity of the PLS-DA model. For further examination of the predictive model's performance, the AUC and ROC curves are plotted.

3.4.1.2 Web Application ("PRISMA App")

For this research, the proponents will develop a website application through Shiny, a website application package in RStudio that will be implemented through RStudio's dedicated web server for Shiny Apps, shinyapps.io. With its ability to directly translate R codes into an interactive website application, users can easily take advantage of R's tantamount of statistical packages in their websites (Github, 2020). Since the proposed study will mainly focus on data statistics and analyzation through chemometrics, the proponents thought of this as the most viable tool for

their website application development. Figures 37 and 38 show the flowchart of the web application.

1) *User Authentication:* To ensure the user's privacy and data security, those who wish to use the application must first login using their credentials that they inputted on the sign up page or use their existing Google Accounts.

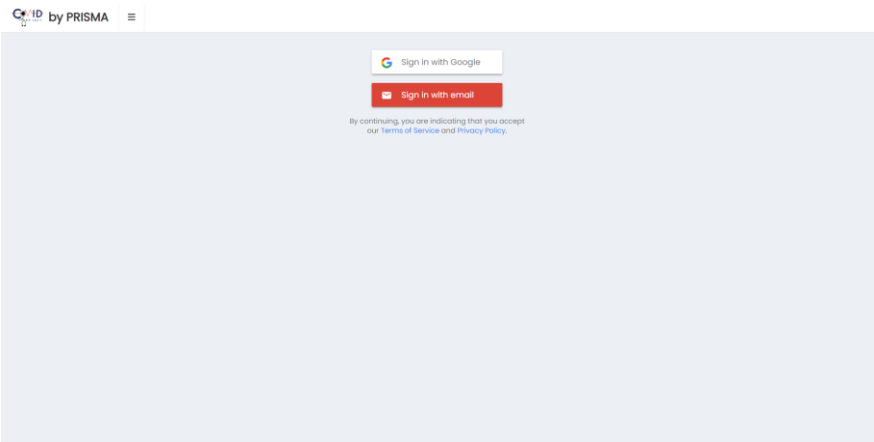


Figure 12: Screenshot of User Authentication Page

2) *User Privacy Data Agreement Notice:* Since the website application will acquire sensitive information from the user, this will inform the user that the acquired data will strictly be for analysis purposes and will be protected under the Data Privacy Act.

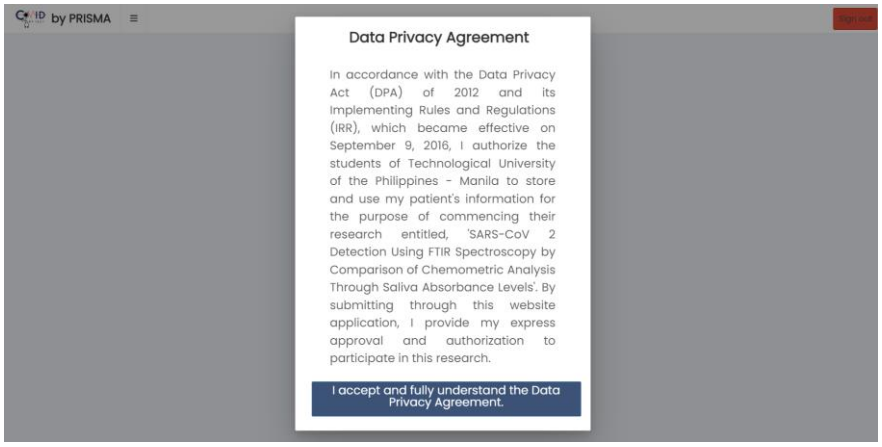


Figure 13: Screenshot of User Privacy Data Agreement Notice

3) *Homepage:* This is where the user will be redirected to after they logged in to the website application. The homepage includes a brief introduction of the project, as well as links for further information of the study. A video walkthrough can also be seen which server as a tour of the website application's entirety.

Figure 14: Screenshot of the Homepage

4) *Predictive Model Calibration:* Since the website application is a predictive model, constant calibration is needed in order to ensure that its output is accurate. Users can choose which dataset they will input to the website application. They also have an option to preview their chosen dataset. The results from the Calibration Set can be seen in Figure 16.

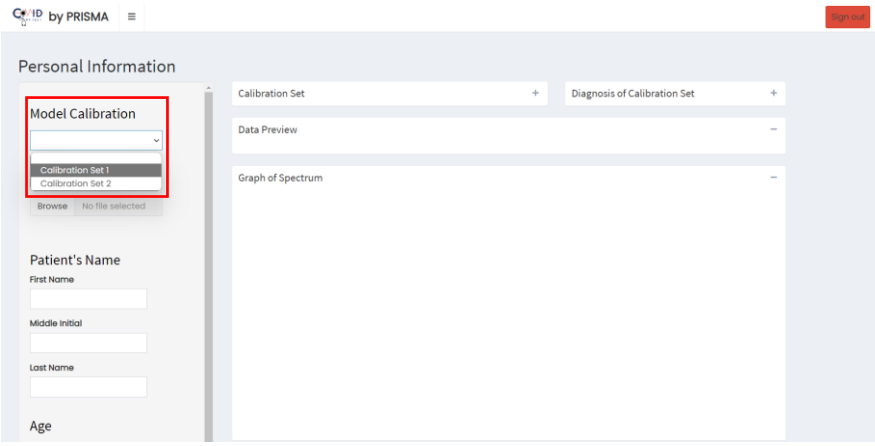


Figure 15: Screenshot of the Testing Tab

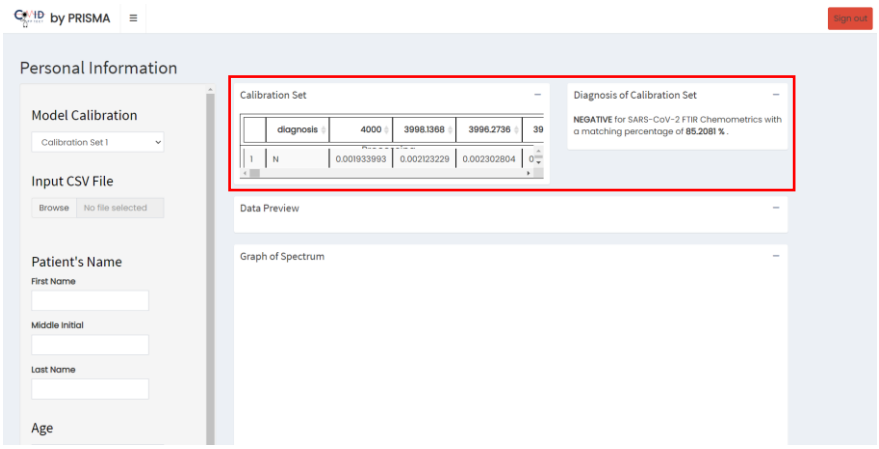


Figure 16: Screenshot of the Testing Tab

5) *Upload CSV File:* Since the raw spectral data must be inputted by the user themselves, a button where the user can upload their files is included.

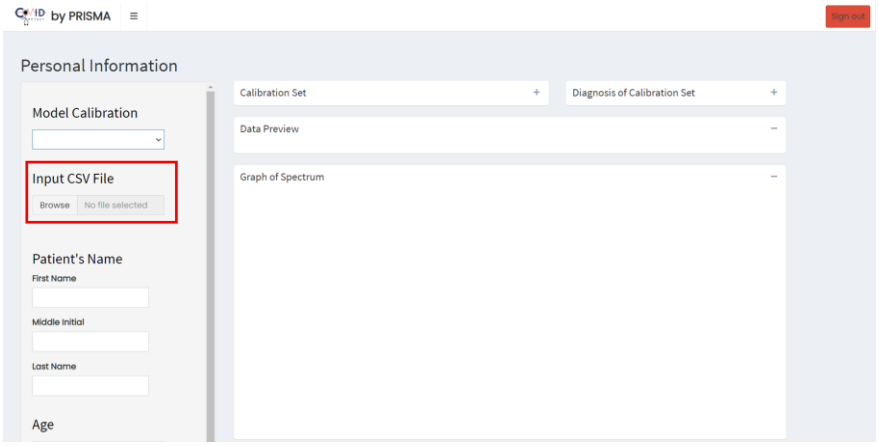


Figure 17: Screenshot of the Testing Tab

6) *Basic Information of the Patient:* The website application has an area wherein it allows the user to input the Patient's Name, Age, Birthdate, Sex, General Location, Comorbidities, and Felt Symptoms for it to be included in the database and in the laboratory report.

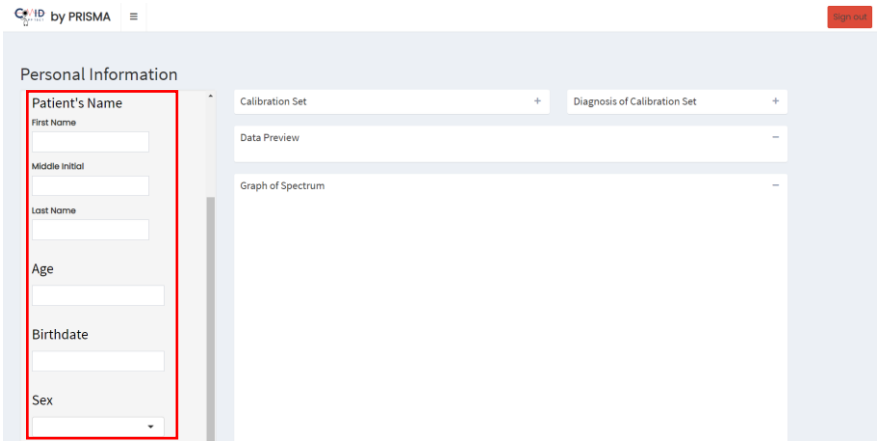


Figure 18: Screenshot of the Testing Tab

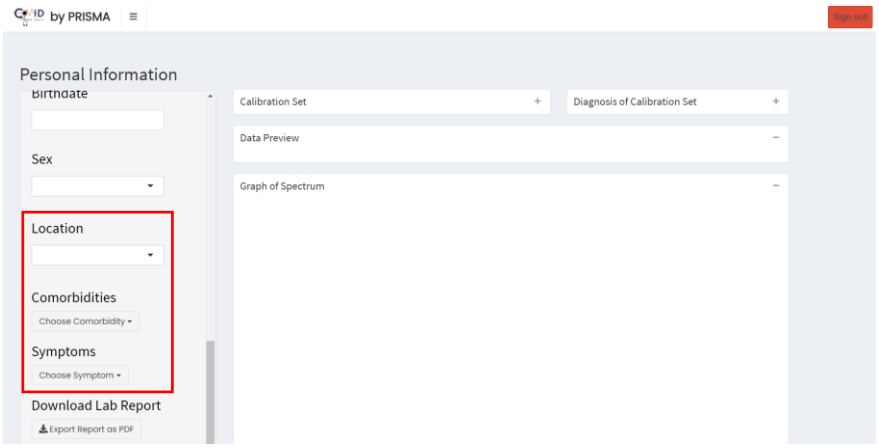


Figure 19: Screenshot of the Testing Tab

7) *MainPanel:* This is where all output is displayed real-time for the user to ensure that their input is correct. Everything displayed within the red box will be reflected on the generated PDF format of the laboratory report, while those in the blue box are for ensuring the consistency of the website application.

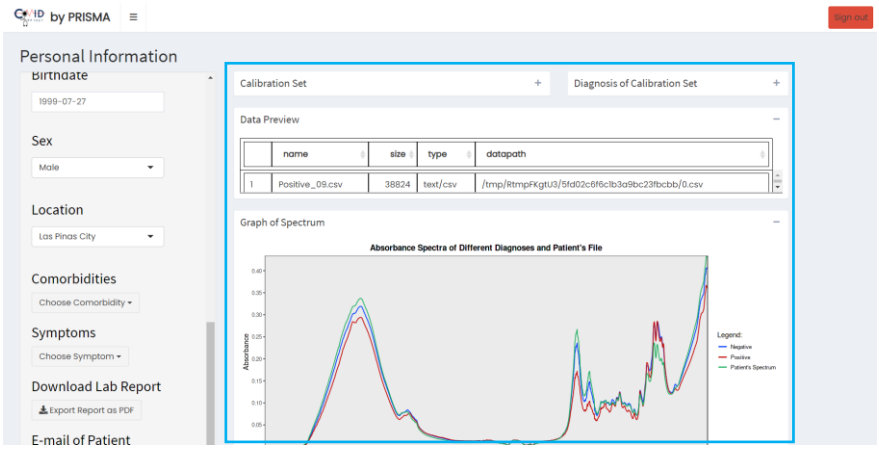


Figure 20: Screenshot of the Testing Tab

Figure 21: Screenshot of the Testing Tab

8) *QR Code Generator:* This feature automatically translates the created laboratory report into a QR Code for easier readability.

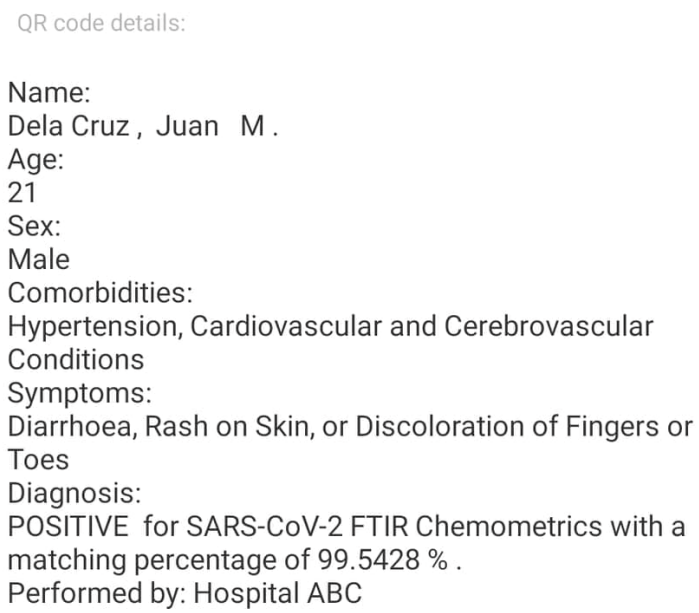


Figure 22: Screenshot of Output from QR Code in Figure 20

9) *Spectra Graph Generator*: This feature allows the user to closely examine the input CSV file of the sample to test in graphical form. It is plotted alongside a Negative and Positive Sample, each with a percentage match of 100%.

Figure 23: Screenshot of the Testing Tab

10) *Laboratory Report Generator:* This feature allows the user to immediately export their created laboratory report into a Portable Document File Format.

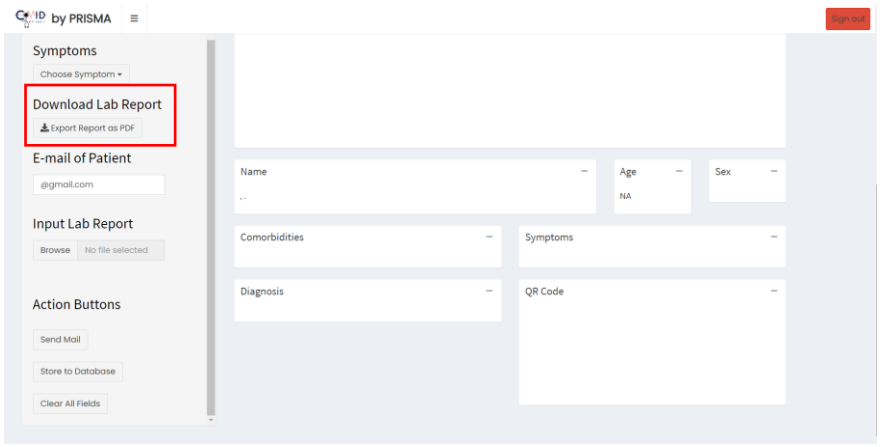


Figure 24: Screenshot of the Testing Tab

Figure 25: Screenshot of Sample Laboratory Report in PDF

10) *E-Mailing System:* This feature allows the user to upload the downloaded laboratory report and then send it to the patient.

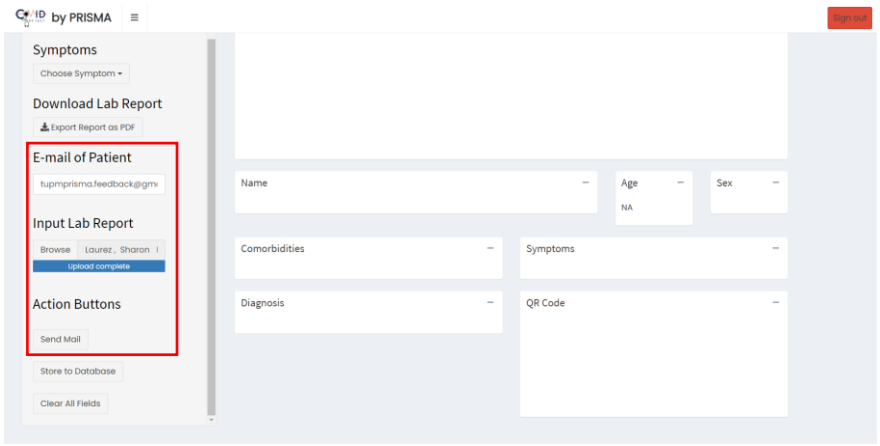


Figure 26: Screenshot of the Testing Tab

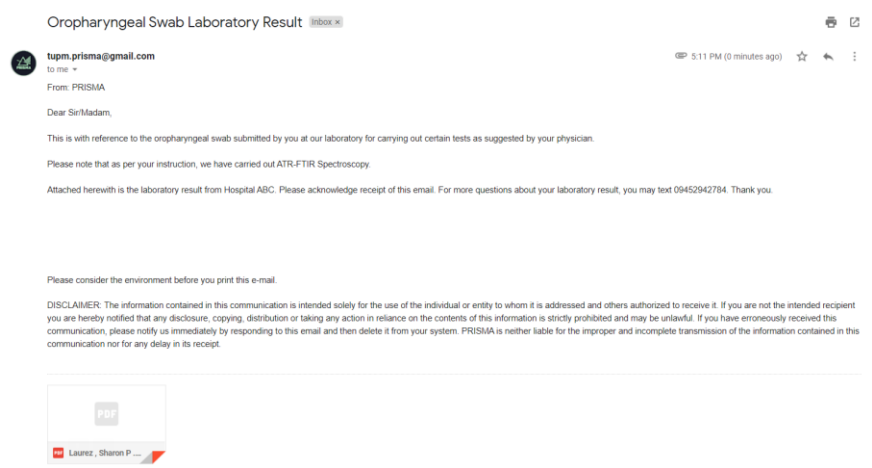


Figure 27: Screenshot of Sample E-Mail of Laboratory Report

9) *Action Buttons*: This feature lets the user to store their generated laboratory report and then clear all fields to prepare for the next laboratory report to create.

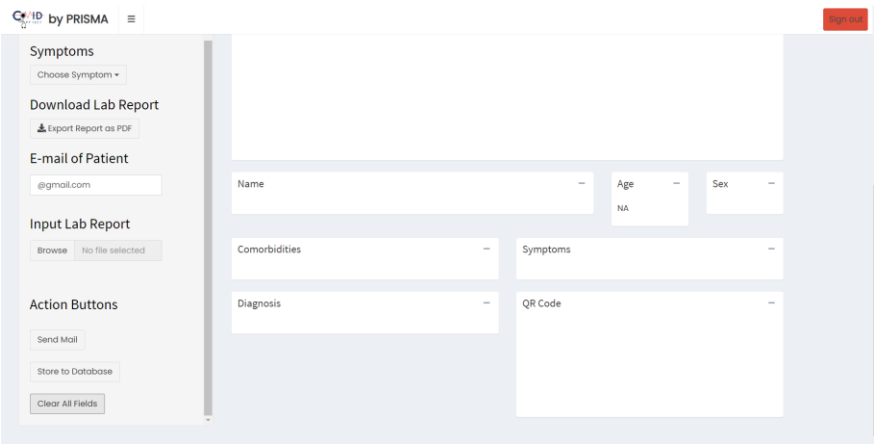
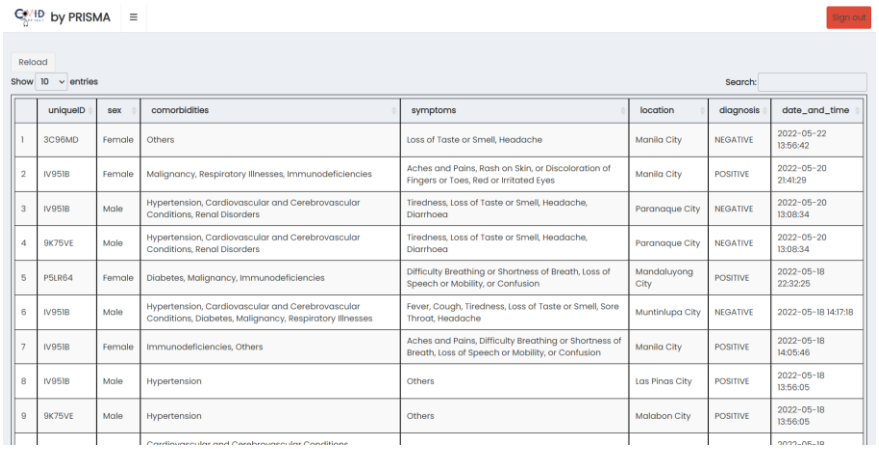


Figure 28: Screenshot of the Testing Tab

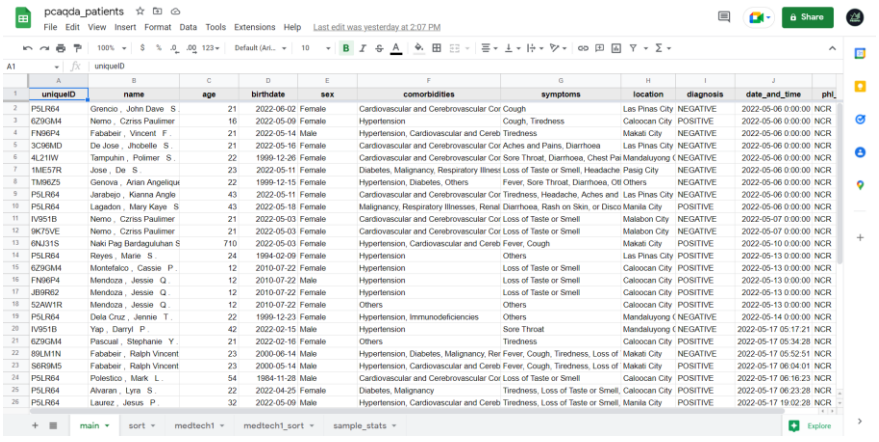
10) *Laboratory Report Inventory:* This feature enables the user to view their past generated laboratory reports. This feature enables the user to view their past generated laboratory reports.



	uniqueID	sex	comorbidities	symptoms	location	diagnosis	date_and_time
1	3C96MD	Female	Others	Loss of Taste or Smell, Headache	Manila City	NEGATIVE	2022-05-22 13:56:42
2	IV951B	Female	Malignancy, Respiratory Illnesses, Immunodeficiencies	Aches and Pains, Rash on Skin, or Discoloration of Fingers or Toes, Red or Irritated Eyes	Manila City	POSITIVE	2022-05-20 21:41:29
3	IV951B	Male	Hypertension, Cardiovascular and Cerebrovascular Conditions, Renal Disorders	Tiredness, Loss of Taste or Smell, Headache, Diarrhoea	Paranaque City	NEGATIVE	2022-05-20 13:08:34
4	9K75VE	Male	Hypertension, Cardiovascular and Cerebrovascular Conditions, Renal Disorders	Tiredness, Loss of Taste or Smell, Headache, Diarrhoea	Paranaque City	NEGATIVE	2022-05-20 13:08:34
5	P5LR64	Female	Diabetes, Malignancy, Immunodeficiencies	Difficulty Breathing or Shortness of Breath, Loss of Speech or Mobility, or Confusion	Mandaluyong City	POSITIVE	2022-05-18 22:32:25
6	IV951B	Male	Hypertension, Cardiovascular and Cerebrovascular Conditions, Diabetes, Malignancy, Respiratory Illnesses	Fever, Cough, Tiredness, Loss of Taste or Smell, Sore Throat, Headache	Muntinlupa City	NEGATIVE	2022-05-18 14:17:18
7	IV951B	Female	Immunodeficiencies, Others	Aches and Pains, Difficulty Breathing or Shortness of Breath, Loss of Speech or Mobility, or Confusion	Manila City	POSITIVE	2022-05-18 14:05:46
8	IV951B	Male	Hypertension	Others	Las Pinas City	POSITIVE	2022-05-18 13:56:05
9	9K75VE	Male	Hypertension	Others	Malabon City	POSITIVE	2022-05-18 13:56:05

Figure 29: Screenshot of the Inventory Page

11) *Hashed Documents:* Since this website application will store sensitive information derived from the patients, it can be seen in Figure 29 that all displayed data does not show sensitive information of patients such as their name, birthdate, etc. These are only visible in the side of the administrator.



	uniqueID	name	age	birthdate	sex	comorbidities	symptoms	location	diagnosis	date_and_time
2	P5LR64	Grencio , John Dave S	21	2022-06-02	Female	Cardiovascular and Cerebrovascular Cor	Cough	Las Pinas City	NEGATIVE	2022-05-06 0:00:00
3	6Z9GM4	Nerno , Czriss Paulmer	16	2022-05-09	Female	Hypertension	Cough, Tiredness	Caloocan City	POSITIVE	2022-05-06 0:00:00
4	FN96P4	Fababeir , Vincent F	21	2022-05-14	Male	Hypertension, Cardiovascular and Cereb	Tiredness	Makati City	NEGATIVE	2022-05-06 0:00:00
5	3C96MD	De Jose , Jhobelle S	21	2022-05-16	Female	Cardiovascular and Cerebrovascular Cor	Aches and Pains, Diarrhoea	Las Pinas City	NEGATIVE	2022-05-06 0:00:00
6	4L21IW	Tampuhin , Polimer S	22	1999-12-26	Female	Cardiovascular and Cerebrovascular Cor	Sore Throat, Diarrhoea, Chest Pai	Mandaluyong (	NEGATIVE	2022-05-06 0:00:00
7	1ME57R	Jose , De S	23	2022-05-11	Female	Diabetes, Malignancy, Respiratory Illnes	Loss of Taste or Smell, Headache	Pasig City	NEGATIVE	2022-05-06 0:00:00
8	TM96Z5	Genova , Arian Angelique	22	1999-12-15	Female	Hypertension, Diabetes, Others	Fever, Sore Throat, Diarrhoea, Oth	Others	NEGATIVE	2022-05-06 0:00:00
9	P5LR64	Jarabejo , Kianna Angle	43	2022-05-11	Female	Cardiovascular and Cerebrovascular Cor	Tiredness, Headache, Aches and	Las Pinas City	NEGATIVE	2022-05-06 0:00:00
10	P5LR64	Lagadon , Mary Kaye S	43	2022-05-18	Female	Malignancy, Respiratory Illnesses, Renal	Diarrhoea, Rash on Skin, or Disco	Manila City	POSITIVE	2022-05-06 0:00:00
11	IV951B	Nerno , Czriss Paulmer	21	2022-05-03	Female	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Malabon City	NEGATIVE	2022-05-07 0:00:00
12	9K75VE	Nerno , Czriss Paulmer	21	2022-05-03	Female	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Malabon City	NEGATIVE	2022-05-07 0:00:00
13	6NJ31S	Naki Pag Bardaguluhan S	710	2022-05-03	Female	Hypertension, Cardiovascular and Cereb	Fever, Cough	Makati City	POSITIVE	2022-05-10 0:00:00
14	P5LR64	Reyes , Marie S	24	1994-02-09	Female	Hypertension	Others	Las Pinas City	POSITIVE	2022-05-13 0:00:00
15	6Z9GM4	Montefalco , Cassie P	12	2010-07-22	Female	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00
16	FN96P4	Mendoza , Jessie Q	12	2010-07-22	Male	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00
17	JB9R62	Mendoza , Jessie Q	12	2010-07-22	Female	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00
18	52AW1R	Mendoza , Jessie Q	12	2010-07-22	Female	Others	Others	Caloocan City	POSITIVE	2022-05-13 0:00:00
19	P5LR64	Dela Cruz , Jennie T	22	1999-12-23	Female	Hypertension, Immunodeficiencies	Others	Mandaluyong (	NEGATIVE	2022-05-14 0:00:00
20	IV951B	Yap , Darryl P	42	2022-02-15	Male	Hypertension	Sore Throat	Mandaluyong (	NEGATIVE	2022-05-17 05:17:21
21	6Z9GM4	Pascual , Stephanie Y	21	2022-02-16	Female	Others	Tiredness	Caloocan City	POSITIVE	2022-05-17 05:34:28
22	89LM1N	Fababeir , Ralph Vincent	23	2000-06-14	Male	Hypertension, Diabetes, Malignancy, Res	Fever, Cough, Tiredness, Loss of	Makati City	NEGATIVE	2022-05-17 05:52:51
23	S6R9M5	Fababeir , Ralph Vincent	23	2000-05-14	Male	Hypertension, Cardiovascular and Cereb	Fever, Cough, Tiredness, Loss of	Makati City	POSITIVE	2022-05-17 06:04:01
24	P5LR64	Polestico , Mark L	54	1984-11-28	Male	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-17 06:16:23
25	P5LR64	Alvaran , Lyra S	22	2022-04-25	Female	Diabetes, Malignancy	Tiredness, Loss of Taste or Smell,	Caloocan City	POSITIVE	2022-05-17 06:23:28
26	P5LR64	Laurez , Jesus P	32	2022-05-09	Male	Hypertension, Cardiovascular and Cereb	Tiredness, Loss of Taste or Smell,	Manila City	POSITIVE	2022-05-17 19:02:28

Figure 30: Screenshot of the Website Application's Remote Database

12) *Login Database:* This feature enables the user of the accounts that currently has access to the website application, as well as the latest date when these accounts have logged in. The administrator can also restrict the activity of selected users here.

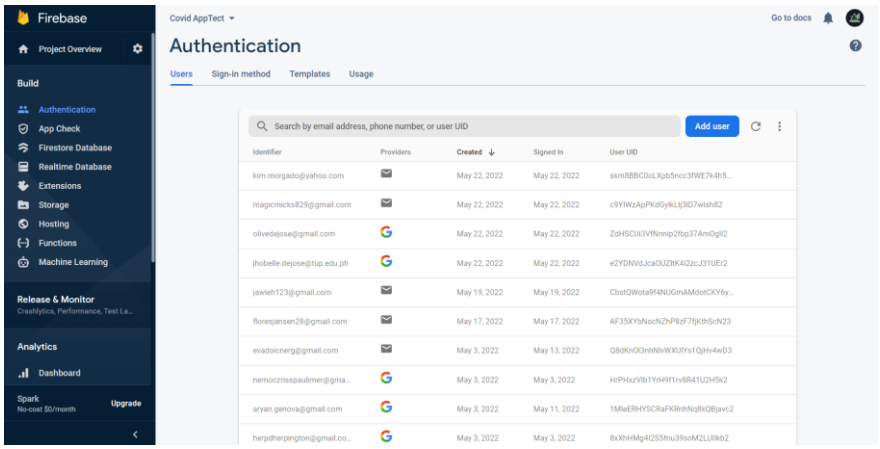


Figure 31: Screenshot of the Login Database

13) *COVID-19 Tracker:* This feature enables the user to view the past results generated using the website application as a graph. They can choose to either view the number of positive cases in the whole NCR, or by city. Take note that all dates used in this tab is in the UTC time zone.

Figure 32: Screenshot of the Statistics Tab

14) *Features Tab:* This provides users with the technical overview of the website application.

Figure 33: Screenshot of the Features Tab

15) *Team Tab:* This provides users with the name and LinkedIn Profile of the people who made this website application possible.

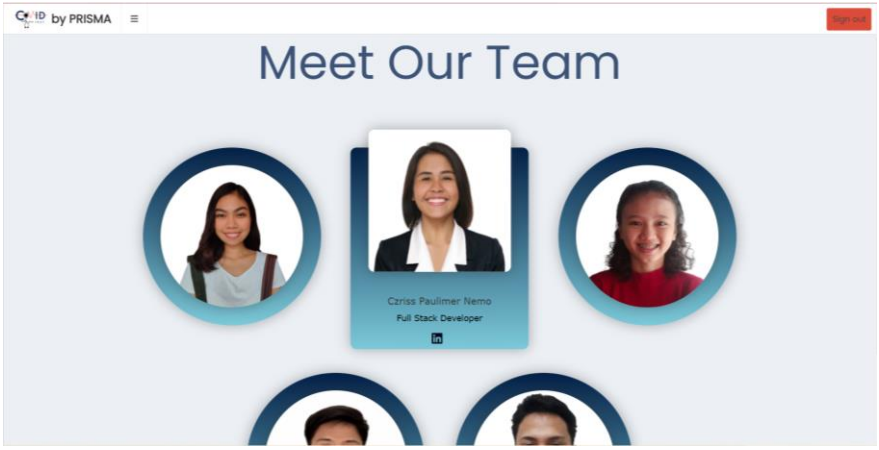


Figure 34: Screenshot of the Teams Tab

16) *Privacy Policy:* Aside from the pop-up box upon login, the website application also encompasses a privacy policy tab wherein the specifics of the data privacy act concerning the website application is discussed in detail.

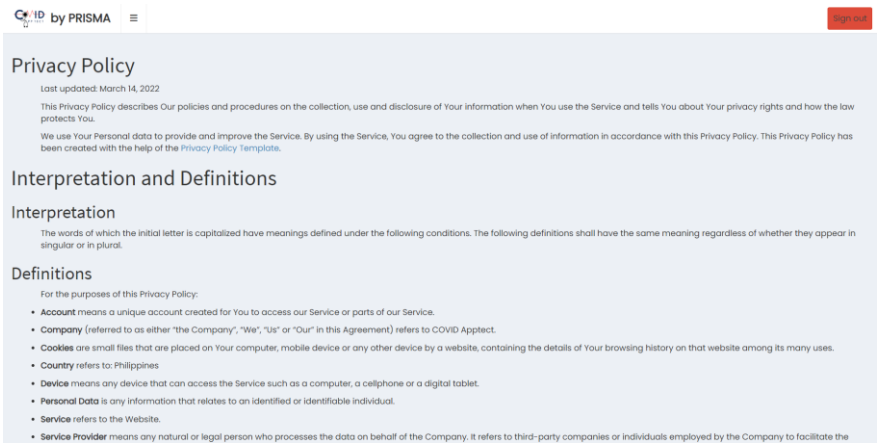


Figure 35: Screenshot of the Privacy Policy Tab

17) *Contact Details and Feedback Form:* This feature displays the contact details where users can directly communicate with the team, as well as a feedback form where all responses are directed to the team's e-mail.

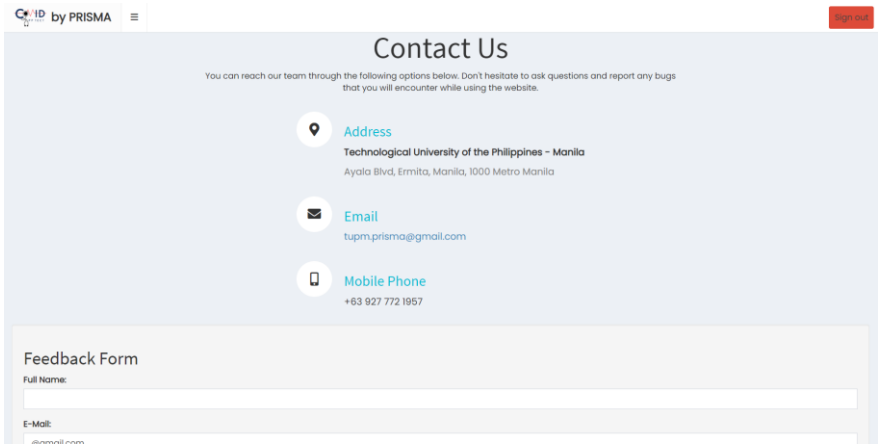


Figure 36: Screenshot of the Contact Us Tab

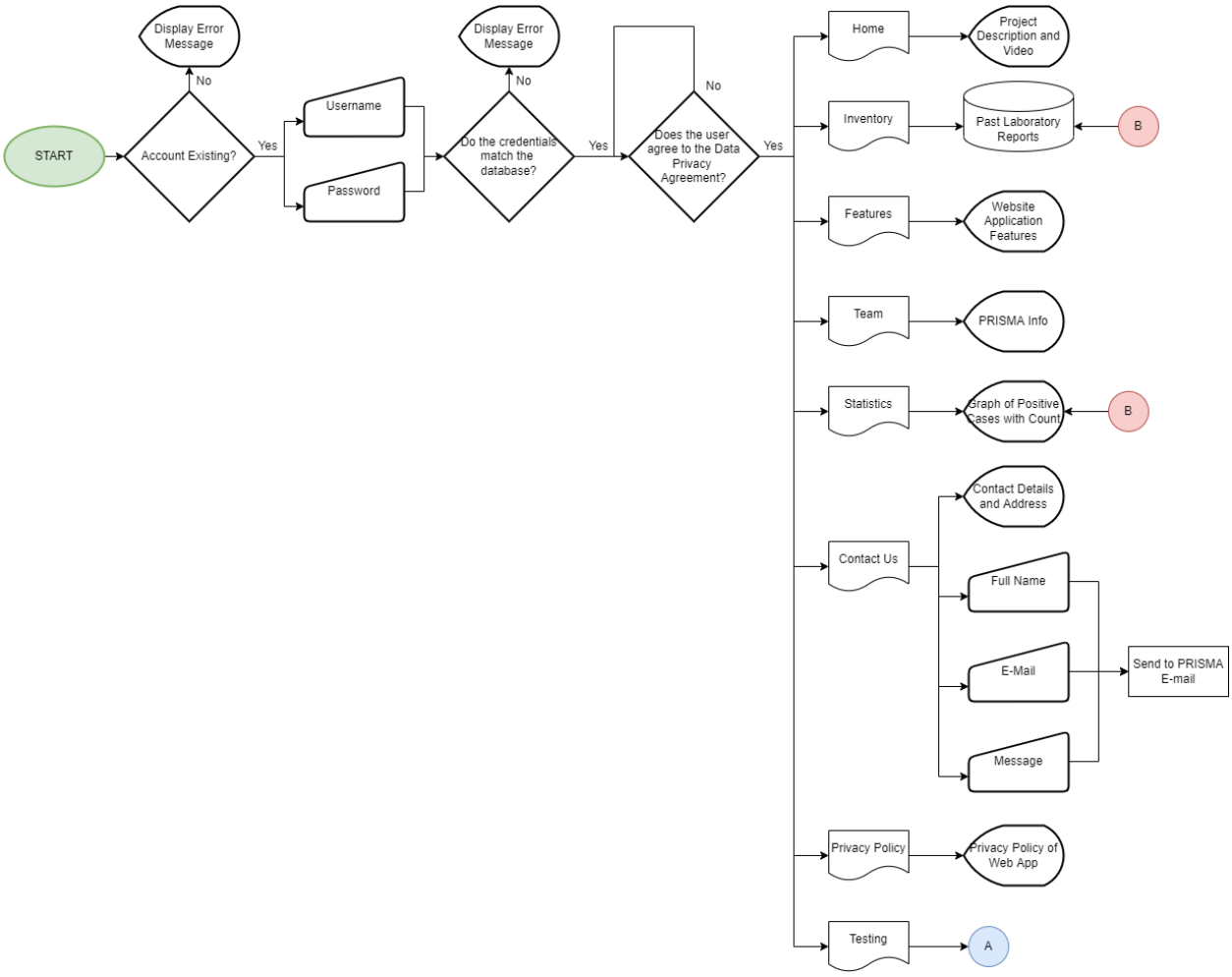


Figure 37: Web Application Flowchart Part 1

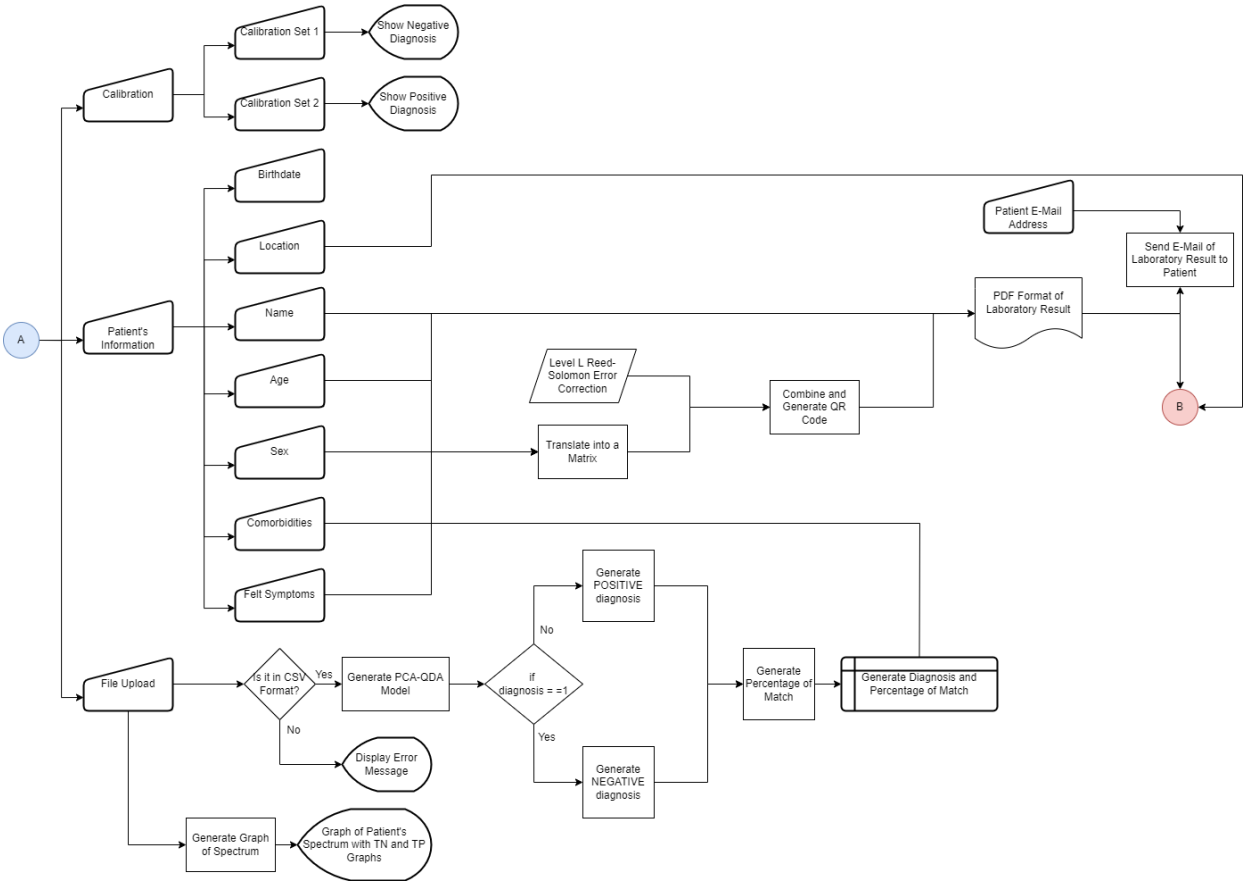


Figure 38: Web Application Flowchart Part 2

3.4.1.3 Program Flowchart

The proceeding figures in this section of the paper will encompass the program flowcharts for the study's proposed website application.

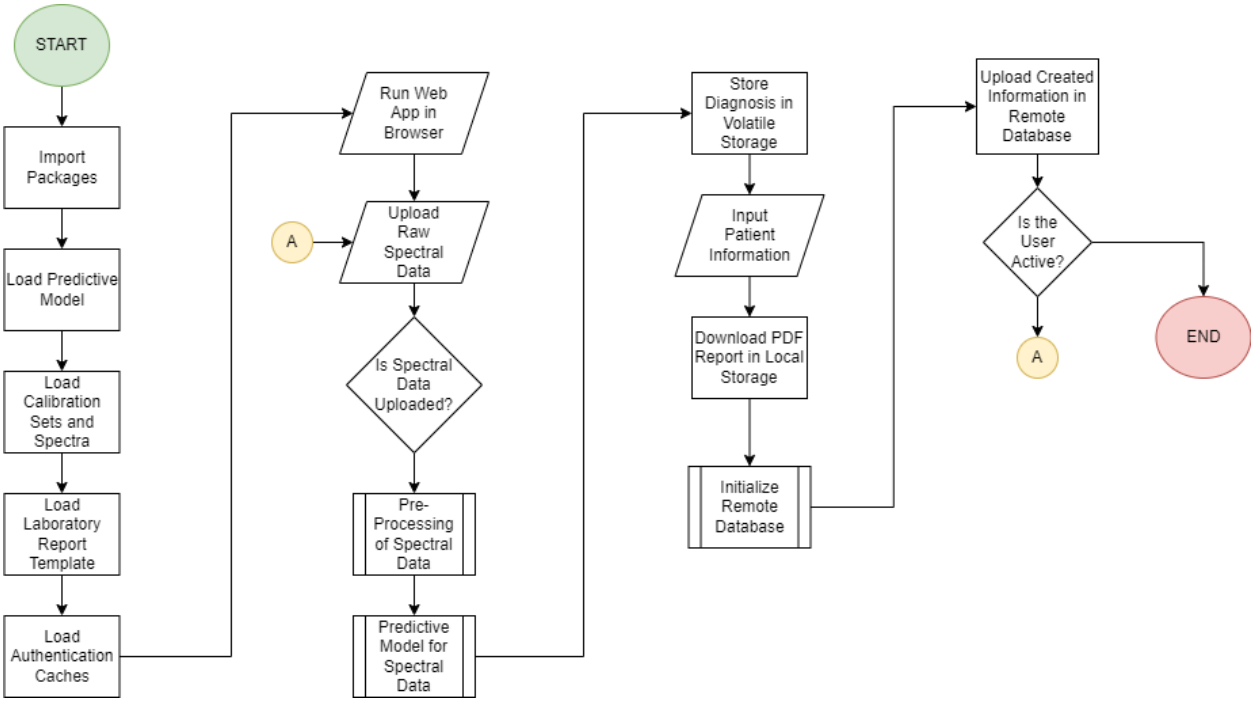


Figure 39: Main Program Flow of the Website Application

Figure 18 shows the general flow of the website application. Upon implementation, the packages necessary for the application will be imported. Necessary data such as the trained predictive model, calibration sets, sample spectra, laboratory template, and the authentication caches will be loaded. The website application can then be accessed by the end user through a browser of their choice. It is also necessary that the raw spectral data must already be available within the local storage of the device. Afterward, the user must upload the raw spectral data into the website application. This will commence the pre-processing

of the spectral data for noise reduction and data simplification purposes. Feeding the pre-processed data into the predictive model comes next. The predictive model will now interpret the processed data and translate the peaks into quantities of biomarkers. The website application will then output these results, alongside the inputted patient information, in a PDF file that is downloadable for personal consumption. These results can be retained upon initializing the remote database through the authentication cache loaded earlier. If the website application does not detect any activity from the user, then the session will be terminated.

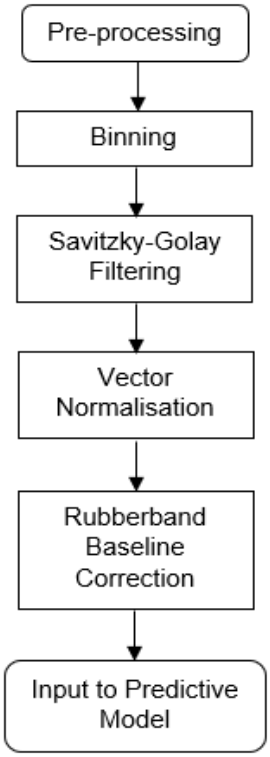


Figure 40: Pre-processing of the Raw Spectral Data (Subroutine)

The figure above shows the flow for the pre-processing of the raw spectral data. Binning will first be employed to help in reducing the points. Afterwards,

Savitzky-Golay filtering will be administered to filter out all noise in the spectrum derived from water. Specifically, the filter will be in the second order polynomial fitting with an average window of nine. After filtering, vector normalization will be done. This will help in reducing the variations that occur within the samples. Rubberband baseline correction will then be used to decrease the scattering present in the spectrum that is brought by radiation from the light source.

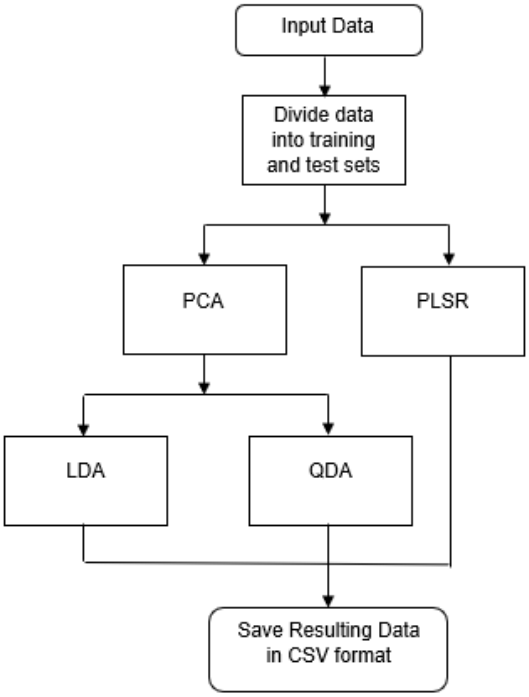


Figure 41: Predictive Models

The predictive models to be used are PCA-LDA, PCA-QDA, and PLS-DA. After the pre-processing, when the spectrum is filtered, the data undergo a multivariate analysis in which classifiers are determined for prediction. This data is first divided into training and test sets. PCA-LDA and PCA-QDA firstly obtain the target variables from the PCA then these variables will be the input variables to

LDA and QDA. PLS-DA is a sole technique used in the model to predict the classifiers.

3.4.1.3 Quality Performance Evaluation

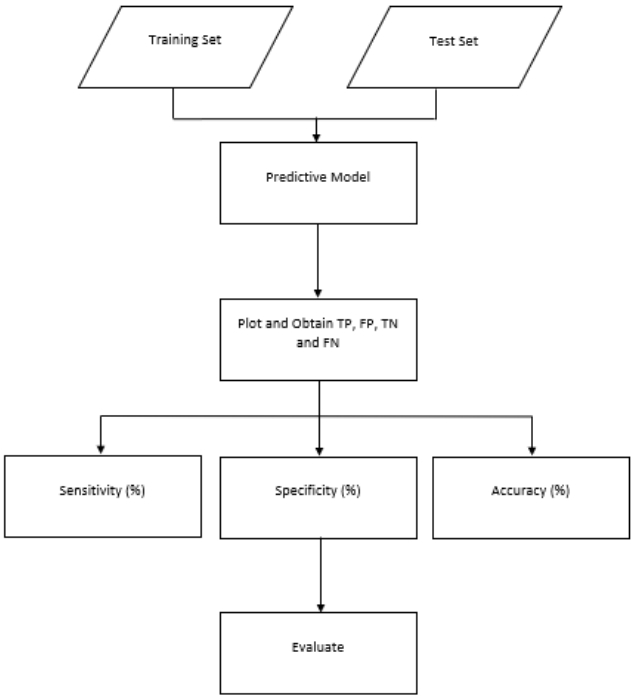


Figure 42: Evaluation of Quality Performance of the model

Table 1: Interpretation of the Area Under the ROC Curve

AUC	Diagnostic Accuracy
0.9 – 1.0	Excellent
0.8 – 0.9	Very good
0.7 – 0.8	Good
0.6 – 0.7	Sufficient
0.5 – 0.6	Bad
< 0.5	Test was not useful

The three parameters in evaluating the quality performance of the models are Sensitivity, Specificity, and Accuracy. Sensitivity is the probability that the model predicts a sample as positive when there is a presence of the virus in the sample. Specificity is the opposite wherein it calculates the probability of a test result that will be negative if there is no presence of the virus. Accuracy is the calculation of the model's ability to correctly classify the samples. Sensitivity, Specificity, and Accuracy follow the equations below:

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

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

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

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

The quality performance of a predictive model is performed both in the training set and test set. These sets will serve as the two measures of classification. The training set is the same samples used in developing and optimizing the model while the test or prediction set is used to test the classification ability of the model. The TP, FP, TN, and FN are all obtained in the program after applying the model to the training and test sets. These values will be needed in calculating the sensitivity, specificity, and accuracy of the model. When all parameters are all calculated in each model, a comparison among them is performed. Results will be documented manually by saving the jpeg format of the plot analyses and results of each parameter will be tabulated. The chemometric model with the best result in its quality performance will be the predictive model to be applied in the web application.

3.5 Testing Procedure

Due to ongoing pandemic and lack of permitted laboratories to experiment viral subjects, the researchers decided to distribute the obtained dataset into testing, training, and evaluation. The evaluation set contains random 10 COVID-19 negative and 10 COVID-19 positive. This set will be used in evaluating the web application. This will also be given to the medical technologists, health practitioners and research specialists who will evaluate further the web application. The figure below provides a visualization of the evaluation procedure of the website application's predictive model.

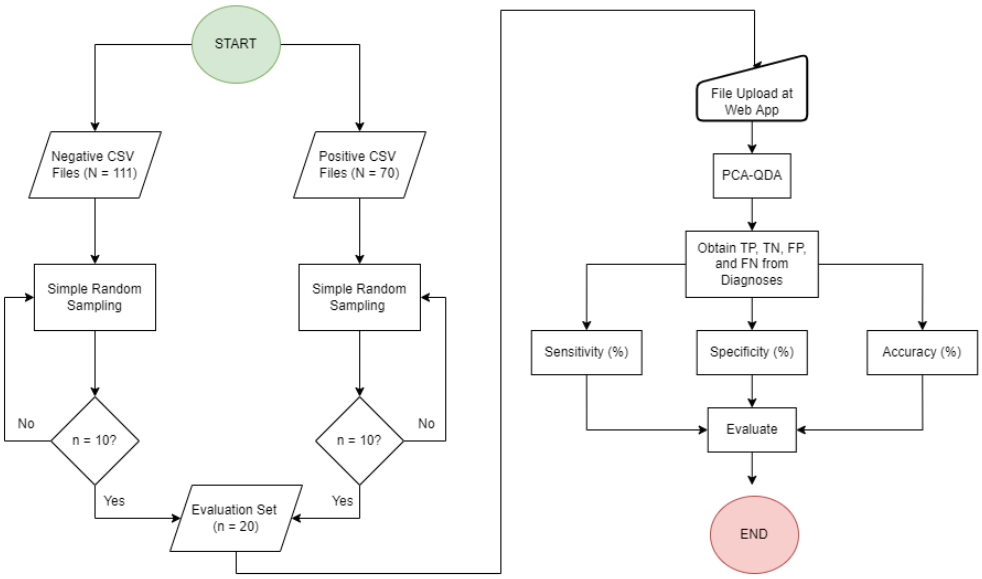


Figure 43: Flowchart of the Evaluation Process of the Predictive Model of the Website Application

The figure above encompasses the process of evaluating the predictive model of the website application. The obtained dataset containing 181 spectra of different diagnoses is separated into two: those whose diagnosis is negative and those that are positive. Afterward, simple random sampling was administered until a sample population of 20 is obtained where 10 is negative and 10 is positive. Now that the optimal number of spectra is obtained for evaluation, each of this spectrum will be inputted one-by-one in the website application. Each of its corresponding diagnosis will be recorded. The resulting diagnoses will be further analyzed by obtaining the total number of True Positives, True Negatives, False Positives, and False Negatives. These numbers will then be used in computing for the specificity, sensitivity, and accuracy which are factors that will be subjected to evaluate the performance of the website application's predictive capability.

3.6 Statistical Treatment

The study uses multivariate statistical analyses using a large dataset generating many dimensions and outputs many variables. These multivariate statistical techniques help the researchers plot the data in a more understandable visual hence the researchers can analyse if the algorithm is performing the expected outcome of the statistical analysis of the dataset.

3.7 Technical Evaluation

Website Application for COVID-19 Patient Diagnosis using FTIR Spectroscopy Proponent Survey					
Introduction: The students involved prior to this study need to conduct a survey for evaluation of different aspects of the proponent, entitled "**SARS-CoV-2 Detection Using FTIR Spectroscopy and Chemometric Analysis Through Saliva Absorbance Levels**".					
Instruction: Please rate whether you strongly disagree or strongly agree. Check one response for the following statements. Rate **1** - if Strongly Disagree **2** - Disagree **3** - Neither Agree nor Disagree **4** - Agree **5** - if Strongly Agree					
Survey Statements	**Rating**				
	1	2	3	4	5
Effectiveness					
1. The website application provided all the information needed in patient diagnosis.					
2. The website application provided information that is easy to understand.					
Efficiency					
3. All information provided by the website application is accurate.					
Satisfaction					
4. The website application is very useful in patient diagnosis for COVID-19.					
5. The website application carried out its tasks as discussed.					
6. The website application provided new information for patient diagnosis for COVID-19.					
7. The overall layout of the website application is very user-friendly and pleasing to the eyes.					
Freedom from Risk					
8. Using the website application helps in reducing the use of resources in the laboratory.					
9. Using the website application lessens the time of exposure to the infected sample.					
Context Coverage					
10. The website application can be accessed in either a laptop or mobile phone.					
11. The website application can be used by someone who does not have much knowledge about infrared spectroscopy and patient diagnosis.					
12. The website application has also other uses aside from patient diagnosis for COVID-19.					

Figure 44: Sample of the Technical Evaluation Form

3.8 Work Plan (Gantt Chart)

	2021											2022							
ACTIVITIES	**FEB**	**MAR**	**APR**	**MAY**	**JUNE**	**JUL**	**AUG**	**SEPT**	**OCT**	**NOV**	**DEC**	**JAN**	**FEB**	**MAR**	**APR**	**MAY**	**JUNE**	**JULY**	**AUG**
Topic Consultation																			
Research on Relevant Studies																			
Formulation and Submission of Chapter 1 to 3																			
Coding for the Pre-processing Portion																			
Coding for the Predictive Models																			
Developing the UI of the Web Application																			
Training and Testing the Predictive Models																			

Comparison of Predictive Models																			
Integration of Database and Other Additional Features																			
Testing and Debugging of Website Application																			
Publication of Website Application																			
Finalization of List of Participants for Website Application Evaluation																			
Project Deployment																			
Formulation of Chapters 4 and 5																			
Finalization of Thesis Paper																			

3.9 Bill of Materials

Table 2: Expenses on Project Deployment

Item Description	Quantity	Unit	Unit Cost	Total Cost
Laptop	1	-	Php 10,600.00	Php 10,600.00
Pocket WiFi	1	-	Php 1,500.00	Php 1,500.00

Table 3: Overall Expenditure of the Project

Item Description	Cost
Laptop	Php 10,600.00
Pocket WiFi	Php 1,500.00
Total	**Php 12,100.00**

CHAPTER 4

Results and Discussion

4.1 Project Technical Description

The project "SARS-CoV 2 Detection Using FTIR Spectroscopy by Comparison of Chemometric Analysis Through Oropharyngeal Swab Samples' Absorbance Levels" is an accessible, novel and faster diagnostic tool for SARS-CoV-2 in which the website application named COVID AppTect, integrated with the capacity for chemometric analysis, is used for the viral detection and spectra analyzation via the Comma-Separated Values (CSV) data format obtained from an oropharyngeal swab sample subjected under a Fourier Transform Infrared Spectrometer. The website application encompasses an authentication system, SARS-CoV 2 diagnosing capabilities, remote data storage, a system for generating PDF documents, and a laboratory report delivery system.

The creation, as well as the comparison, of the predictive models used for chemometric analysis were performed using R while the development of the website application was developed through Shiny, a built-in package of R dedicated for web development. The website application's integrated predictive model is the best performing chemometric method among the three analyses used in the comparison. The three chemometric analyses namely PLS-DA, PCA-LDA, and PCA-QDA, were performed using the open source dataset from the study "Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity" by V. G, Barauna et . al. Evaluation sets were first obtained from dataset prior to splitting it into training and test sets. Afterward, a total of 101 negative for COVID-19 and 60 positive for COVID-19 was split into 80% training and 20% testing.

For the PLS-DA, the datasets first underwent pre-processing which comprises of Binning, Savitzy-Golay filtering, Vector Normalisation, and Rubberband Baseline Correction. On the other hand, Principal Component Analysis (PCA) served as the preprocess for both the PCA-LDA and PCA-QDA. The comparison of these chemometric analyses were based on the quality performances. These were observed using the 'caret' package in R which included the accuracy, sensitivity, and specificity. For further inspection and comparison of the predictive models, their respective AUC curves were evaluated.

The COVID AppTect was developed using Shiny and is hosted by shinyapps.io. The authentication is cookie-based. It employs local persistence wherein users are not automatically logged out even if they close the website application. If the user is new in the website application, then they can create their own credentials to access the site. The app administrators have the authority to either delete the account from the pool of users or restrict their activity in the website application.

For the diagnosing sector of the application, the best performing chemometric model, the PCA-QDA, was implemented. This is only dedicated to the user input's Comma-Separated Values file format of the FTIR spectral data of the patient's oropharyngeal swab. For further analyzation of the spectral data, users can also inspect the input spectrum data in its resulting graphical form alongside the percentage of match that is also an output of the integrated chemometric model.

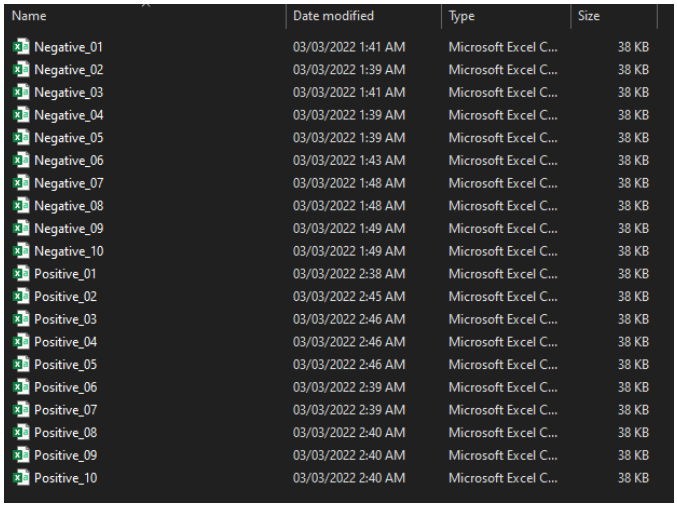
The website application's data storage that houses the users' past inputs are visible in the its inventory, and is also stored in the cloud. To ensure data privacy, users cannot view the patients' sensitive information such as their name. This information is only visible to the app administrator's end which is the cloud-based application that stores the entirety of the database.

For the document preparation system, this provides users a Portable Document File of their generated diagnosis report after they input all the necessary information and prompt the site to start rendering the report. The diagnosis report also includes a QR code which contains the summary of the generated diagnosis report's contents. Lastly, a report delivery system is another feature in which the generated laboratory report in PDF can be sent directly to the patient's electronic mail.

4.2 Project Organizational Structure

4.2.1 CSV File from FTIR

Partnerships with laboratories such as Admatel who have FTIR spectroscopy have been stopped due to a reason that there is no protocol yet implemented to make the viral disease, COVID-19, be a sample to such machinery in the Philippines. Due to this reason, the researchers use the data samples that are provided online through the Centers for Disease Control and Prevention (CDC). The data samples are saved as CSV files since it underwent to FTIR spectroscopy from different country, it is in a form of tabulated data consisting of the wavelengths of the samples. The researchers have separated each patients' OPS' absorbance levels in the evaluation sets and have changed the filenames.



Name	Date modified	Type	Size
Negative_01	03/03/2022 1:41 AM	Microsoft Excel C...	38 KB
Negative_02	03/03/2022 1:39 AM	Microsoft Excel C...	38 KB
Negative_03	03/03/2022 1:41 AM	Microsoft Excel C...	38 KB
Negative_04	03/03/2022 1:39 AM	Microsoft Excel C...	38 KB
Negative_05	03/03/2022 1:39 AM	Microsoft Excel C...	38 KB
Negative_06	03/03/2022 1:43 AM	Microsoft Excel C...	38 KB
Negative_07	03/03/2022 1:48 AM	Microsoft Excel C...	38 KB
Negative_08	03/03/2022 1:48 AM	Microsoft Excel C...	38 KB
Negative_09	03/03/2022 1:49 AM	Microsoft Excel C...	38 KB
Negative_10	03/03/2022 1:49 AM	Microsoft Excel C...	38 KB
Positive_01	03/03/2022 2:38 AM	Microsoft Excel C...	38 KB
Positive_02	03/03/2022 2:45 AM	Microsoft Excel C...	38 KB
Positive_03	03/03/2022 2:46 AM	Microsoft Excel C...	38 KB
Positive_04	03/03/2022 2:46 AM	Microsoft Excel C...	38 KB
Positive_05	03/03/2022 2:46 AM	Microsoft Excel C...	38 KB
Positive_06	03/03/2022 2:39 AM	Microsoft Excel C...	38 KB
Positive_07	03/03/2022 2:39 AM	Microsoft Excel C...	38 KB
Positive_08	03/03/2022 2:40 AM	Microsoft Excel C...	38 KB
Positive_09	03/03/2022 2:40 AM	Microsoft Excel C...	38 KB
Positive_10	03/03/2022 2:40 AM	Microsoft Excel C...	38 KB

Figure 45: CSV Filenames of the Evaluation Set

4.2.2 Devices Used for Opening the Web Application

4.2.2.1 Researchers' Devices

Figure 46: One of the researcher's own tablet used for deploy

Figure 47: One of the researcher's own laptop used for deployment

The web application is a user-friendly interface which can be launched through any devices such as laptop, PC, or tablet. The researchers used their own

tablet and laptops for deployment if the participants prefer to evaluate the web application personally with the researchers or have no available devices.

4.2.2.2 Participant's own devices

For the Participants who evaluated the web application without the researchers around, their own devices will be used. The researchers have provided the participants of the procedures and forms to be answered.

4.3 Experimental Results and Data Analysis

4.3.1 Preprocess Results

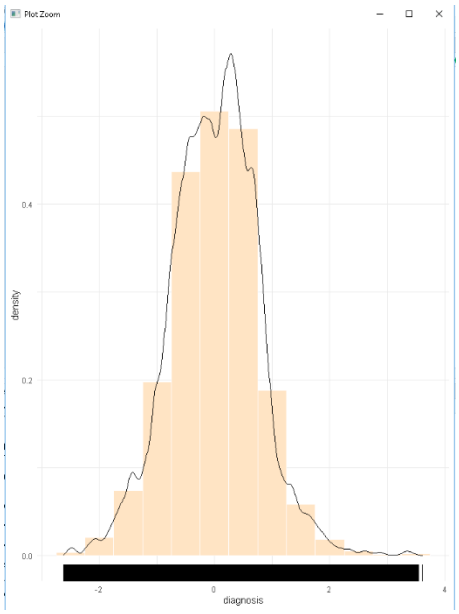


Figure 48: Plot of the obtained dataset

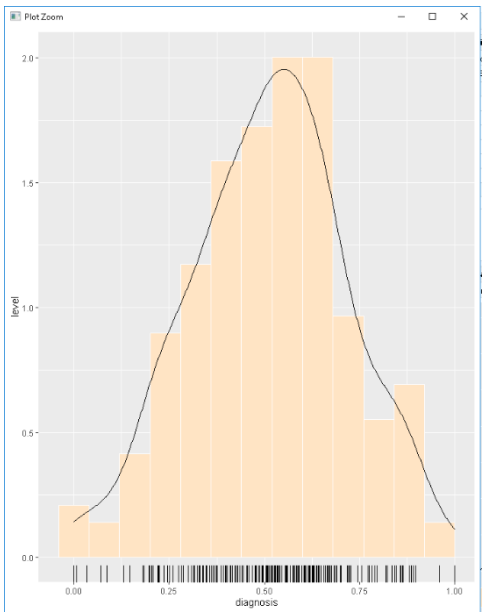


Figure 49: Plot of the obtained dataset after preprocessing

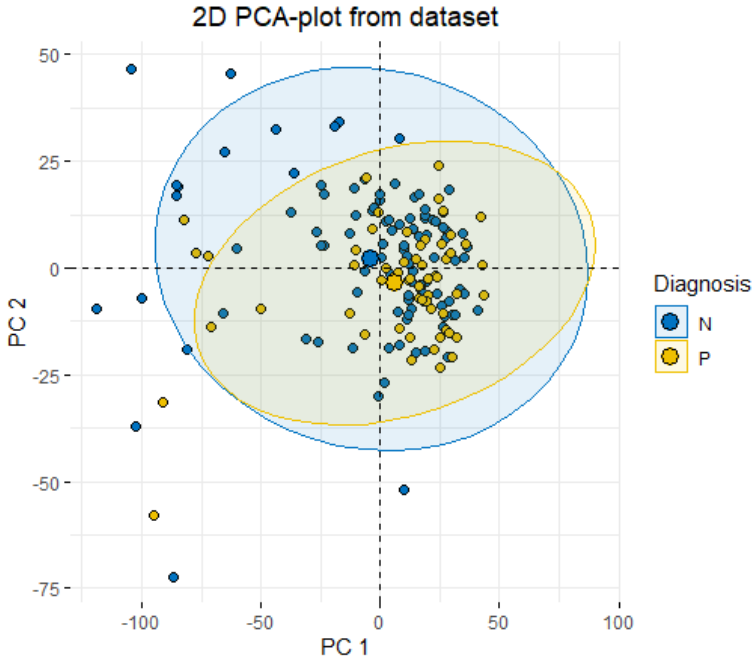


Figure 50: 2D PCA-plot of the obtained dataset with PC2 (15.37% total variance) versus PC1 (70.85%)

The preprocess in PLS-DA included Savitzy-Golay filter, vector normalization, and binning which is shown in Figure7. Inherent noise were reduced and peak morphology was maintained after it was preprocessed. For the classification of the dataset, PCA technique is used. Upon observing Figure 8 which depicts the scores of the 1st and 2nd PCs, the discrimination pattern is lacking. Moreover, this suggests to further classify using the supervised discriminant analyses which are the LDA and QDA.

4.3.2 Experimental Parametric Procedures Results

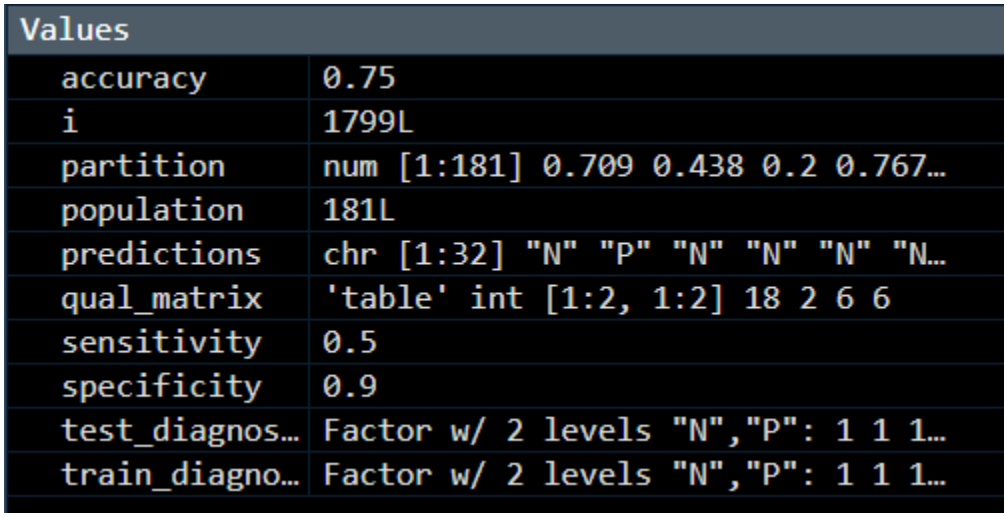
4.3.2.1 Results in PLS DA

Values	
accuracy	0.769230769230769
i	1799L
predictions	chr [1:52] "N" "N" "N" "N" "P" "N" "N" "N" "N"…
qual_matrix	'table' int [1:2, 1:2] 30 1 11 10
seg_number	num [1:181] 0.185 0.702 0.573 0.168 0.944 ...
sensitivity	0.476190476190476
specificity	0.967741935483871

Figure 51: Quality Performance of PLS-DA After Applying 80% Threshold for Vector Normalization

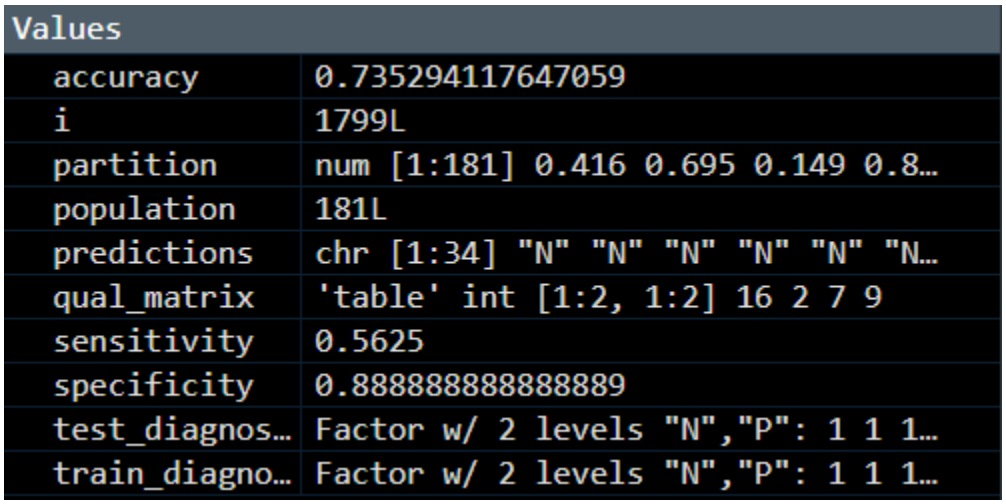
Values	
accuracy	0.75
i	1799L
partition	num [1:181] 0.185 0.702 0.573 0.168 0.944 ...
population	181L
predictions	chr [1:52] "N" "N" "N" "N" "P" "N" "P" "P" "N" "N" "N" "N…
qual_matrix	'table' int [1:2, 1:2] 24 7 6 15
sensitivity	0.714285714285714
specificity	0.774193548387097
test_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1 1 ...
train_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1 1 ...

Figure 52: Quality Performance of PLS-DA After Applying Savitky-Golay Filter and set.seed



Values	
accuracy	0.75
i	1799L
partition	num [1:181] 0.709 0.438 0.2 0.767…
population	181L
predictions	chr [1:32] "N" "P" "N" "N" "N" "N…
qual_matrix	'table' int [1:2, 1:2] 18 2 6 6
sensitivity	0.5
specificity	0.9
test_diagnos…	Factor w/ 2 levels "N","P": 1 1 1…
train_diagno…	Factor w/ 2 levels "N","P": 1 1 1…

Figure 53: Quality Performance of PLS-DA After Applying Savitsky-Golay Filter and set.seed 50



Values	
accuracy	0.735294117647059
i	1799L
partition	num [1:181] 0.416 0.695 0.149 0.8…
population	181L
predictions	chr [1:34] "N" "N" "N" "N" "N" "N…
qual_matrix	'table' int [1:2, 1:2] 16 2 7 9
sensitivity	0.5625
specificity	0.888888888888889
test_diagnos…	Factor w/ 2 levels "N","P": 1 1 1…
train_diagno…	Factor w/ 2 levels "N","P": 1 1 1…

Figure 54. Quality Performance of PLS-DA After Applying Savitsky-Golay Filter and set.seed 25

Values	
accuracy	0.619047619047619
i	1799L
partition	num [1:181] 0.266 0.372 0.573 0.9…
population	181L
predictions	chr [1:42] "N" "N" "P" "P" "P" "N…
qual_matrix	'table' int [1:2, 1:2] 20 10 6 6
sensitivity	0.5
specificity	0.666666666666667
test_diagnos…	Factor w/ 2 levels "N","P": 1 1 1…
train_diagno…	Factor w/ 2 levels "N","P": 1 1 1…

Figure 55: Quality Performance of PLS-DA After Applying Savitsky-Golay Filter and set.seed 1

Values	
accuracy	0.790697674418605
conft_matrix	'table' int [1:2, 1:2] 17 1 9 …
diagnosis	num [1:161] 1 1 1 1 1 1 1 1 1 …
i	1799L
N	161L
partition	num [1:161] 0.185 0.702 0.573 …
plsda_train…	Factor w/ 2 levels "N","P": 1 …
population	161L
pr.var	num [1:181] 1259.8 286.3 139.4…
predictions	chr [1:43] "N" "N" "N" "P" "N"…
qual_matrix	'table' int [1:2, 1:2] 22 4 5 …
rvec	num [1:181] 0.5075 0.3068 0.42…
rvec1	num [1:161] 0.8229 0.7102 0.96…
rvec2	num [1:161] 0.446 0.395 0.484 …
sarscov.qda…	Factor w/ 2 levels "0","1": 1 …
sensitivity	0.705882352941177
specificity	0.846153846153846

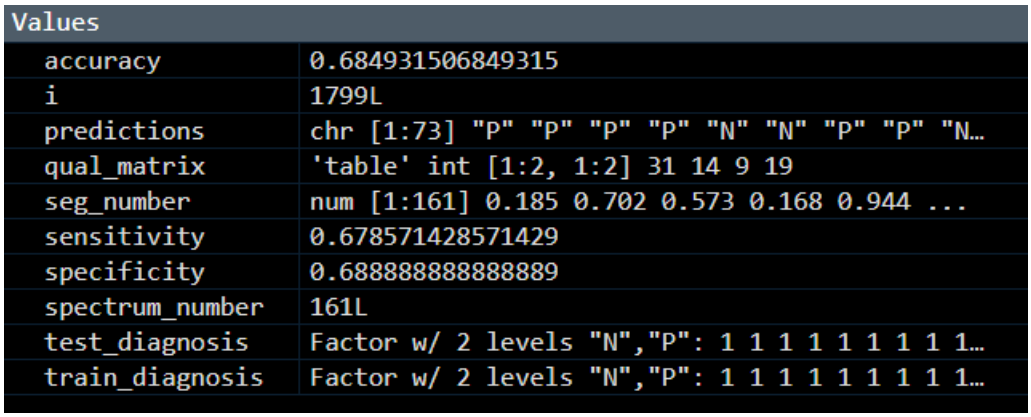
Figure 56: Quality Performance of PLS-DA After Applying All Pre-Process Methods at set.seed 2

Values	
accuracy	0.526315789473684
partition	num [1:161] 0.185 0.702 0.573 …
population	161L
predictions	chr [1:38] "P" "P" "P" "N" "P"…
qual_matrix	'table' int [1:2, 1:2] 11 12 6…
sensitivity	0.6
specificity	0.478260869565217
test_diagno…	Factor w/ 2 levels "N","P": 1 …
train_diagn…	Factor w/ 2 levels "N","P": 1 …

Figure 57: Quality Performance of PLS-DA with no Pre-Process Methods at set.seed 2

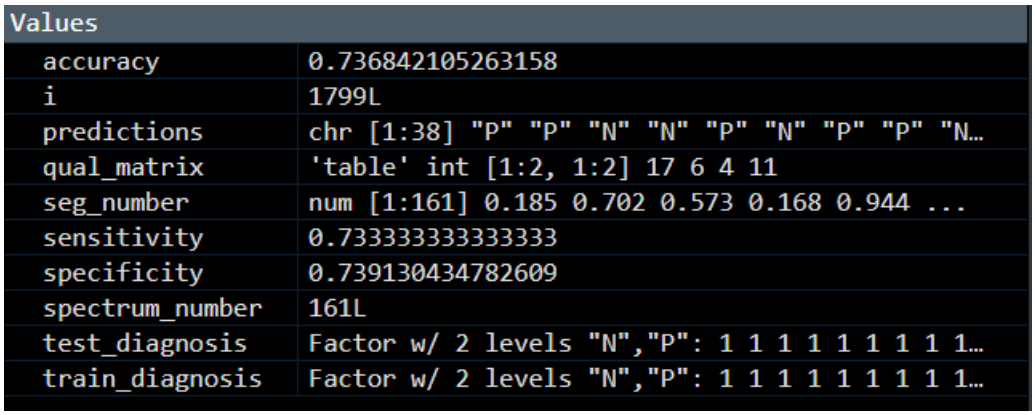
Values	
accuracy	0.790697674418605
conft_matrix	'table' int [1:2, 1:2] 17 1 9 …
diagnosis	num [1:161] 1 1 1 1 1 1 1 1 1 …
i	1799L
N	161L
partition	num [1:161] 0.185 0.702 0.573 …
plsda_train…	Factor w/ 2 levels "N","P": 1 …
population	161L
pr.var	num [1:181] 1259.8 286.3 139.4…
predictions	chr [1:43] "N" "N" "N" "P" "N"…
qual_matrix	'table' int [1:2, 1:2] 22 4 5 …
rvec	num [1:181] 0.5075 0.3068 0.42…
rvec1	num [1:161] 0.8229 0.7102 0.96…
rvec2	num [1:161] 0.446 0.395 0.484 …
sarscov.qda…	Factor w/ 2 levels "0","1": 1 …
sensitivity	0.705882352941177
specificity	0.846153846153846

Figure 58: Quality Performance of PLS-DA using 75%-Train and 25%-Test Data Partition



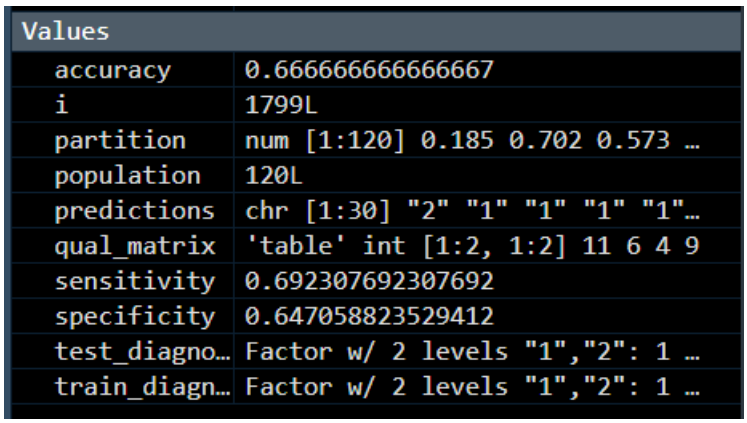
Values	
accuracy	0.684931506849315
i	1799L
predictions	chr [1:73] "P" "P" "P" "P" "N" "N" "P" "P" "N…
qual_matrix	'table' int [1:2, 1:2] 31 14 9 19
seg_number	num [1:161] 0.185 0.702 0.573 0.168 0.944 ...
sensitivity	0.678571428571429
specificity	0.688888888888889
spectrum_number	161L
test_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…
train_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…

Figure 59: Quality Performance of PLS-DA using 50%-Train and 50%-Test Data Partition



Values	
accuracy	0.736842105263158
i	1799L
predictions	chr [1:38] "P" "P" "N" "N" "P" "N" "P" "P" "N…
qual_matrix	'table' int [1:2, 1:2] 17 6 4 11
seg_number	num [1:161] 0.185 0.702 0.573 0.168 0.944 ...
sensitivity	0.733333333333333
specificity	0.739130434782609
spectrum_number	161L
test_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…
train_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…

Figure 60: Quality Performance of PLS-DA using 80%-Train and 20%-Test Data Partition



Values	
accuracy	0.666666666666667
i	1799L
partition	num [1:120] 0.185 0.702 0.573 …
population	120L
predictions	chr [1:30] "2" "1" "1" "1" "1"…
qual_matrix	'table' int [1:2, 1:2] 11 6 4 9
sensitivity	0.692307692307692
specificity	0.647058823529412
test_diagno…	Factor w/ 2 levels "1","2": 1 …
train_diagn…	Factor w/ 2 levels "1","2": 1 …

Figure 61: Quality Performance of PLS-DA using 60 Positive and 60 Negative Datasets at 80%-Train and 20%-Test Data Partition

Values	
accuracy	0.565217391304348
i	1799L
predictions	chr [1:23] "N" "N" "P" "P" "N" "P" "N" "P" "P…
qual_matrix	'table' int [1:2, 1:2] 6 8 2 7
seg_number	num [1:100] 0.185 0.702 0.573 0.168 0.944 ...
sensitivity	0.777777777777778
specificity	0.428571428571429
spectrum_number	100L
test_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…
train_diagnosis	Factor w/ 2 levels "N","P": 1 1 1 1 1 1 1 1 1…

Figure 62: Quality Performance of PLS-DA using 50 Positive and 50 Negative Datasets at 80%-Train and 20%-Test Data Partition

4.3.2.2 Results in PCA LDA

```
Confusion Matrix and Statistics


sarscov.lda.predict.class  0  1
                        0  7  5
                        1  1 17

               Accuracy : 0.8
                 95% CI : (0.6143, 0.9229)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.2751

                  Kappa : 0.5588

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.8750
            Specificity : 0.7727
         Pos Pred Value : 0.5833
         Neg Pred Value : 0.9444
             Prevalence : 0.2667
         Detection Rate : 0.2333
   Detection Prevalence : 0.4000
      Balanced Accuracy : 0.8239

       'Positive' Class : 0
```

Figure 63: Quality Performance of PCA LDA using 75%-Train and 25%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.lda.predict.class  0  1
                        0 15  7
                        1 11 42

               Accuracy : 0.76
                 95% CI : (0.6475, 0.8511)
    No Information Rate : 0.6533
    P-Value [Acc > NIR] : 0.03172

                  Kappa : 0.4503

 Mcnemar's Test P-Value : 0.47950

            Sensitivity : 0.5769
            Specificity : 0.8571
         Pos Pred Value : 0.6818
         Neg Pred Value : 0.7925
             Prevalence : 0.3467
         Detection Rate : 0.2000
   Detection Prevalence : 0.2933
      Balanced Accuracy : 0.7170

       'Positive' Class : 0
```

Figure 64: Quality Performance of PCA LDA using 50%-Train and 50%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.lda.predict.class  0  1
                        0  7  4
                        1  1 12

               Accuracy : 0.7917
                 95% CI : (0.5785, 0.9287)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.1383

                  Kappa : 0.5714

 Mcnemar's Test P-Value : 0.3711

            Sensitivity : 0.8750
            Specificity : 0.7500
         Pos Pred Value : 0.6364
         Neg Pred Value : 0.9231
             Prevalence : 0.3333
         Detection Rate : 0.2917
   Detection Prevalence : 0.4583
      Balanced Accuracy : 0.8125

       'Positive' Class : 0
```

Figure 65: Quality Performance of PCA LDA using 80%-Train and 20%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.lda.predict.class 0 1
                        0 6 4
                        1 5 5

               Accuracy : 0.55
                 95% CI : (0.3153, 0.7694)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 0.5914

                  Kappa : 0.1

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.5455
            Specificity : 0.5556
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.5000
             Prevalence : 0.5500
         Detection Rate : 0.3000
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.5505

       'Positive' Class : 0
```

Figure 66: Quality Performance of PCA LDA using 60 Positive and 60 Negative Datasets at 80%-Train and 20%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.lda.predict.class 0 1
                        0 5 2
                        1 2 7

               Accuracy : 0.75
                 95% CI : (0.4762, 0.9273)
    No Information Rate : 0.5625
    P-Value [Acc > NIR] : 0.102

                  Kappa : 0.4921

 Mcnemar's Test P-Value : 1.000

            Sensitivity : 0.7143
            Specificity : 0.7778
         Pos Pred Value : 0.7143
         Neg Pred Value : 0.7778
             Prevalence : 0.4375
         Detection Rate : 0.3125
   Detection Prevalence : 0.4375
      Balanced Accuracy : 0.7460

       'Positive' Class : 0
```

Figure 67: Quality Performance of PCA LDA using 50 Positive and 50 Negative Datasets at 80%-Train and 20%-Test Data Partition

4.3.2.3 Results in PCA QDA

```
Confusion Matrix and Statistics


sarscov.qda.predict.class  0  1
                        0 14  8
                        1  3 21

               Accuracy : 0.7609
                 95% CI : (0.6123, 0.8741)
    No Information Rate : 0.6304
    P-Value [Acc > NIR] : 0.04338

                  Kappa : 0.5163

 Mcnemar's Test P-Value : 0.22780

            Sensitivity : 0.8235
            Specificity : 0.7241
         Pos Pred Value : 0.6364
         Neg Pred Value : 0.8750
             Prevalence : 0.3696
         Detection Rate : 0.3043
   Detection Prevalence : 0.4783
      Balanced Accuracy : 0.7738

       'Positive' Class : 0
```

Figure 68: Quality Performance of PCA QDA using 75%-Train and 25%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.qda.predict.class  0  1
                        0 21 11
                        1 11 37

               Accuracy : 0.725
                 95% CI : (0.6138, 0.819)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.01365

                  Kappa : 0.4271

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.6562
            Specificity : 0.7708
         Pos Pred Value : 0.6562
         Neg Pred Value : 0.7708
             Prevalence : 0.4000
         Detection Rate : 0.2625
   Detection Prevalence : 0.4000
      Balanced Accuracy : 0.7135

       'Positive' Class : 0
```

Figure 69: Quality Performance of PCA QDA using 50%-Train and 50%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.qda.predict.class  0  1
                        0 13  4
                        1  3 16

               Accuracy : 0.8056
                 95% CI : (0.6398, 0.9181)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.001559

                  Kappa : 0.6087

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.8125
            Specificity : 0.8000
         Pos Pred Value : 0.7647
         Neg Pred Value : 0.8421
             Prevalence : 0.4444
         Detection Rate : 0.3611
   Detection Prevalence : 0.4722
      Balanced Accuracy : 0.8063

       'Positive' Class : 0
```

Figure 70: Quality Performance of PCA QDA using 80%-Train and 20%-Test Data Partition

```
Confusion Matrix and Statistics


sarscov.qda.predict.class  0  1
                        0 10  6
                        1  3  4

               Accuracy : 0.6087
                 95% CI : (0.3854, 0.8029)
    No Information Rate : 0.5652
    P-Value [Acc > NIR] : 0.4205

                  Kappa : 0.1753

 Mcnemar's Test P-Value : 0.5050

            Sensitivity : 0.7692
            Specificity : 0.4000
         Pos Pred Value : 0.6250
         Neg Pred Value : 0.5714
             Prevalence : 0.5652
         Detection Rate : 0.4348
   Detection Prevalence : 0.6957
      Balanced Accuracy : 0.5846

       'Positive' Class : 0
```

Figure 71: Quality Performance of PCA QDA using 60 Positive and 60 Negative Datasets at 80%-Train and 20%-Test Data Partition

```
Confusion Matrix and Statistics

sarscov.qda.predict.class 0 1
                        0 9 4
                        1 2 5

               Accuracy : 0.7
                 95% CI : (0.4572, 0.8811)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 0.1299

                  Kappa : 0.3814

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.8182
            Specificity : 0.5556
         Pos Pred Value : 0.6923
         Neg Pred Value : 0.7143
             Prevalence : 0.5500
         Detection Rate : 0.4500
   Detection Prevalence : 0.6500
      Balanced Accuracy : 0.6869

       'Positive' Class : 0
```

Figure 71: Quality Performance of PCA QDA using 50 Positive and 50 Negative Datasets at 80%-Train and 20%-Test Data Partition

4.3.2.4 AUC Results

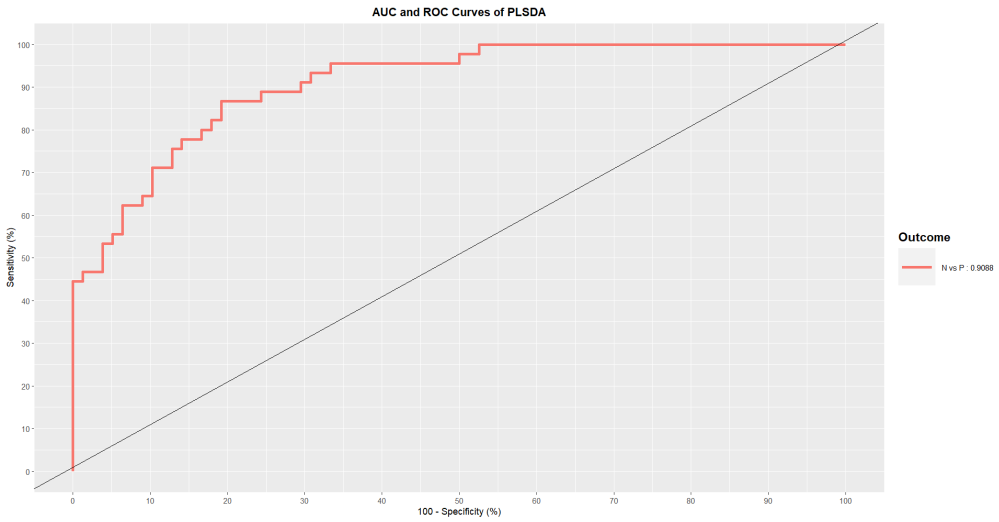


Figure 72: AUC of PLS-DA using 80%-Train and 20%-Test Data Partition

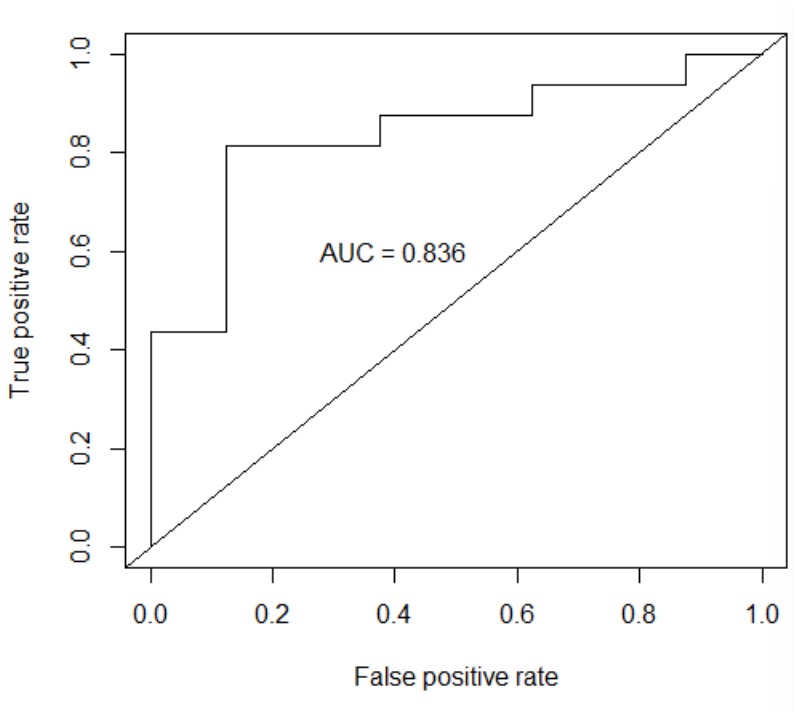


Figure 73: AUC of PCA LDA using 80%-Train and 20%-Test Data Partition

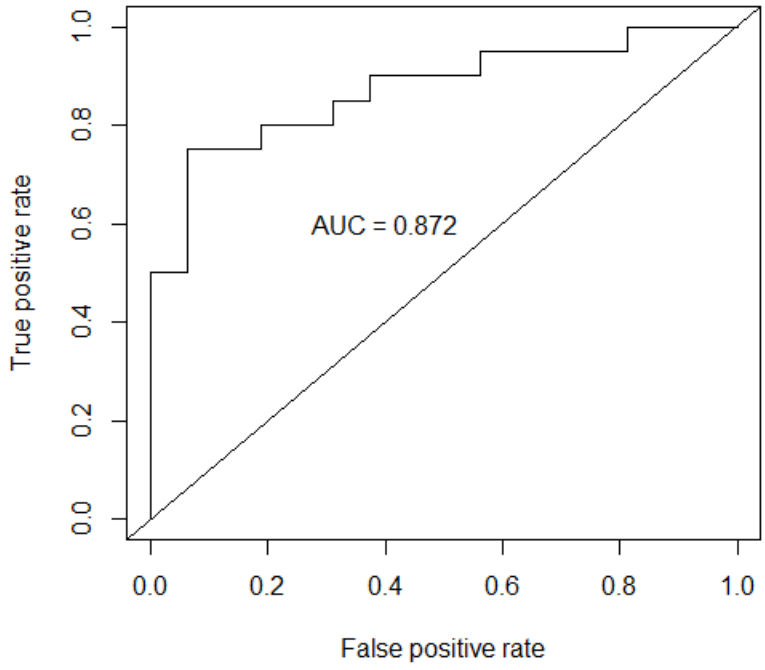


Figure 74: AUC of PCA QDA using 80%-Train and 20%-Test Data Partition

The highest accuracy garnered in experimental parametric procedures was the 80.65% using 80-20% partition of datasets. It can be observed that PLS DA performed 73.68% accuracy, 73.33% sensitivity, and 73.91% specificity. PCA LDA showed 79.17% accuracy, 87.5% sensitivity, and 75% specificity. And PCA QDA resulted to 80.65% accuracy, 81.25% sensitivity, and 80% specificity. This parameter is selected thus the researchers also obtain the AUC for the three prediction models. In addition to this, PLS DA's AUC shows 0.909 which is the highest, PCA LDA is 0.836, and PCA QDA is 0.872. PCA QDA showed the highest accuracy. For sensitivity, PLS DA performed best resulting to 84.62% and PCA LDA has the highest sensitivity with 87.5%.

4.3.3 Spectral Graphs of the dataset

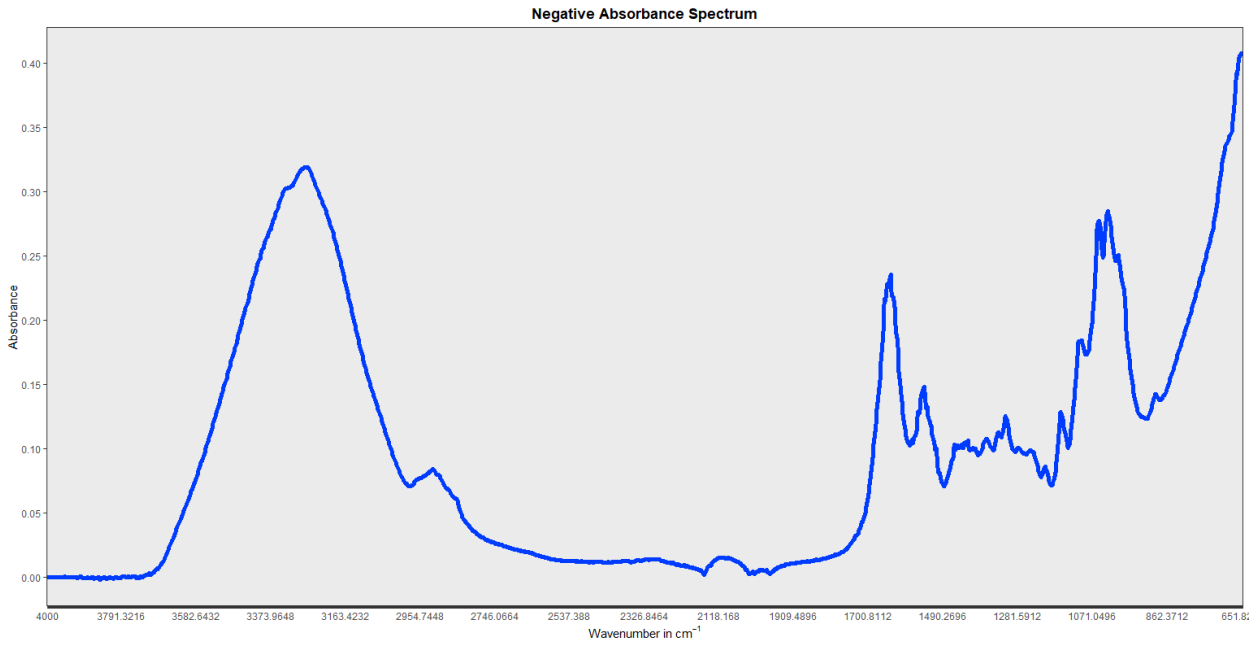


Figure 75: Spectral Graph of the obtained dataset of COVID-19 Negative

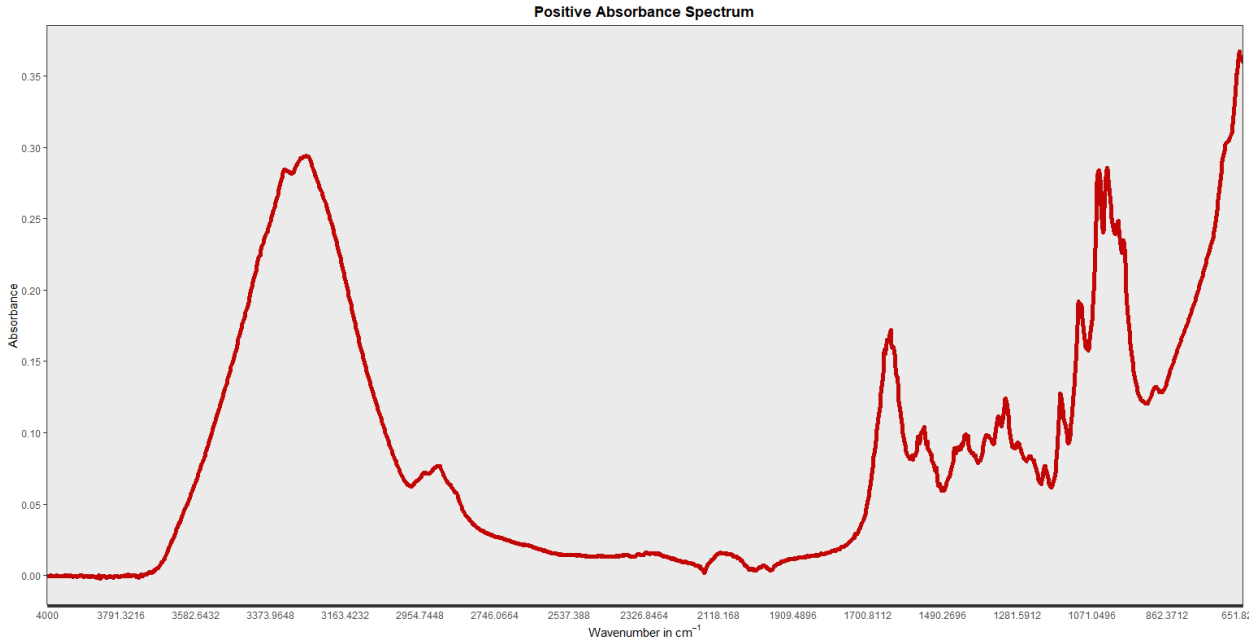


Figure 76: Spectral Graph of the obtained dataset of COVID-19 Positive

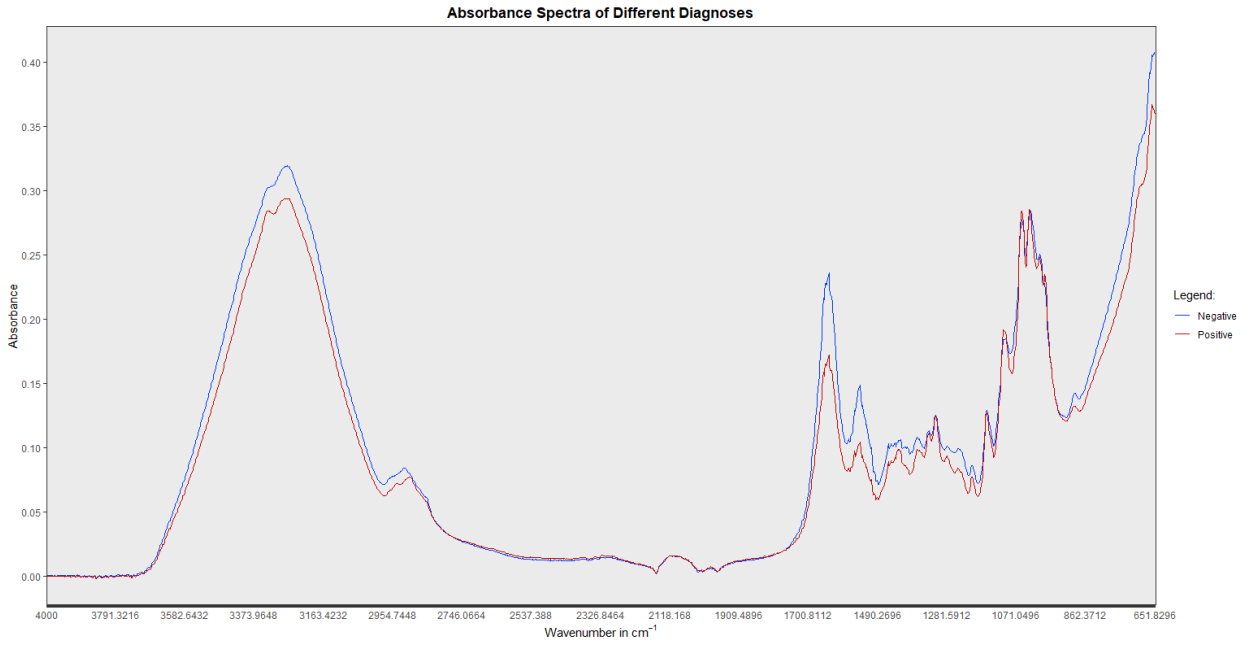


Figure 77: Spectral Graph of the obtained dataset of COVID-19 Negative and Positive

4.3.4 Diagnosis Analysis Results

Filename	RT PCR Diagnosis	COVID AppTect Diagnosis (PCA-QDA)	Matching Percentage
Negative_01	Negative	Negative	51.0111
Negative _02	Negative	Positive	63.7042
Negative _03	Negative	Negative	86.0509
Negative _04	Negative	Negative	55.7347
Negative _05	Negative	Negative	87.0975
Negative _06	Negative	Negative	71.9492
Negative _07	Negative	Negative	50.9433
Negative _08	Negative	Negative	89.1132
Negative _09	Negative	Negative	99.7351
Negative _10	Negative	Negative	98.3148
Positive_01	Positive	Positive	98.276
Positive_02	Positive	Negative	98.3803
Positive_03	Positive	Positive	93.7173
Positive_04	Positive	Positive	59.9311
Positive_05	Positive	Negative	64.7082
Positive_06	Positive	Positive	99.942
Positive_07	Positive	Positive	97.1096
Positive_08	Positive	Positive	96.7688
Positive_09	Positive	Positive	99.5428
Positive_10	Positive	Positive	94.3943

Table 4: COVID AppTect's Diagnosis of the Evaluation Set

Table 4 displays the resulting diagnoses of the COVID AppTect of the evaluation set and their respective matching percentage to the prediction model. Among the 10 diagnosed negative from COVID-19 via RT-PCR, 9 was diagnosed by the web application correctly while 8 out of 10 in COVID-19. As for the

matching percentage, it is observed that all of them obtained 50% above showing considerable match against the model prediction.

4.3.4.1 Graphs of evaluation sets using the PCA QDA prediction model

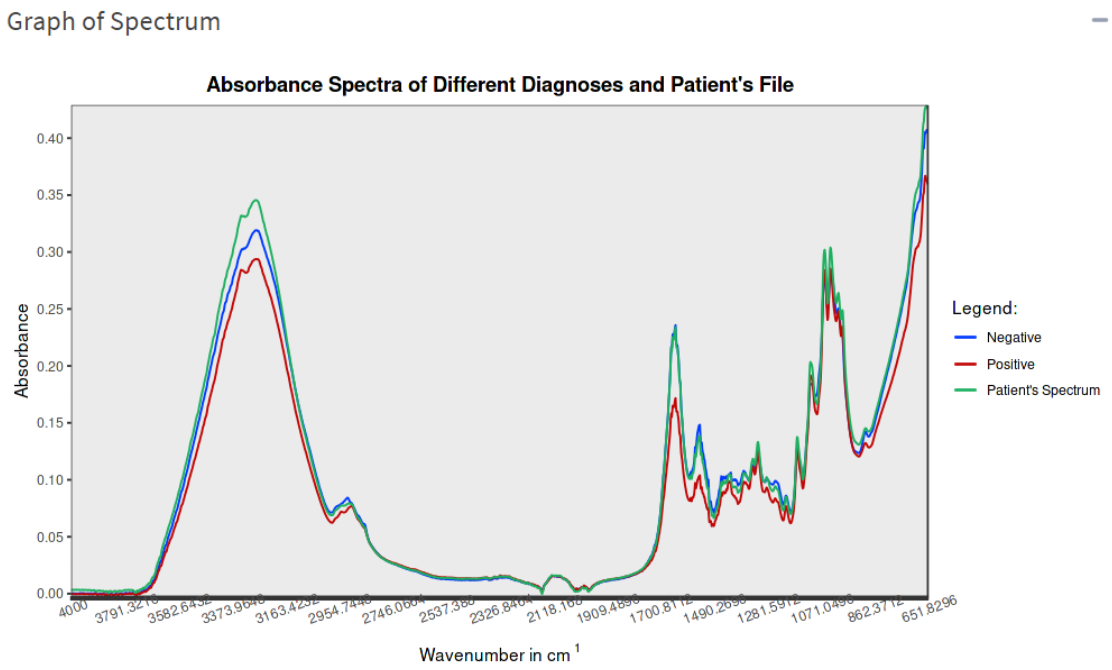


Figure 78: Graph Spectrum the file Negative_01 with color green as the legend

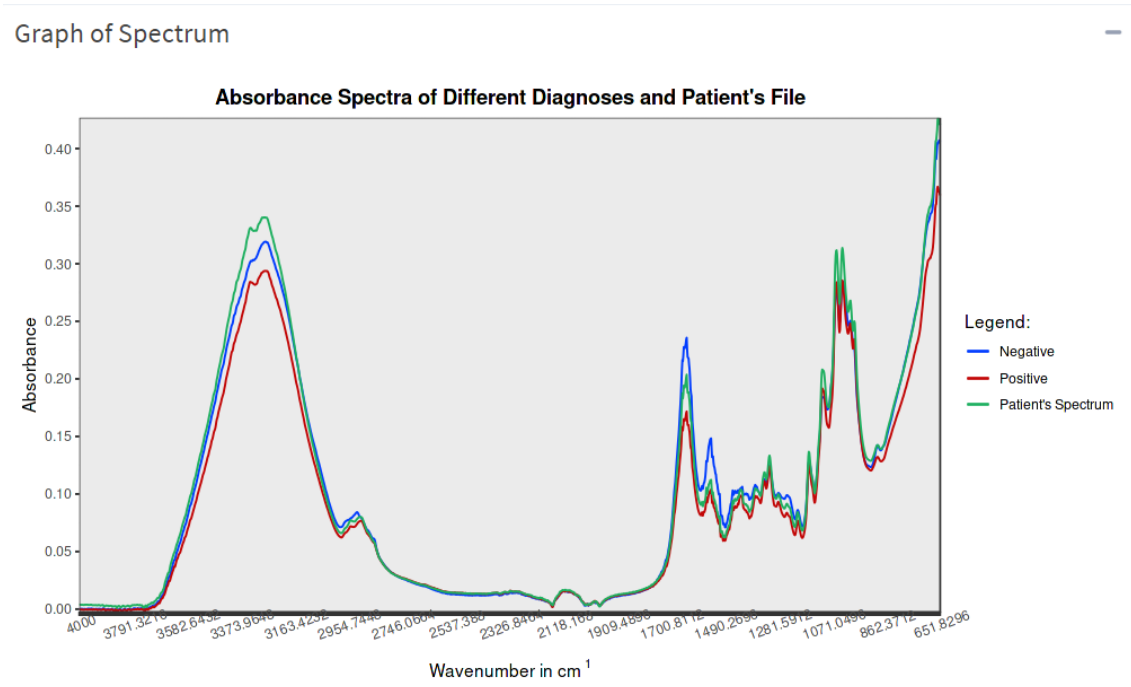


Figure 79: Graph Spectrum the file Negative_02 with color green as the legend

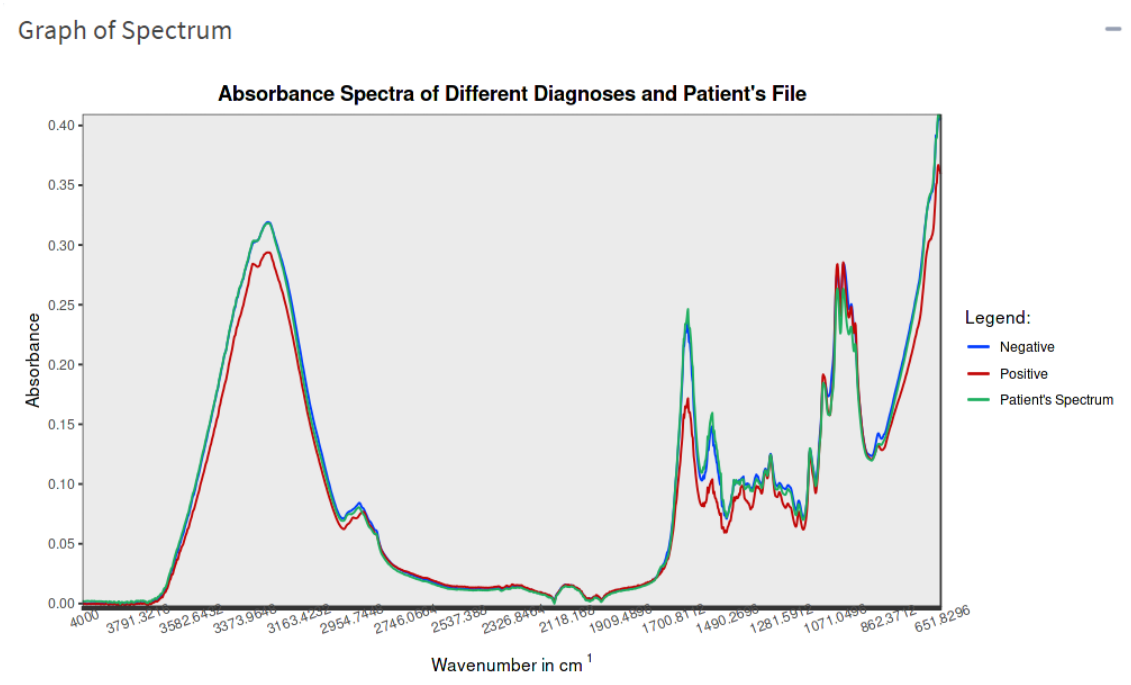


Figure 80: Graph Spectrum the file Negative_03 with color green as the legend

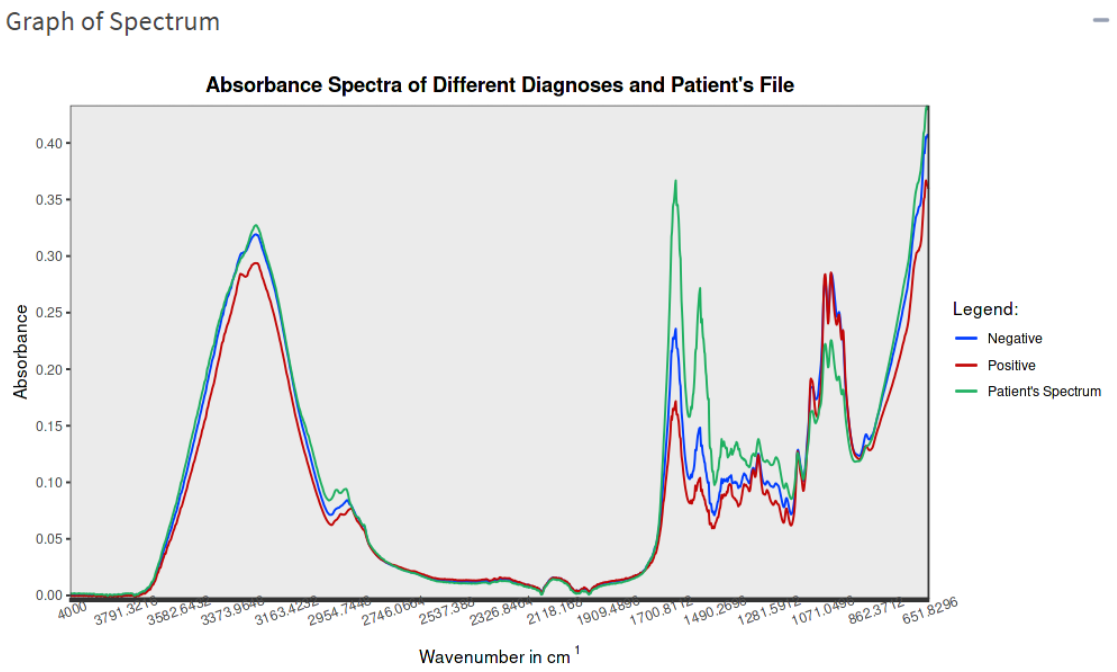


Figure 81: Graph Spectrum the file Negative_04 with color green as the legend

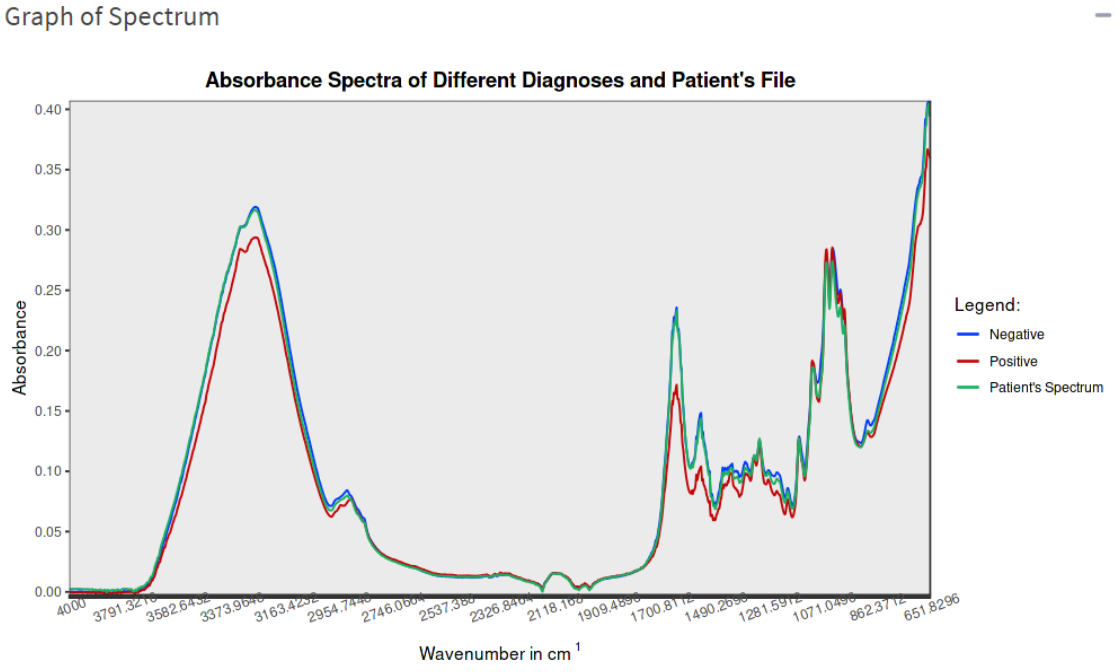


Figure 82: Graph Spectrum the file Negative_05 with color green as the legend

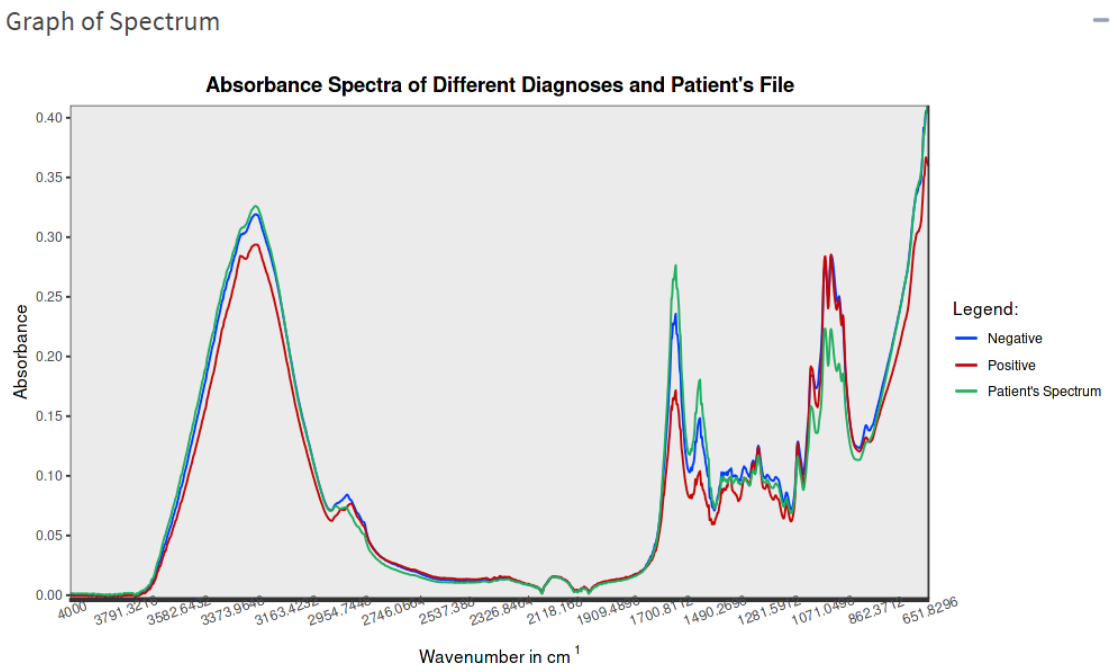


Figure 83: Graph Spectrum the file Negative_06 with color green as the legend

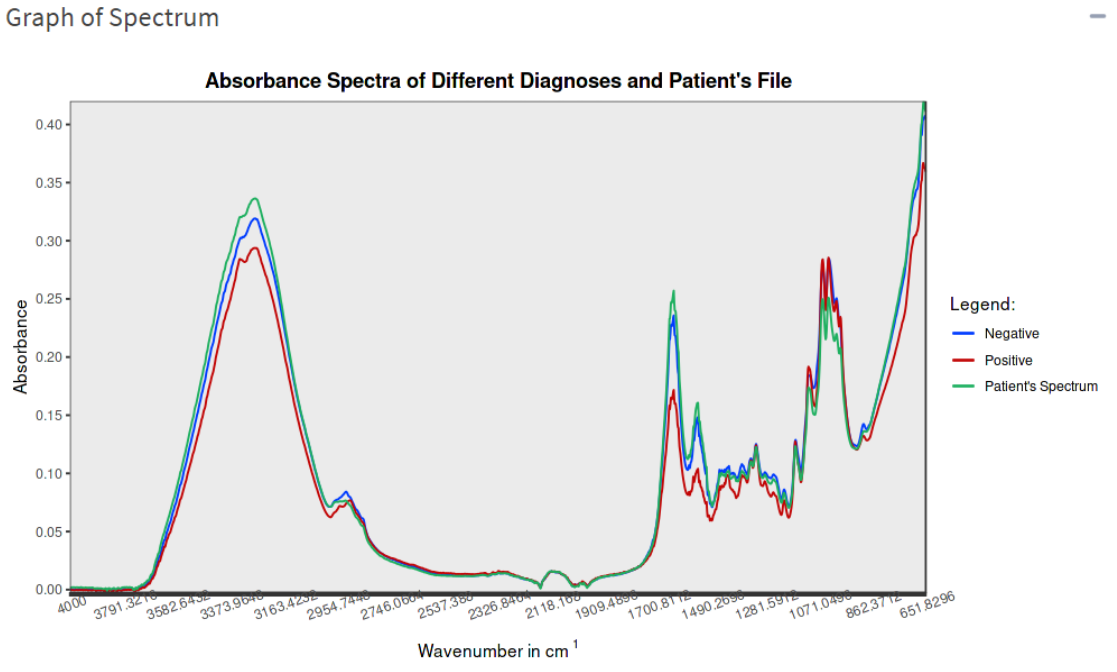


Figure 84: Graph Spectrum of the file Negative_07 with color green as the legend

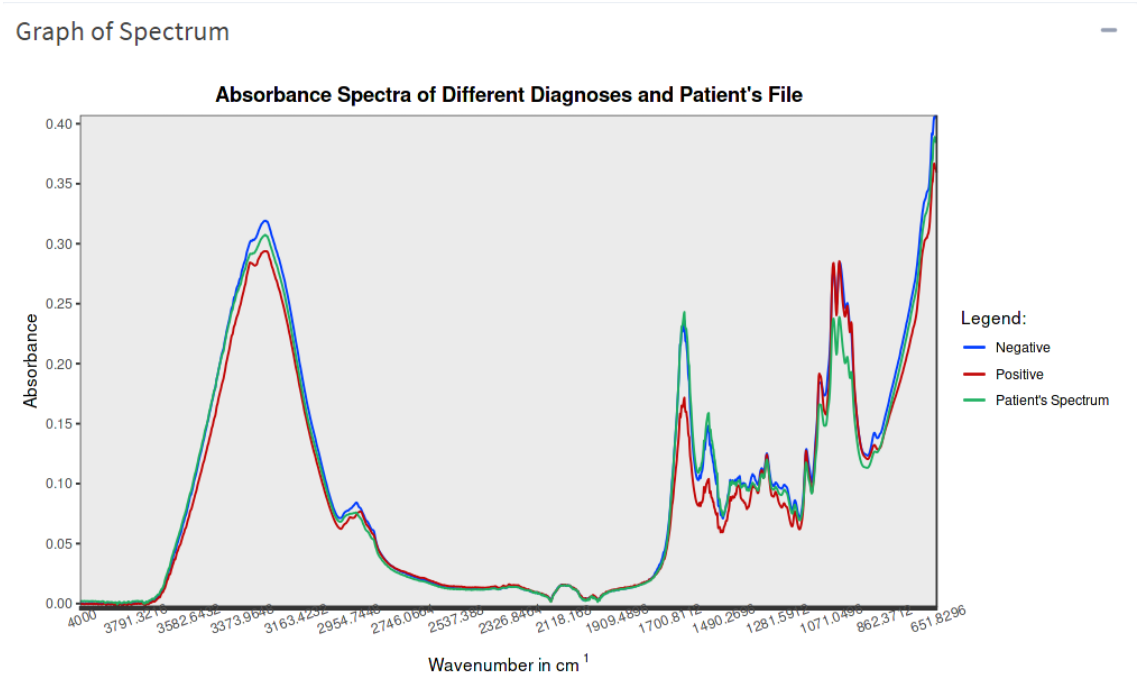


Figure 85: Graph Spectrum of the file Negative_08 with color green as the legend

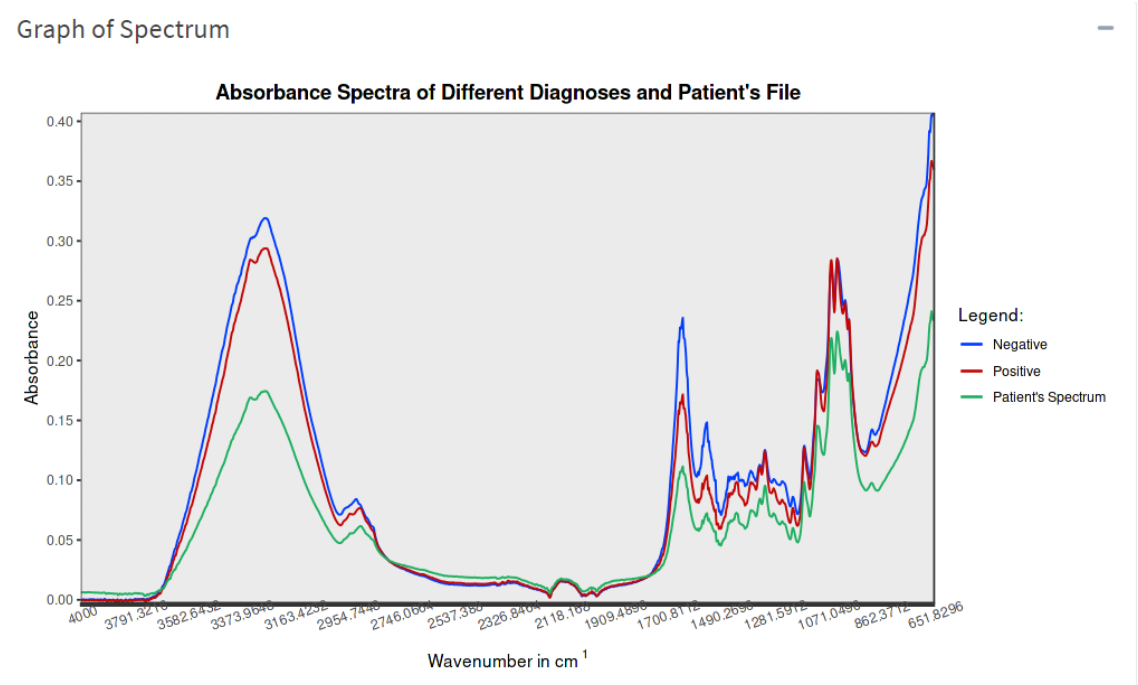


Figure 86: Graph Spectrum the file Negative_09 with color green as the legend

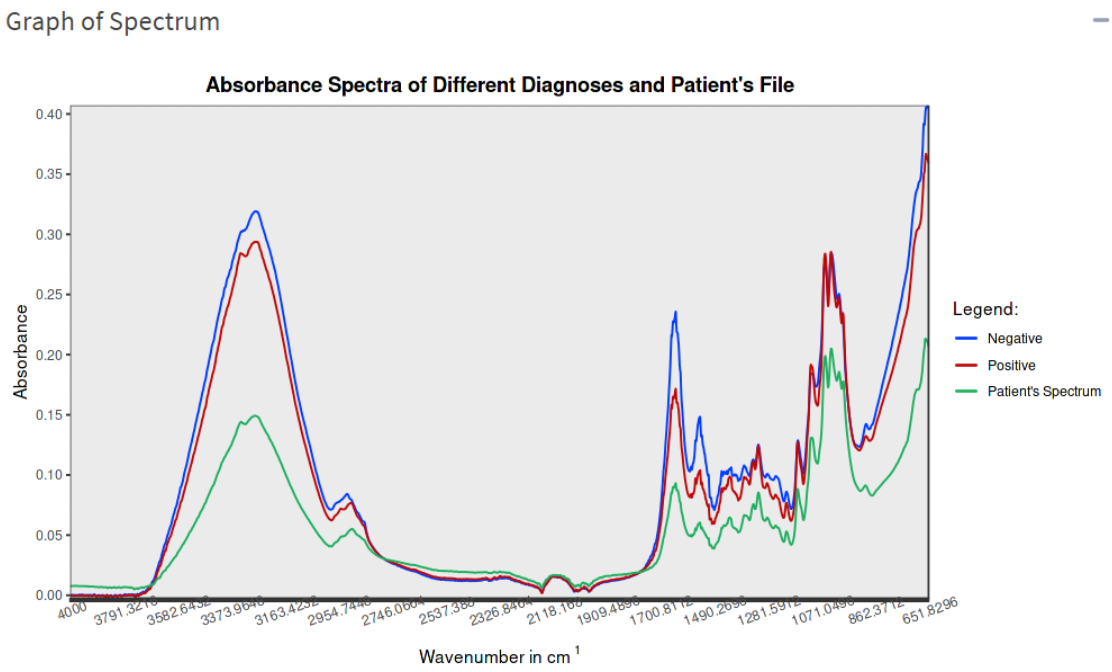


Figure 87: Graph Spectrum the file Negative_10 with color green as the legend

The graphs of the first eight diagnosed as COVID-19 negative via RT-PCR show that patients' spectra are close to the negative spectrum between wavenumbers 1638 to 1204 in cm^{-1}. The CSV file Negative_02 which was diagnosed incorrectly by the app has absorbance levels close to the positive spectrum in the aforementioned wavenumbers. Moreover, for graphs of Negative_09 and Negative_10, both having more than 98% percentage match, show great distances from the negative and positive spectra.

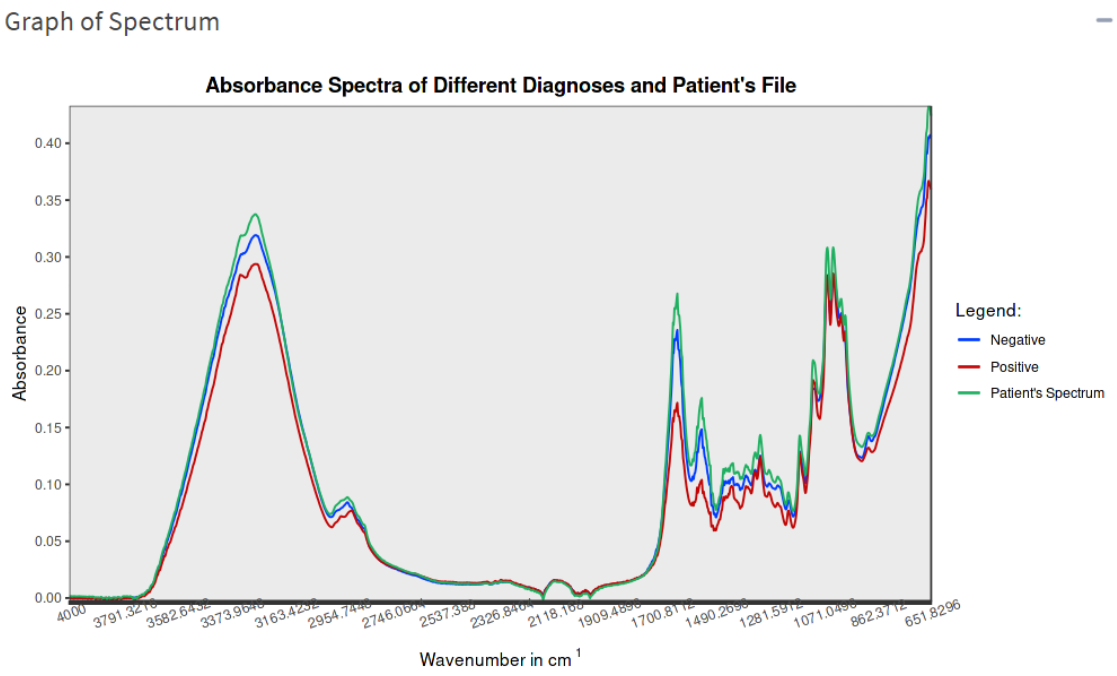


Figure 88: Graph Spectrum the file Positive_01 with color green as the legend

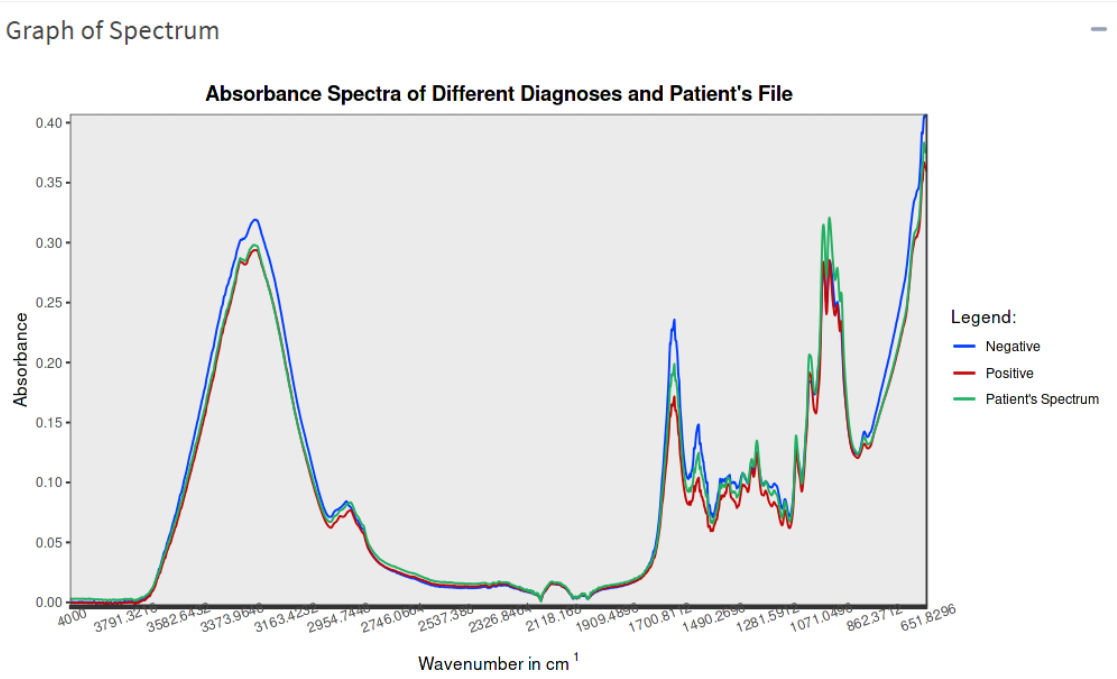


Figure 89: Graph Spectrum the file Positive_02 with color green as the legend

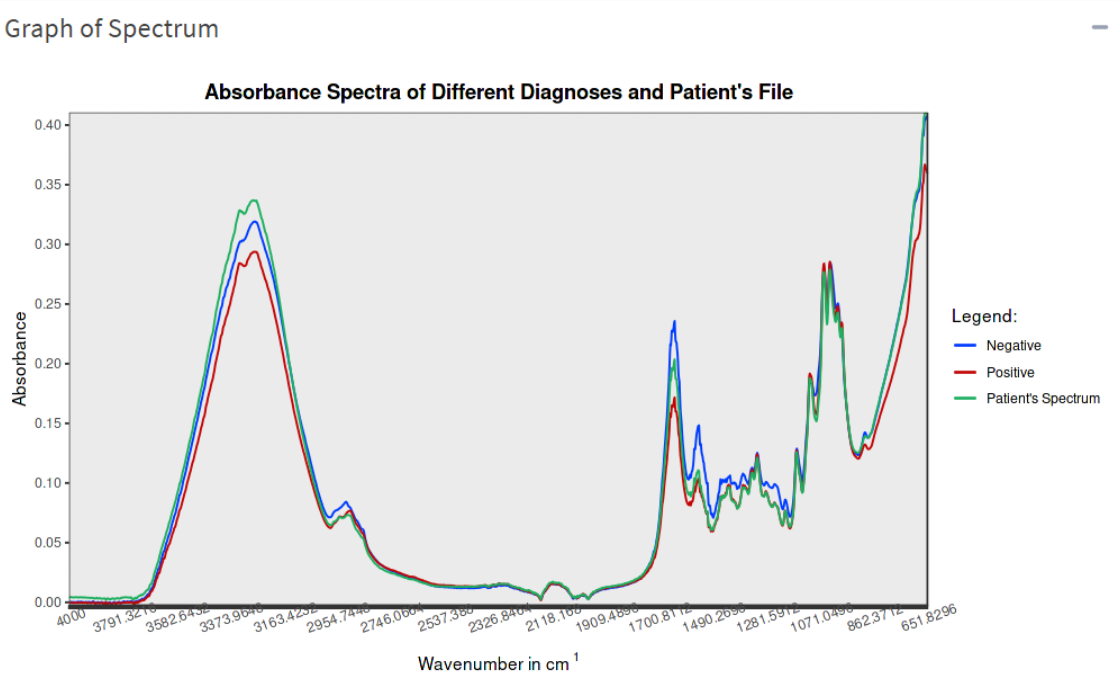


Figure 90: Graph Spectrum the file Positive_03 with color green as the legend

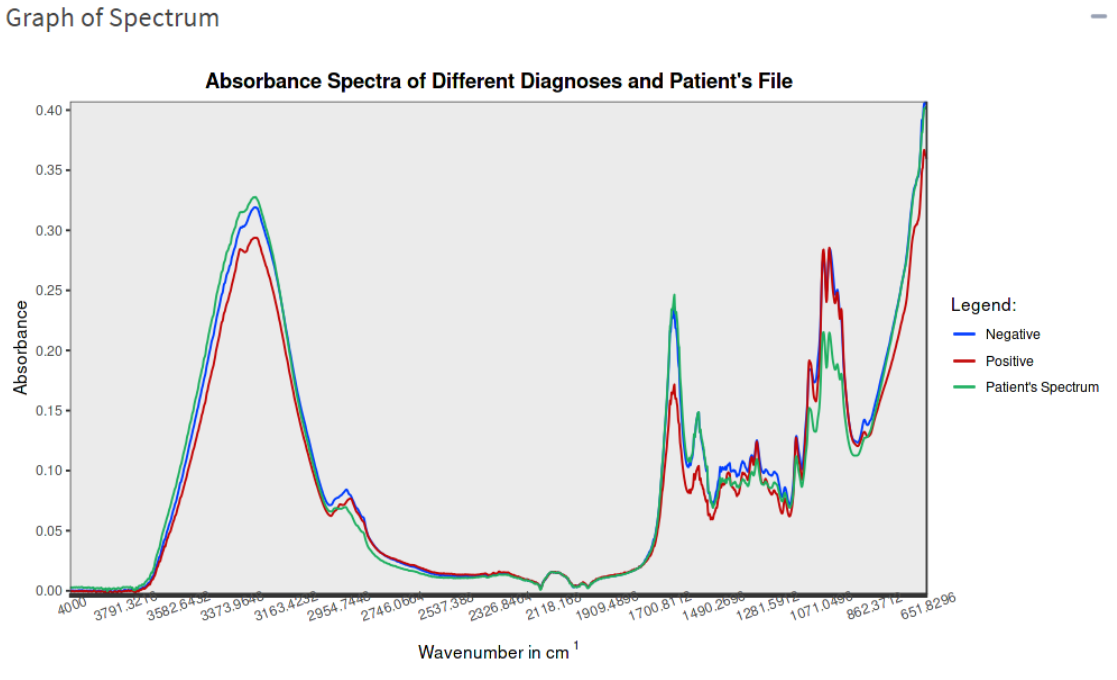


Figure 91: Graph Spectrum the file Positive_04 with color green as the legend

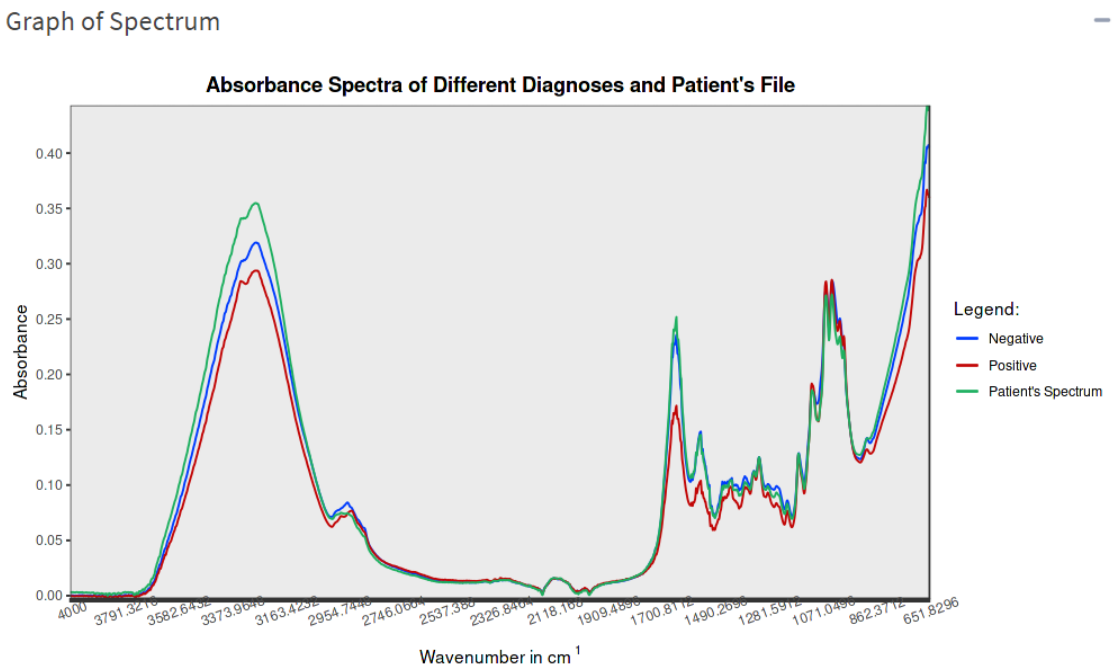


Figure 92: Graph Spectrum the file Positive_05 with color green as the legend

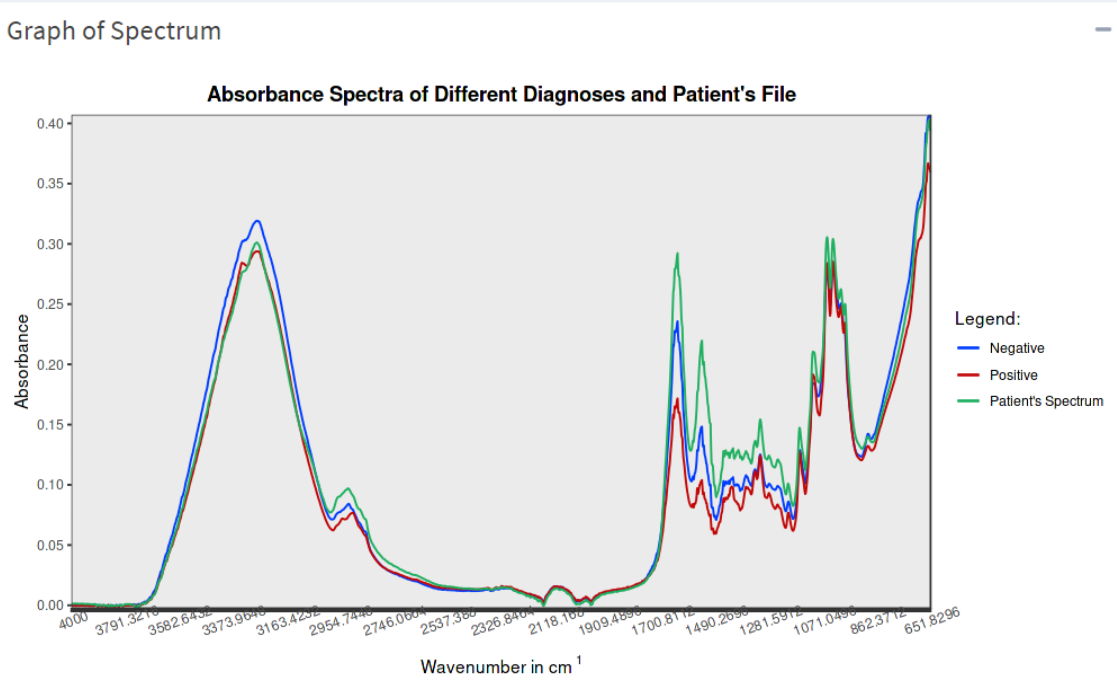


Figure 93: Graph Spectrum the file Positive_06 with color green as the legend

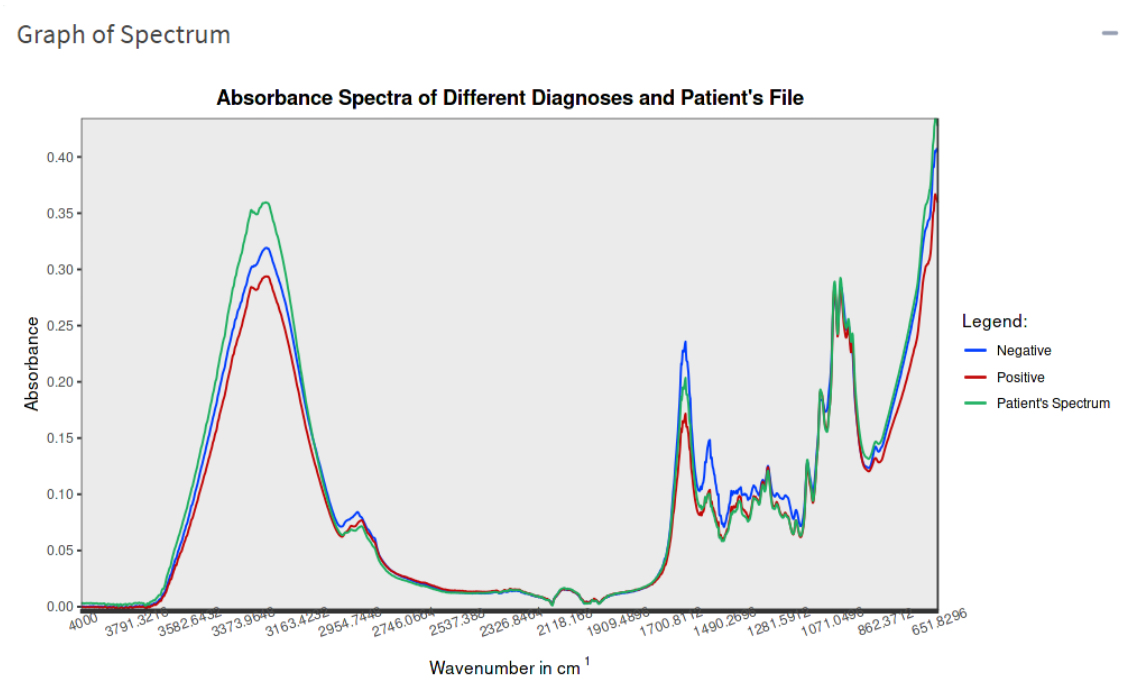


Figure 94: Graph Spectrum the file Positive_07 with color green as the legend

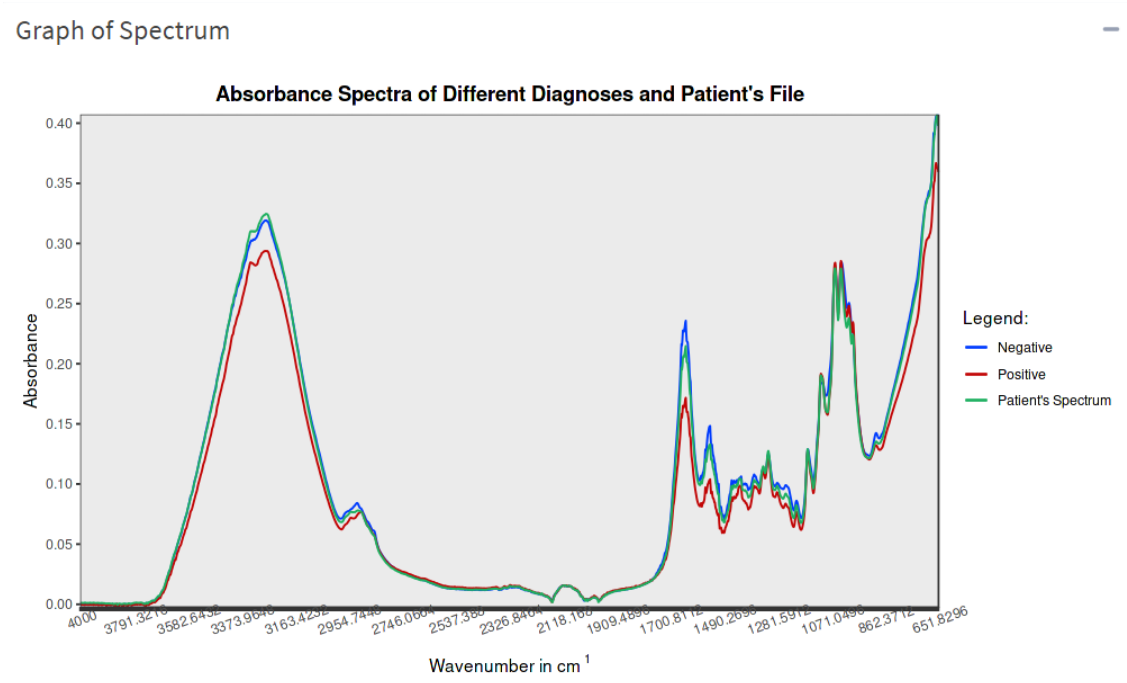


Figure 95: Graph Spectrum the file Positive_08 with color green as the legend

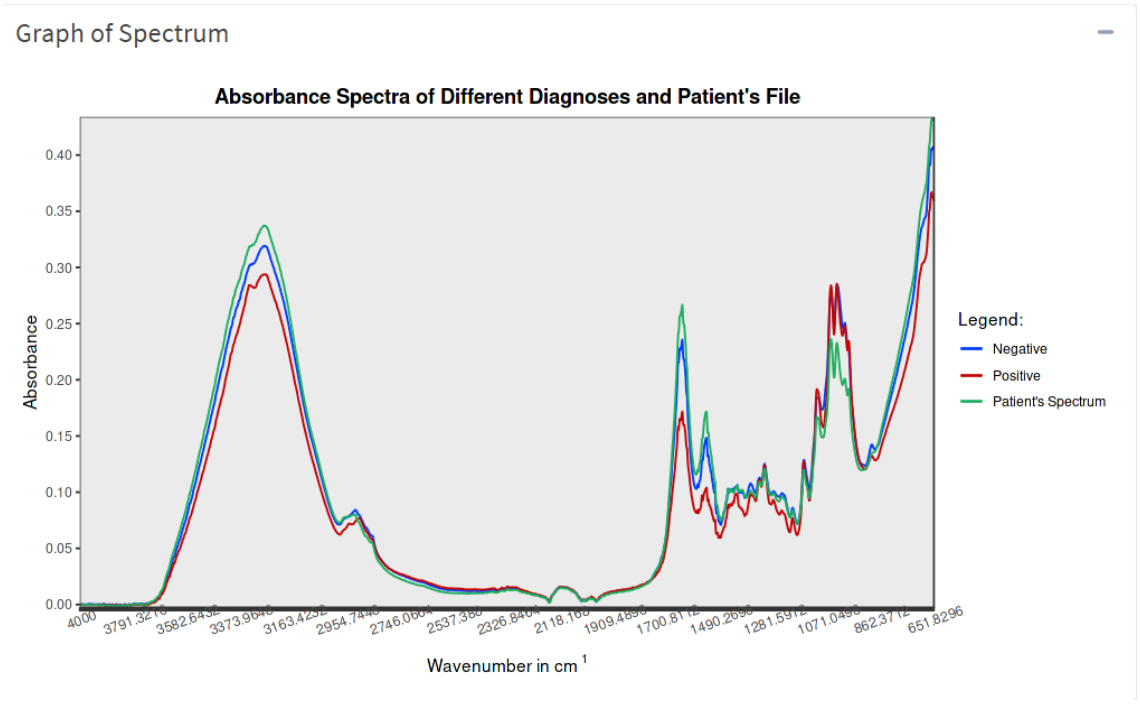


Figure 96: Graph Spectrum the file Positive_09 with color green as the legend

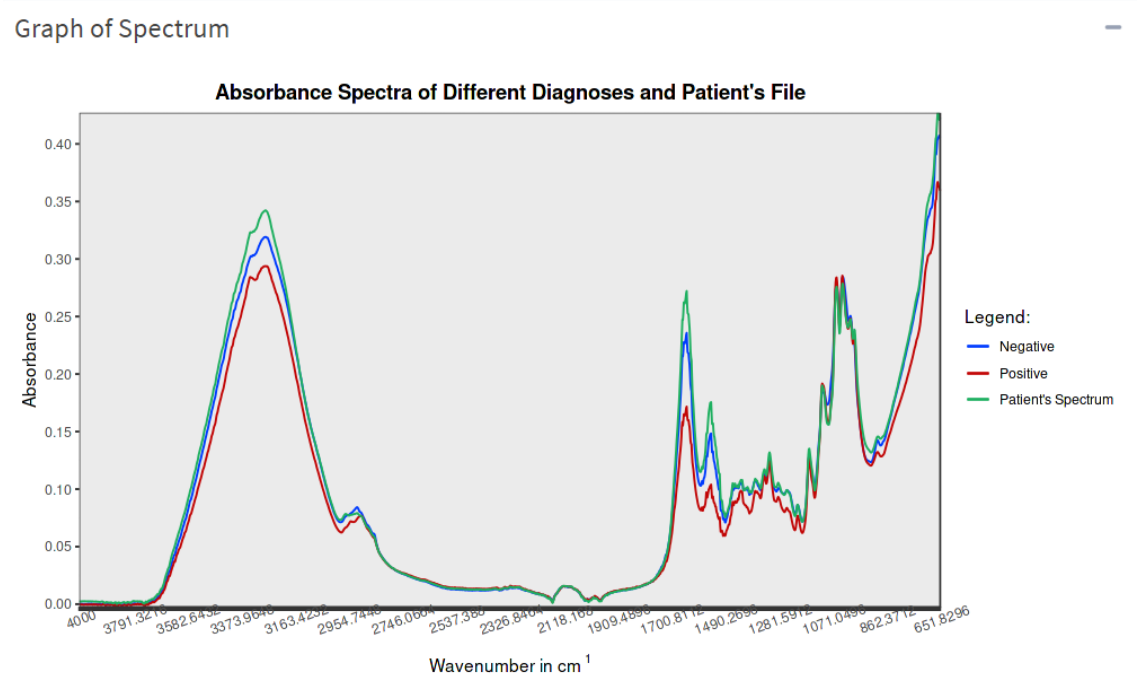
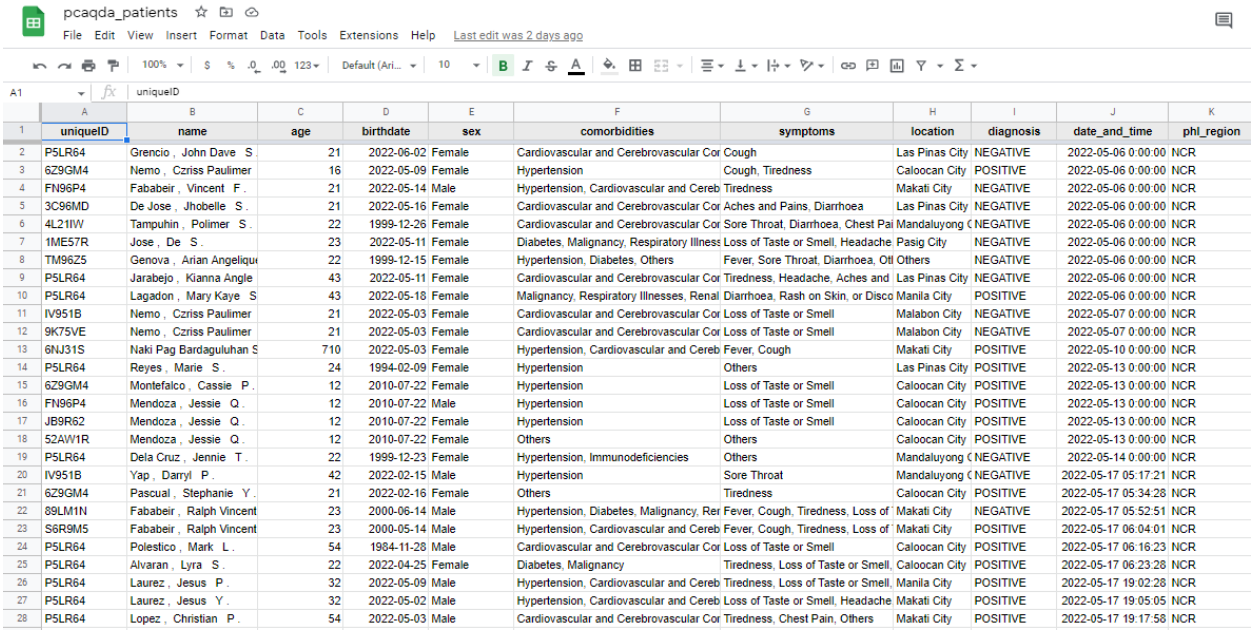


Figure 97: Graph Spectrum the file Positive_10 with color green as the legend

The spectrum of COVID-19 positive shows significant peaks at wavenumbers 1683 to 983 in cm^{-1}. The generated spectra of CSV files Positive_03 and Positive_07 display close distances of the significant peaks. For files Positive_04, Positive_06, Positive_08, and Positive_10, a few specific peaks were observed. Positive_04 has short distances of peaks at 1459 to 1140 in cm^{-1}. For Positive_06, it only shows a close peak absorbance at 3279 and the rest is neither close negative nor positive spectrum. Positive_08 has close peaks at 1140 to 1043 in cm^{-1} while Positive_10 is at 1084 to 983 in cm^{-1}. Furthermore, Positive_01 with matching percentage of 98% has no observable close peaks while Positive_09 with matching percentage of 99% is close to negative spectrum. Positive_02 and Positive_05, which were incorrectly diagnosed by the app, have peaks close to negative spectrum.

4.3.5 Database of the web application



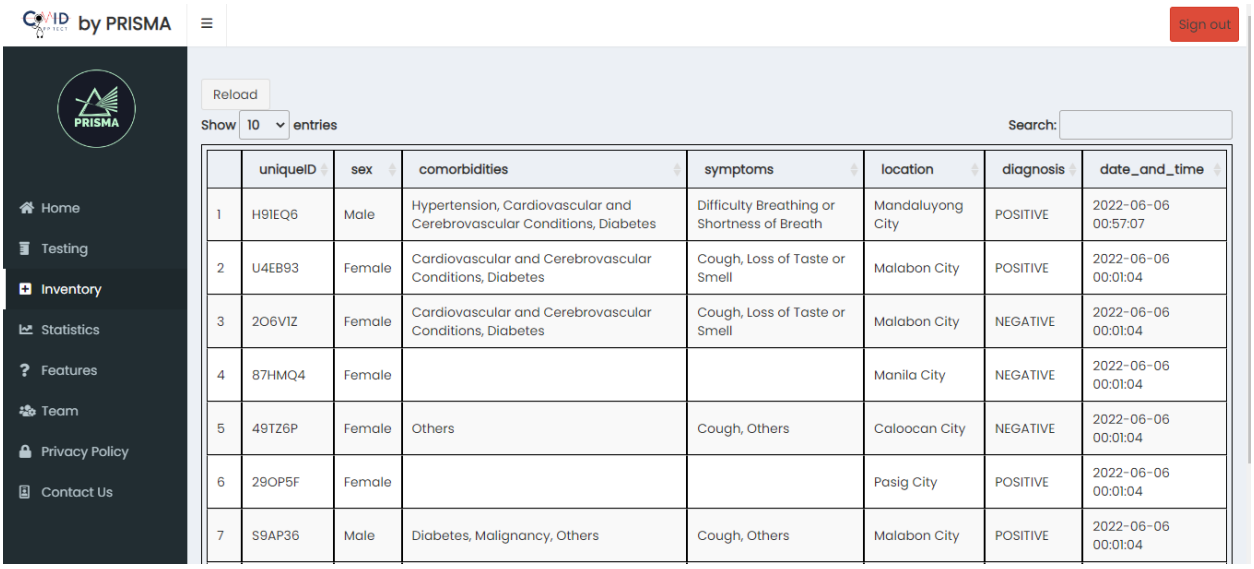
pcaqda_patients

File Edit View Insert Format Data Tools Extensions Help Last edit was 2 days ago

	uniqueID	name	age	birthdate	sex	comorbidities	symptoms	location	diagnosis	date_and_time	phl_region
2	P5LR64	Grencio , John Dave S	21	2022-06-02	Female	Cardiovascular and Cerebrovascular Cor	Cough	Las Pinas City	NEGATIVE	2022-05-06 0:00:00	NCR
3	6Z9GM4	Nemo , Czriss Paulimer	16	2022-05-09	Female	Hypertension	Cough, Tiredness	Caloocan City	POSITIVE	2022-05-06 0:00:00	NCR
4	FN96P4	Fababeir , Vincent F .	21	2022-05-14	Male	Hypertension, Cardiovascular and Cereb	Tiredness	Makati City	NEGATIVE	2022-05-06 0:00:00	NCR
5	3C96MD	De Jose , Jhobelle S .	21	2022-05-16	Female	Cardiovascular and Cerebrovascular Cor	Aches and Pains, Diarrhoea	Las Pinas City	NEGATIVE	2022-05-06 0:00:00	NCR
6	4L21IW	Tampuhin , Polimer S .	22	1999-12-26	Female	Cardiovascular and Cerebrovascular Cor	Sore Throat, Diarrhoea, Chest Pai	Mandaluyong C	NEGATIVE	2022-05-06 0:00:00	NCR
7	1ME57R	Jose , De S .	23	2022-05-11	Female	Diabetes, Malignancy, Respiratory Illness	Loss of Taste or Smell, Headache	Pasig City	NEGATIVE	2022-05-06 0:00:00	NCR
8	TM96Z5	Genova , Arian Angeliqu	22	1999-12-15	Female	Hypertension, Diabetes, Others	Fever, Sore Throat, Diarrhoea, Ot	Others	NEGATIVE	2022-05-06 0:00:00	NCR
9	P5LR64	Jarabejo , Kianna Angle	43	2022-05-11	Female	Cardiovascular and Cerebrovascular Cor	Tiredness, Headache, Aches and	Las Pinas City	NEGATIVE	2022-05-06 0:00:00	NCR
10	P5LR64	Lagadon , Mary Kaye S	43	2022-05-18	Female	Malignancy, Respiratory Illnesses, Renal	Diarrhoea, Rash on Skin, or Disco	Manila City	POSITIVE	2022-05-06 0:00:00	NCR
11	IV951B	Nemo , Czriss Paulimer	21	2022-05-03	Female	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Malabon City	NEGATIVE	2022-05-07 0:00:00	NCR
12	9K75VE	Nemo , Czriss Paulimer	21	2022-05-03	Female	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Malabon City	NEGATIVE	2022-05-07 0:00:00	NCR
13	6NJ31S	Naki Pag Bardaguluhan S	710	2022-05-03	Female	Hypertension, Cardiovascular and Cereb	Fever, Cough	Makati City	POSITIVE	2022-05-10 0:00:00	NCR
14	P5LR64	Reyes , Marie S .	24	1994-02-09	Female	Hypertension	Others	Las Pinas City	POSITIVE	2022-05-13 0:00:00	NCR
15	6Z9GM4	Montefalco , Cassie P .	12	2010-07-22	Female	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00	NCR
16	FN96P4	Mendoza , Jessie Q .	12	2010-07-22	Male	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00	NCR
17	JB9R62	Mendoza , Jessie Q .	12	2010-07-22	Female	Hypertension	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-13 0:00:00	NCR
18	52AW1R	Mendoza , Jessie Q .	12	2010-07-22	Female	Others	Others	Caloocan City	POSITIVE	2022-05-13 0:00:00	NCR
19	P5LR64	Dela Cruz , Jennie T .	22	1999-12-23	Female	Hypertension, Immunodeficiencies	Others	Mandaluyong C	NEGATIVE	2022-05-14 0:00:00	NCR
20	IV951B	Yap , Darryl P .	42	2022-02-15	Male	Hypertension	Sore Throat	Mandaluyong C	NEGATIVE	2022-05-17 05:17:21	NCR
21	6Z9GM4	Pascual , Stephanie Y .	21	2022-02-16	Female	Others	Tiredness	Caloocan City	POSITIVE	2022-05-17 05:34:28	NCR
22	89LM1N	Fababeir , Ralph Vincent	23	2000-06-14	Male	Hypertension, Diabetes, Malignancy, Rer	Fever, Cough, Tiredness, Loss of	Makati City	NEGATIVE	2022-05-17 05:52:51	NCR
23	S6R9M5	Fababeir , Ralph Vincent	23	2000-05-14	Male	Hypertension, Cardiovascular and Cereb	Fever, Cough, Tiredness, Loss of	Makati City	POSITIVE	2022-05-17 06:04:01	NCR
24	P5LR64	Polestico , Mark L .	54	1984-11-28	Male	Cardiovascular and Cerebrovascular Cor	Loss of Taste or Smell	Caloocan City	POSITIVE	2022-05-17 06:16:23	NCR
25	P5LR64	Alvaran , Lyra S .	22	2022-04-25	Female	Diabetes, Malignancy	Tiredness, Loss of Taste or Smell,	Caloocan City	POSITIVE	2022-05-17 06:23:28	NCR
26	P5LR64	Laurez , Jesus P .	32	2022-05-09	Male	Hypertension, Cardiovascular and Cereb	Tiredness, Loss of Taste or Smell,	Manila City	POSITIVE	2022-05-17 19:02:28	NCR
27	P5LR64	Laurez , Jesus Y .	32	2022-05-02	Male	Hypertension, Cardiovascular and Cereb	Loss of Taste or Smell, Headache	Makati City	POSITIVE	2022-05-17 19:05:05	NCR
28	P5LR64	Lopez , Christian P .	54	2022-05-03	Male	Cardiovascular and Cerebrovascular Cor	Tiredness, Chest Pain, Others	Makati City	POSITIVE	2022-05-17 19:17:58	NCR

Figure 98: Database of Users' inputs to COVID AppTect via Google Sheet

4.3.6 Stored Inputs in the Inventory Tab



	uniqueID	sex	comorbidities	symptoms	location	diagnosis	date_and_time
1	H9IEQ6	Male	Hypertension, Cardiovascular and Cerebrovascular Conditions, Diabetes	Difficulty Breathing or Shortness of Breath	Mandaluyong City	POSITIVE	2022-06-06 00:57:07
2	U4EB93	Female	Cardiovascular and Cerebrovascular Conditions, Diabetes	Cough, Loss of Taste or Smell	Malabon City	POSITIVE	2022-06-06 00:01:04
3	2O6V1Z	Female	Cardiovascular and Cerebrovascular Conditions, Diabetes	Cough, Loss of Taste or Smell	Malabon City	NEGATIVE	2022-06-06 00:01:04
4	87HMQ4	Female			Manila City	NEGATIVE	2022-06-06 00:01:04
5	49TZ6P	Female	Others	Cough, Others	Caloocan City	NEGATIVE	2022-06-06 00:01:04
6	29OP5F	Female			Pasig City	POSITIVE	2022-06-06 00:01:04
7	S9AP36	Male	Diabetes, Malignancy, Others	Cough, Others	Malabon City	POSITIVE	2022-06-06 00:01:04

Figure 99: Data Storage in the Inventory

Figure 98 displays the google sheet that serves as the main database of the inputs of the users in COVID AppTect. Also, database of users in the inventory tab are shown in Figure 99 that is generated immediately after storing the information by the user. The selected information such as unique ID, sex, comorbidities, symptoms, locations, diagnosis, and date and time stored is tabulated in the inventory tab after storing by the user in the testing tab.

4.3.7 Displayed Graphs in the Statistics Tab

Figure 100: NCR Level Graph with May 15, 2022 as the minimum date

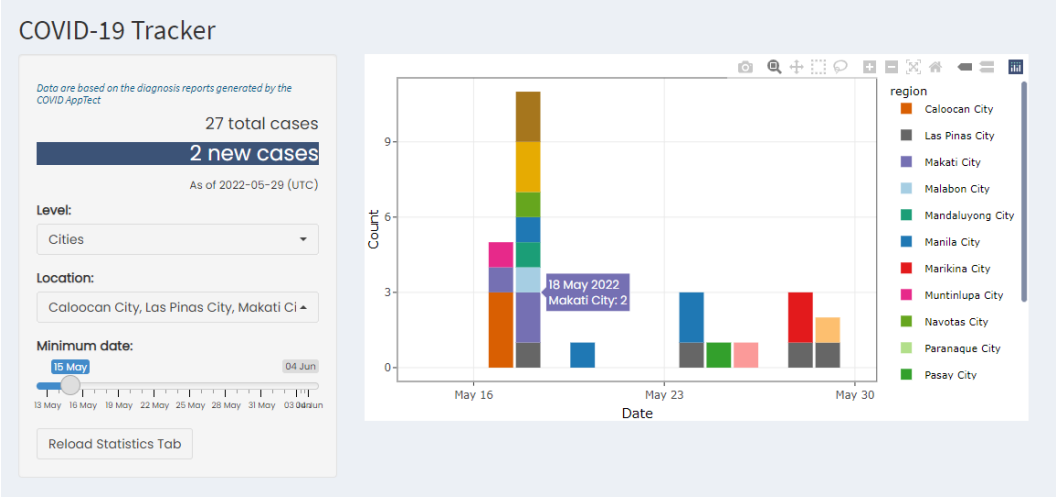


Figure 101: City Level Graph with all selected cities in NCR and May 15, 2022 as the minimum date

The figures shown above are the results of generated number of COVID-19 positive cases diagnosed by the COVID Apptect that was stored by the user. Total number of cases and new cases change depending on the diagnoses stored by the users.

4.4 Project Evaluation

Due to the increase of cases brought by COVID-19, health protocols regarding research to be conducted by external organizations aside from hospitals are halted. Hence, the actual use of OPS to be subjected in FTIR was not done. In alternative, the dataset from Centers of Disease Control and Prevention (CDC) was used and subjected to random sampling to make a mock CSV file to be used as dataset for evaluation of the web application.

For the evaluation, the proponents decided to divide the technical evaluation in terms of its Effectiveness. Efficiency, Satisfaction, Freedom from Risk, and Context Coverage. It uses Likert scales from 1 to 5, where 1 means the user strongly disagrees and 5 means strongly agrees to the statement.

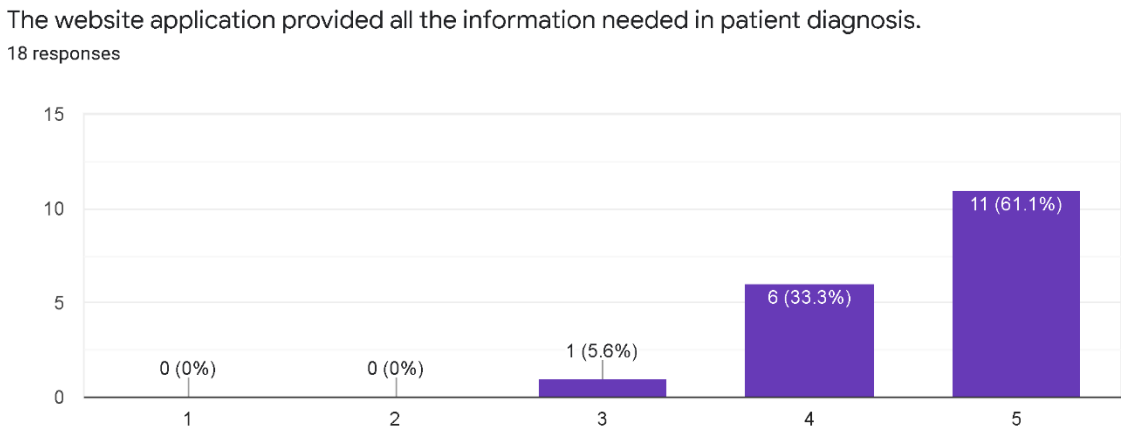


Figure 102: Bar Graph of the Likert-Scale Ratings to the Web Application's Effectiveness answering the first question of the Technical Evaluation Form from 17 participants

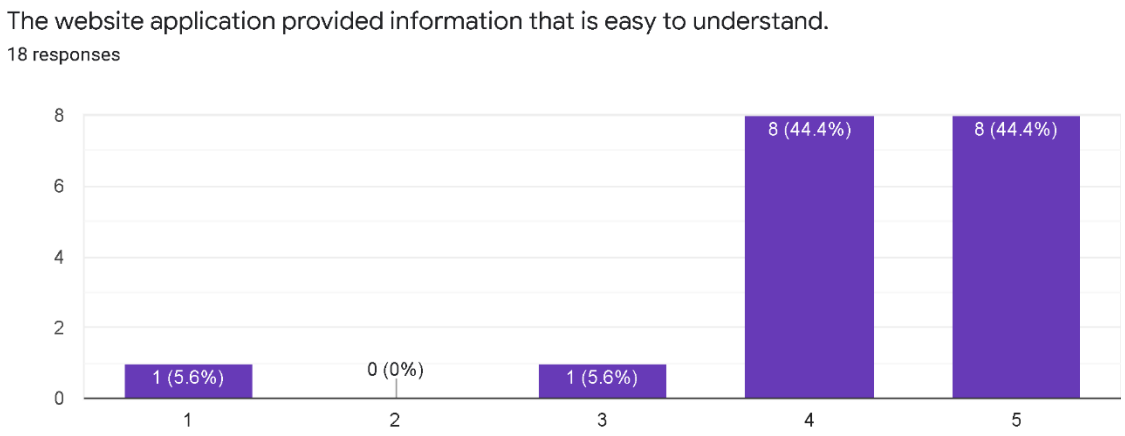


Figure 103: Bar Graph of the Likert-Scale Ratings to the Web Application's Effectiveness answering the second question of the Technical Evaluation Form from 17 participants

In terms of effectiveness, the graph above shows that the web application can give the essential information for the patient's diagnosis (refer to Figure 102). Furthermore, most participants believed that the information presented by the web application is simple to interpret (refer to Figure 103). However, there is a singularity in the graph who considers the web application complex in terms of the information it provides.

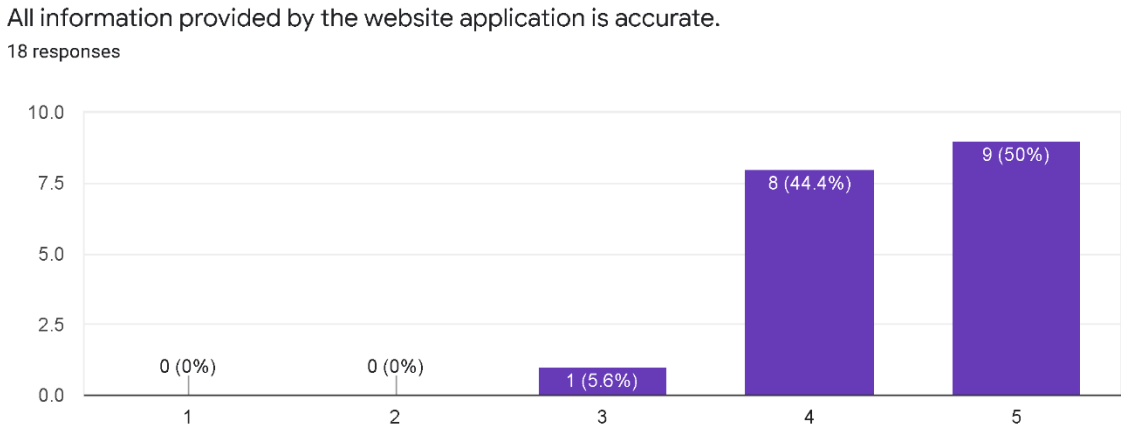


Figure 104: Bar Graph of the Likert-Scale Ratings to the Web Application's Efficiency from 17 participants

As for the efficiency, the users agreed that the information presented by the web application gives an accurate result for the diagnosis of the patients (refers to Figure 3). However, one of the feedbacks we've received is taking the window time of the OPS into account which can be a factor in diagnosing the patient and having a much more accurate result.

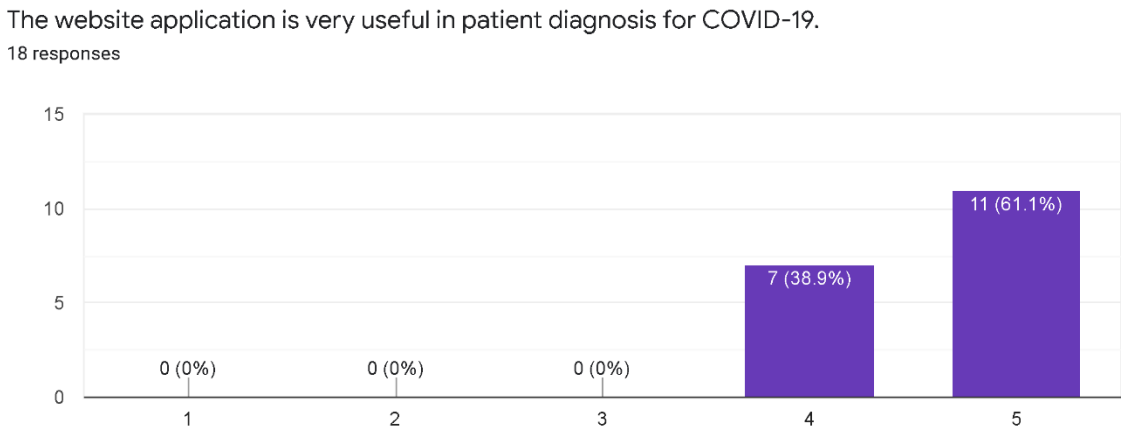


Figure 105: Bar Graph of the Likert-Scale Ratings about the Participant's Satisfaction answering the first question Technical Evaluation Form from 17 participants

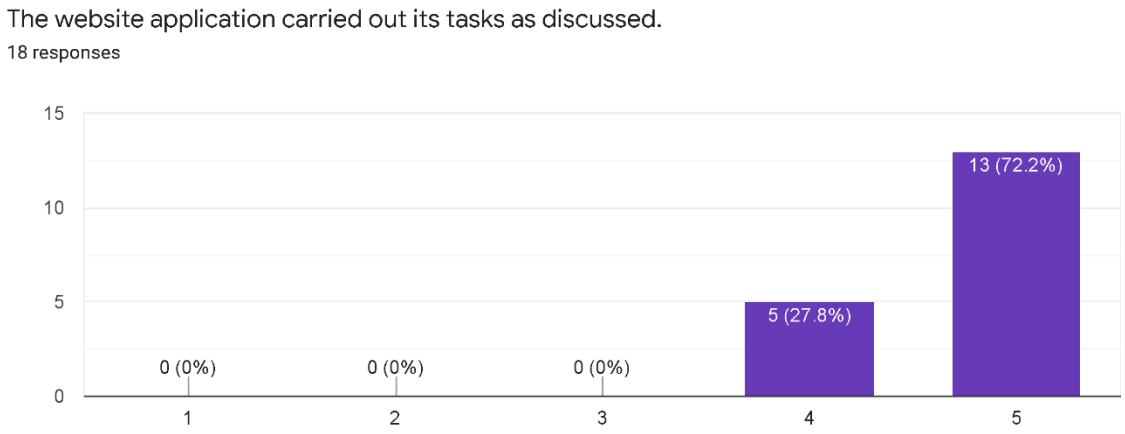


Figure 106: Bar Graph of the Likert-Scale Ratings about the Participant's Satisfaction answering the second question of the Technical Evaluation Form question from 17 participants

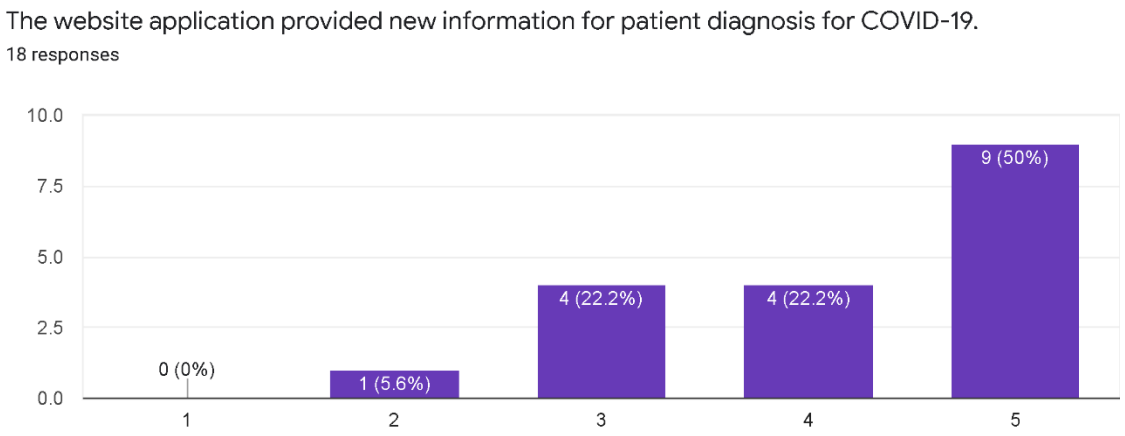


Figure 107: Bar Graph of the Likert-Scale Ratings about the Participant's Satisfaction answering the third question of the Technical Evaluation Form question from 17 participants

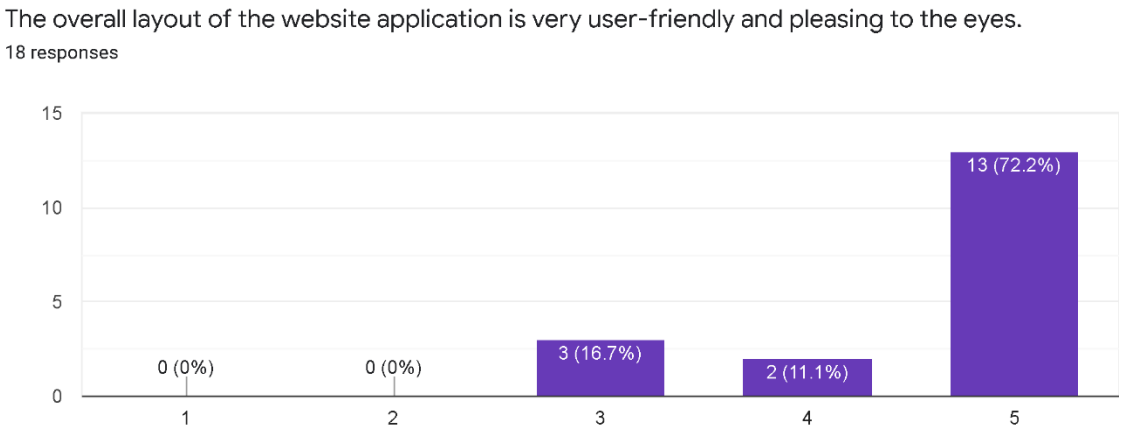


Figure 108: Bar Graph of the Likert-Scale Ratings about the Participant's Satisfaction answering the fourth question of the Technical Evaluation Form question from 17 participants

Majority of the users are satisfied with the layout and functionality of the web application. They find the web application helpful in diagnosing the patients for Sars-COV-2. In addition, with an average score of 4.72 out of 5, the participants agree that functionality of the web application was carried out properly as discussed. Users believed that the layout of the web application is pleasing and user-friendly.

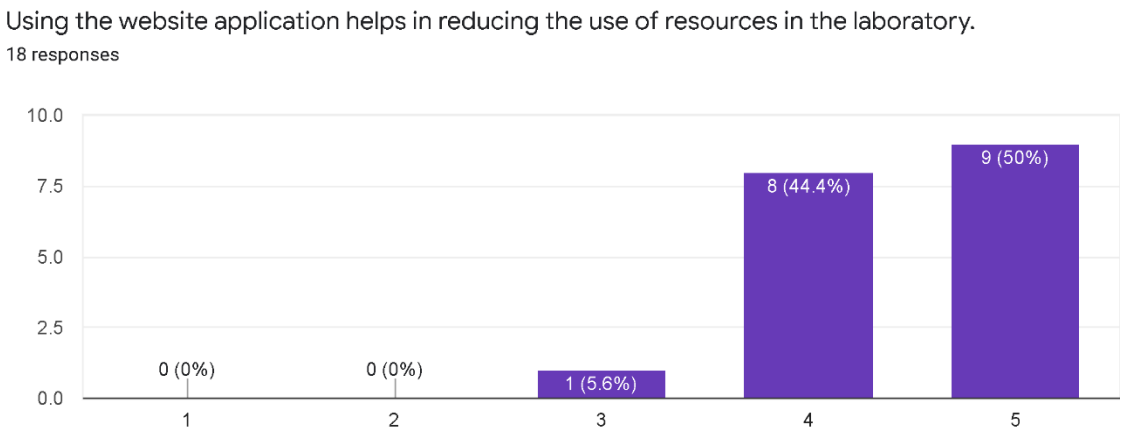


Figure 109: Bar Graph of the Likert-Scale Ratings about the Freedom From Risk part answering the first question of the Technical Evaluation Form question from 17 participants

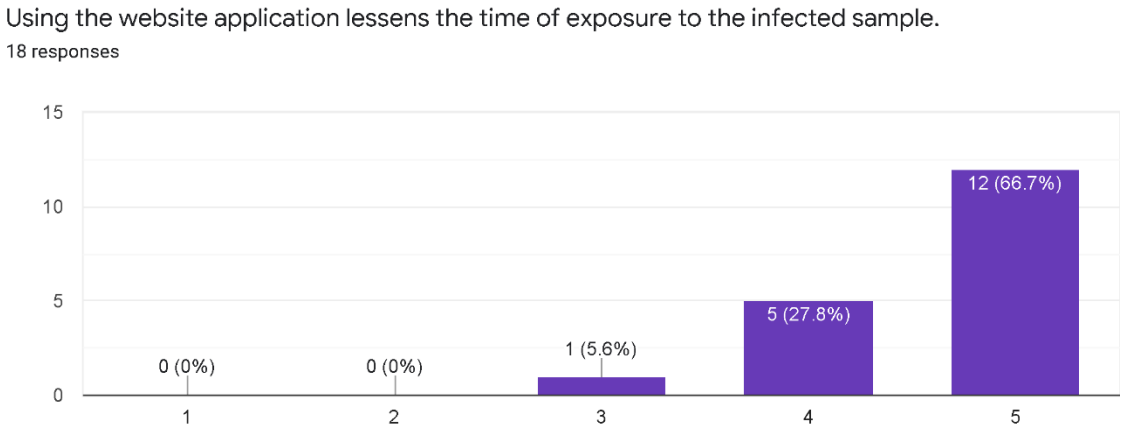


Figure 110: Bar Graph of the Likert-Scale Ratings about the Freedom From Risk part answering the second question of the Technical Evaluation Form question from 17 participants

The users agree that the web application provides a job in reducing the use of laboratory resources since CSV file of OPS from FTIR spectrometer, electronic devices such as computer, tablet and smartphone, and internet connectivity is the thing that is needed for the diagnosis of the patient. With the same reason, the users agree that it also lessens the time of exposure to the infected sample which reduces the risk of being infected by Sar-COV-2.

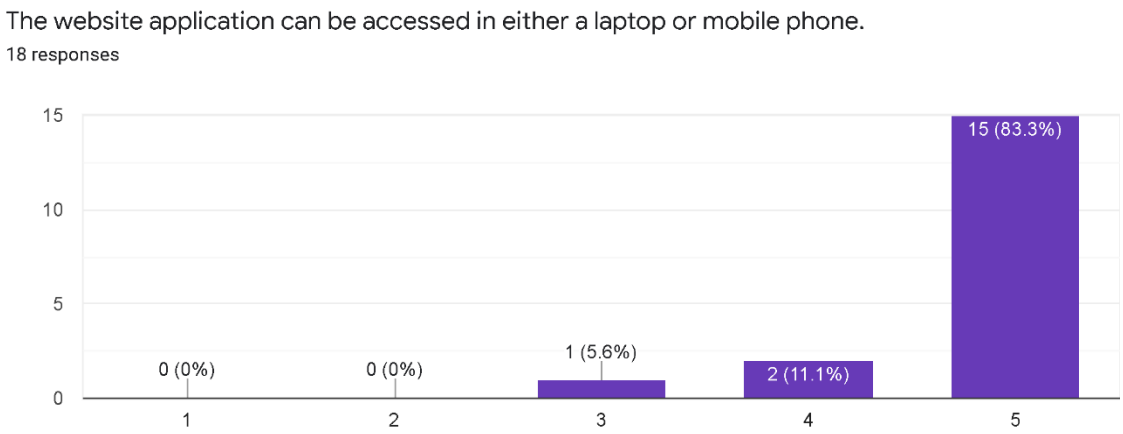


Figure 111: Bar Graph of the Likert-Scale Ratings about the Context Coverage part answering the first question of the Technical Evaluation Form question from 17 participants

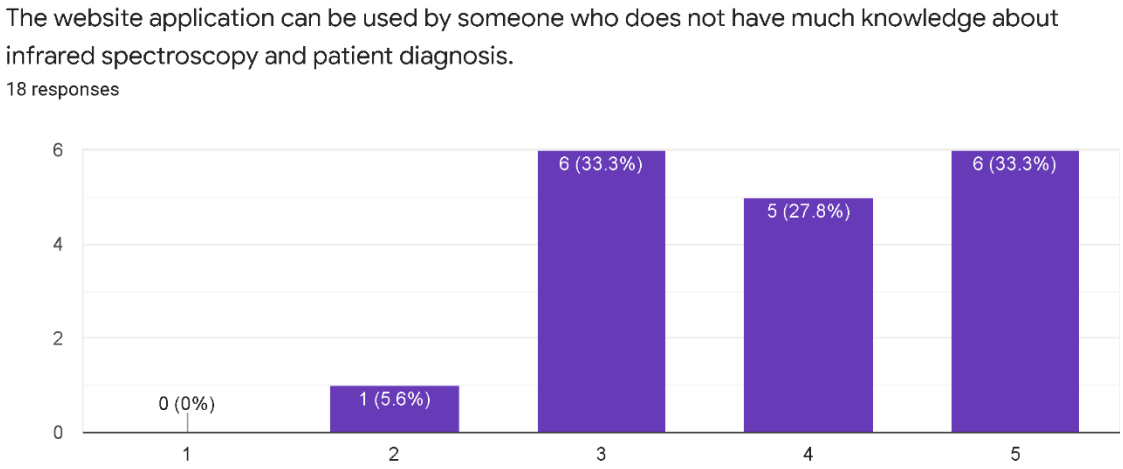


Figure 112: Bar Graph of the Likert-Scale Ratings about the Context Coverage part answering the second question of the Technical Evaluation Form question from 17 participants

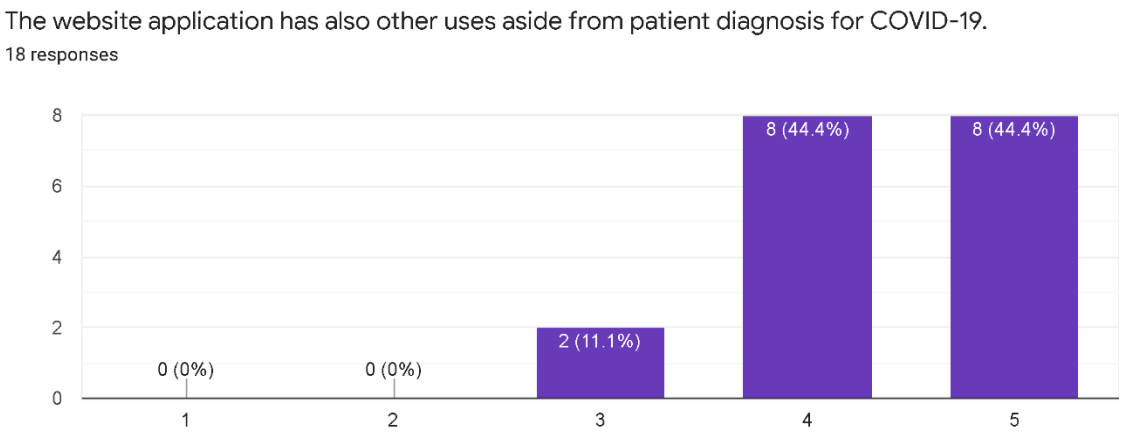


Figure 113: Bar Graph of the Likert-Scale Ratings about the Context Coverage part answering the third t question of the Technical Evaluation Form question from 17 participants

The web application being user-friendly was proven as 15 out of 17 users strongly agree that the web application can be accessed with any device, such as laptop, smartphone, or tablet. The web application still functions normally whatever electronic device uses as long as there is internet connectivity. However, most of the users stay neutral if the web application can be used by people who lack knowledge in FTIR spectroscopy and patients' diagnosis since the diagnosis still consists of medical factors to be accurate. On the other hand, the users agree that aside from diagnosing, the web application offers other uses such as generating and sending laboratory reports.

CHAPTER 5

Summary of Findings, Conclusion, and Recommendation

5.1 Summary of Findings

PLS DA				
Experiment Parameters	TP	TN	FP	FN
75% Train 25% Test	12	22	4	5
50% Train 50% Test	19	31	14	9
80% Train 20% Test	11	17	6	4
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	9	11	6	4
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	7	6	8	2

Table 5: Binary Classification Of The Experimental Parameters Using PLS DA

PLS DA			
Experiment Parameters	Accuracy / %	Sensitivity / %	Specificity / %
75% Train 25% Test	79.07	70.59	84.61
50% Train 50% Test	68.49	67.86	68.89
80% Train 20% Test	73.68	73.33	73.91
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	66.67	69.23	64.71
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	56.52	77.78	42.86

Table 6: Experimental Quality Performance Parameters Found For PLS-DA For Diagnosing Covid-19

PCA LDA				
Experiment Parameters	TP	TN	FP	FN
75% Train 25% Test	17	7	1	5
50% Train 50% Test	42	15	11	7
80% Train 20% Test	12	7	1	4
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	5	6	5	4
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	7	5	2	2

Table 7: Binary Classification Of The Experimental Parameters Using PCA LDA

PCA LDA			
Experiment Parameters	Accuracy / %	Sensitivity / %	Specificity / %
75% Train 25% Test	80	77.27	87.5
50% Train 50% Test	76	85.71	57.69
80% Train 20% Test	79.17	75	87.5
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	55	55.56	54.55
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	56.52	77.78	42.86

Table 8: Experimental Quality Performance Parameters Found For PCA-LDA For Diagnosing Covid-19

PCA QDA				
Experiment Parameters	TP	TN	FP	FN
75% Train 25% Test	21	14	3	8

50% Train 50% Test	37	21	11	11
80% Train 20% Test	16	13	3	4
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	4	10	3	6
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	5	9	2	4

Table 9: Binary Classification Of The Experimental Parameters Using PCA QDA

PCA QDA			
Experiment Parameters	Accuracy / %	Sensitivity / %	Specificity / %
75% Train 25% Test	76.09	72.41	82.35
50% Train 50% Test	72.5	77.08	65.63
80% Train 20% Test	80.56	80	81.25
80% Train 20% Test with 60 Positive and 60 Negative in the dataset	60.87	40	76.92
80% Train 20% Test with 50 Positive and 50 Negative in the dataset	70	55.56	81.82

Table 10: Experimental Quality Performance Parameters Found For PCA-QDA For Diagnosing Covid-19

Parameters	Models		
	PLS DA	PCA LDA	PCA QDA
Accuracy using Testing Set / %	73.38	79.17	80.56
Accuracy using Evaluation Set / %	85	35	85
Sensitivity / %	73.33	87.5	81.25
Specificity / %	73.91	75	80
AUC / %	90.88	83.6	87.2

Table 11: Summary Of Quality Performance Parameters Found For PLS-DA, PCA-LDA And PCA-QDA Models For Diagnosing Covid-19

PCA-QDA shows superiority in terms of accuracy through 80-20% partition of dataset. It has garnered an 80.56% accuracy. Hence, this method is used to verify further the quality performance by obtaining the AUC rate pf each prediction model. Table 11 shows the comparison of the AUC results of the prediction models. Despite having the least accuracy, PLS DA's AUC rate resulted the highest value among the three models. With its sensitivity and specificity as 73.33% and 73.91% respectively, showing a very near correlation, the AUC resulted to 90.88. Using the evaluation sets, both PLS DA and PCA QDA resulted to 85% accuracy. Either of the two models can be used in the application. However, PCA QDA's sensitivity and specificity are higher than PLS DA, thus, this model is preferred in the application.

Survey Statements	**Average Rating**
Effectiveness	
1. The website application provided all the information needed in patient diagnosis.	4.53
2. The website application provided information that is easy to understand.	4.24

Efficiency	
3. All information provided by the website application is accurate.	4.41
Satisfaction	
4. The website application is very useful in patient diagnosis for COVID-19.	4.59
5. The website application carried out its tasks as discussed.	4.71
6. The website application provided new information for patient diagnosis for COVID-19.	4.24
7. The overall layout of the website application is very user-friendly and pleasing to the eyes.	4.53
Freedom from Risk	
8. Using the website application helps in reducing the use of resources in the laboratory.	4.41
9. Using the website application lessens the time of exposure to the infected sample.	4.59
Context Coverage	
10. The website application can be accessed in either a laptop or mobile phone.	4.76
11. The website application can be used by someone who does not have much knowledge about infrared spectroscopy and patient diagnosis.	3.89
12. The website application has also uses aside from patient diagnosis for COVID-19.	4.35
OVERALL AVERAGE	4.44

Table 12: Web Technical Evaluation Average Rating

All aspects garnered an average of 4.44 depicting that all proponents agreed that the web application is effective, efficient, and safe. The proponents were also satisfied in using the

application. Moreover, they agreed that it is accessible using either a laptop or a mobile phone, it can be used by someone who is not knowledgeable in FTIR, and that it has other uses.

5.2 Conclusion

The following are the conclusions reached by the proponents based on the study's findings and results:

1. Being successful in developing predictive models using chemometrics packages of R programming, the predictive models, namely, PLS DA, PCA-LDA, and PCA-QDA were compared and which where PCA-QDA garnered the highest quality performance of 85%.
2. The web app (COVID AppTect) is capable of adapting to the screen size of any device which makes it a user-friendly web application. It can be used through smartphones, tablet, and laptop. It is capable of not only generating Portable Document Format (PDF) File but as well sending it to the patient's email address. In addition, the graphical representation of the CSV file from the FTIR spectrometer can be shown at the Testing tab when subjected to the web application.
3. Upon testing the web application using the dataset from CDC, the PCA-QDA accumulated the most balanced performance with 80.56% accuracy, 81.25% sensitivity, 80% specificity, and 87.2% AUC.
4. As the web application’s evaluation form adheres to the ISO 25010:2020 software evaluation standard, 16 out of 17 participants agree or strongly agree that the web application offers an accurate result for patient diagnosis. Furthermore, the web application performed its functionality, to the delight of the users.

5.3 Recommendations

Some of the medical practitioners who volunteer to try and evaluate the web application have some recommendations or suggestions that will help to improve the web application below are the list:

1. Consider the window time of the OPS.
2. Bulk input of the CSV Files and a “Next” button.

   There is a chance that the user might click incorrectly and repeat the same CSV file and send it to a wrong patient. It is better to input all the CSV file and if the user is done with that patient, it will only click the “Next” button and repeat the same process.
3. Severity of the Disease

   It is better to put the severity of the disease to know which patient will be treating first and it also informs the patient.
4. Enhanced user authentication

The following below is the research team recommendation

1. Instead of a Video Walkthrough in the homepage it is better to have a pop-up instructions per tab and the user will just do next button.

References

Barauna, V. G., Singh, M. N., Barbosa, L. L., Marcarini, W. D., Vassallo, P. F., Mill, J. G., Ribeiro-Rodrigues, R., Campos, L. C., Warnke, P. H., & Martin, F. L. (2021). Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Analytical Chemistry, 93*(5), 2950–2958. https://doi.org/10.1021/acs.analchem.0c04608

Barker, M. & Rayens, W. (2002). Partial Least Squares for Discrimination. *Journal of Chemometrics, 17(3), 166-173*. https://doi.org/10.1002/cem.785

Barth, A. (2007). Infrared Spectroscopy of Proteins. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1767(9)*. https://doi.org/10.1016/j.bbabio.2007.06.004

Biancolilo, A. & Marini, F. (2018). Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. *Frontiers in Chemistry, 6:576*. doi: 10.3389/fchem.2018.00576

Binnicker, M. J. (2020). Challenges and Controversies to Testing for COVID-19. *Journal of Clinical Microbiology, 58.* https://doi.org/10.1128/JCM.01695-20

Butler, H. J., Smith, B. R., Fritzsch, R., Radhakrishnan, P., Palmer, D. S., & Baker, M. J. (2018). Optimised Spectral Pre-Processing for Discrimination of Biofluids via ATR-FTIR Spectroscopy. *Analyst, 24*. https://doi.org/10.1039/C8AN01384E

Chai, W., Labbe, M. & Stedman, C. (2021). Big Data Analytics. https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Costa, F. F. (2014). Big Data in Biomedicine. *Drug Discovery Today, Volume 19 Issue 4*. https://doi.org/10.1016/j.drudis.2013.10.012

Farifteh, J. et al. (2006). Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). Retrieved from https://doi.org/10.1016/j.rse.2007.02.005huerta

Ferreira, S. L. C. (2019). Chemometric and Statistics | Experimental Design. *Encyclopedia of Analytical Science (Third Edition)*. https://www.sciencedirect.com/topics/chemistry/chemometrics

Food and Drug Association (FDA). (2021). Coronavirus Disease 2019 Testing Basics. Retrieved from https://www.fda.gov/consumers/consumer-updates/coronavirus-disease-2019- testing-basics

Huerta, M.T.H., Mayoral, L.P.C., Navarro, L.M.S., Andrade, G.M.A., Mayoral, E.P.C., Zenteno, E. & Campos, E.P. (2020). Should RT-PCR be considered a Gold Standard in the Diagnosis of COVID-19?. *J Med Virol, 11*. DOI: 10.1002/jmv.26228

Kucheryavskiy, S. (2021). PLS Discriminant Analysis. Retrieved from https://mdatools.com/docs/plsda.html

Lasch, P. (2012). Spectral Pre-Processing for Biomedical Vibrational Spectroscopy and Microspectroscopic Imaging. *Chemometrics and Intelligent Laboratory Systems*. https://doi.org/10.1016/j.chemolab.2012.03.011

Laser Interferometer Gravitational-Wave Observatory (LIGO). (n.d.). What is an Interferometer?. https://www.ligo.caltech.edu/page/what-is-interferometer?

Leaverton, P. E. & Zhu, Y. (2017). International Encyclopedia of Public Health (Second Edition). https://www.sciencedirect.com/topics/medicine-and-dentistry/biostatistics

Libretexts. (2020, August 15). *Infrared: Application*. Chemistry LibreTexts. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy/Infrared%3A_Application.

Liu, K., Shi, M. & Mantsch, H. (2005). Molecular and chemical characterization of blood cells by infrared spectroscopy: A new optical tool in hematology. *Blood Cells, Molecules and Diseases, 35 (3)*. https://doi.org/10.1016/j.bcmd.2005.06.009

Mevik, B. & Wehrens, R. (2020). Introduction to the pls Package. Retrieved from https://cran.r.project.org/web/packages/pls/vignettes/pls-manual.pdf

Morais, C.L.M. & Lima, K.M. (2018). Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *J. Braz, Chem. Soc., Volume 29 (3).* http://dx.doi.org/10.21577/0103-5053.20170159

Mortuza, G. (2020). Gaussian Discriminant Analysis. https://gmortuza.medium.com/gaussian-discriminative-analysis-e5701f12f3e9

Nocairi, H., Qannari, E. M., Vigneau, E., & Bertrand, D. Discrimination on Latent Components with Respect to Patterns: Application to Multicollinear Data. *Computational Statistics & Data Analysis, 48(1), 139-147.* 10.1016/j.csda.2003.09.008

Nogueira, M. S., Leal, L. B., Marcarini, W. D., Pimentel, R. L., Muller, M., Vassallo, P. F., Campos, L. C. G., dos Santos, L., Luiz, W. B., Mill, J. G., Barauna, V. G., & de Carvalho, L. F. das C. e S. (2021). *Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning*. Nature News. Retrieved February 10, 2022, from https://www.nature.com/articles/s41598-021-93511-2

Ollesch, J., Drees, S.L., Heise, H.M., Behrens, T., Bruning, T. & Gerwert, K. (2013). FTIR Spectroscopy of Biofluids Revisited: An Automated Approach to Spectral Biomarker Identification. *Analyst, 138.* https://10.1039/c3an00337j

Othman, N., Lee, K. Y., Radzol, A. R. M., W. Mansor, & N. I. (2019). *PCA-QDA Model Selection for Detecting NS1 Related Diseases from SERS Spectra of Salivary Mixtures*. Research Gate. Retrieved February 10, 2022, from

https://www.researchgate.net/publication/325452611_PCA-QDA_Model_Selection_for_Detecting_NS1_Related_Diseases_from_SERS_Spectra_of_Salivary_Mixtures

Perry, A. K. (2013). What is the Role of Biostatistics in Modern Medicine?. *Discovery Health*. health.howstuffworks.com/medicine/modern-treatments/biostatistics-in-modern-medicine.htm/printable

Rai, A. (2020). What is Big Data - Characteristics, Types, Benefits & Examples. https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/

Rasmussen, S.H. (2021). Quadratic Discriminant Analysis. https://towardsdatascience.com/quadratic-discriminant-analysis-ae55d8a8148a

Sánchez-Brito, M., Mata-Miranda, M. M., Martínez-Cuazitl, A., López-Mezquita, D. J., Guerrero-Ruiz, M., & Vázquez-Zapién, G. J. (2021, February 5). *Saliva analysis using FTIR spectroscopy to detect possible SARS-CoV-2 (COVID-19) virus carriers*. Revista mexicana de ingeniería biomédica. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-95322021000100001.

Santos, M. C. D., Morais, C. L. M., & Lima, K. M. G. (2021). ATR-FTIR spectroscopy for virus identification: A powerful alternative. *Biomedical Spectroscopy and Imaging*, *9*(3-4), 103–118. https://doi.org/10.3233/bsi-200203

Sartorius. (2020). What is Principal Component Analysis (PCA) and How It Is Used?. https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

Siqueira, L. F. S., Júnior, R. F. A., Araújo, A. A. de, Morais, C. L. M., & Lima, K. M. G. (2017,). *Lda vs. Qda for FT-mir prostate cancer tissue classification*. Chemometrics and Intelligent Laboratory Systems. Retrieved February 10, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0169743916303318

TestingXperts. (2020). 7 Key Benefits of Big Data Analytics in Healthcare. https://www.testingxperts.com/blog/Big-Data-Analytics-HSpalding, Kealthcare

Tobias, R.D. (n.d.). *An Introduction to Partial Least Squares Regression*. https://stats.idre.ucla.edu/wp-content/uploads/2016/02/pls.pdf

Wang, C., Wu, P., Yan, L., Ye, Z., Chen, H., Ling, H. (2020). Image Classification Based on Principal Component Analysis Optimized Generative Adversarial Networks. *Multimed Tools Appl 80, 9687-9701).* https://doi.org/10.1007/s11042-020-10137-8

Whittaker, E.T. & Robinson, G. (1924). The Calculus of Observations. Blackie & Son. Pp. 291-6. OCLC 11879… “Graduation Formulae Obtained by Fitting a Polynomial.

Wold, S., Sjostro, M. & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Retrieved from https://doi.org/10.1016/S0169-7439(01)00155-1

World Health Organization (WHO). 2021. Philippines Situation in COVID-19. https://covid19.who.int/region/wpro/country/ph

Zach. (2020). Introduction to Quadratic Discriminant Analysis. https://www.statology.org/quadratic-discriminant-analysis/