Detection of SARS-CoV-2 from raman spectroscopy data using machine learning models

*Nkgaphe Tsebesebe1,2*, Kelvin Mpofu1, Sphumelele Ndlovu1, Sudesh Sivarasu2 and Patience Mthunzi-Kufa1,2,3
1Council for Scientific and Industrial Research, National Laser Centre, P.O Box 395, Building 46A, Pretoria 0001, South Africa
2Department of Human Biology, Division of Biomedical Engineering, University of Cape Town, Cape Town 7935, South Africa
3School of Chemistry and Physics, University of KwaZulu-Natal, Durban 4001, South Africa

Abstract. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a member of the coronaviruses that caused the COVID-19 pandemic. The pathogenic SARS-CoV-2 virus can act as a miRNA sponge to lower cellular miRNA levels, making it a more dangerous human coronavirus. Diagnostic testing of the virus is intended to identify current infection in individuals and is performed when a person exhibits symptoms that are compatible with COVID-19. In this work, machine learning models (artificial neural network, decision tree, and support vector machine) are used to classify Raman spectroscopy samples as healthy or infected with SARS-CoV-2. The aim of the work is to introduce an alternative method for detecting SARS-CoV-2. The accuracy of the artificial neural network, the support vector machine and the decision tree were 94%, 90%, and 87%, respectively. The algorithms produced evidence of high recall and specificity. Hence, integrating Raman spectroscopy with machine learning has the potential to serve as an alternative diagnostic tool.

1 Introduction

Coronavirus disease 2019 (COVID-19) is a novel coronavirus of the genus β, made of a single-stranded RNA [1]. The outbreak of the virus was the result of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2]. The SARS-CoV-2 produces unidentified open reading frames (ORFs) with fusion, deletion, and/or frameshift [3]. The pathogenic SARS-CoV-2 virus can act as a sponge of miRNA to lower cellular miRNA levels, making it a more dangerous human coronavirus [3]. Together, the SARS-CoV-2 and COVID-19 introduced a huge challenge worldwide, with approximately 100 million infected cases and more than 20 million fatalities confirmed in 2021 [2, 4]. The rapid spread of COVID-19 was due to human-to-human transmissions, indefinite latency and nonspecific symptoms of the SARS-CoV-2 [2, 5]. These factors have led to the need to develop reliable...
and quick methods to diagnose SARS-CoV-2 [6, 7, 8]. Diagnostic testing of the virus is intended to identify current infection in individuals and is performed when a person exhibits symptoms that are compatible with COVID-19. The method must be quick and efficient in detecting SARS-CoV-2 [1]. As such, most of the attention has been focused on point-of-care, portable and easy-to-use devices.

Currently, polymerase chain reaction (PCR) is mainly used method to detect novel coronavirus. However, the method is associated with disadvantages such as a high false-negative rate and high capital [9]. An alternative method of diagnosing the coronavirus is a spectroscopic technique called Raman spectroscopy which works by determining the vibrational modes of molecules based on the inelastic scattering of photons [10]. Considering the nature of detecting diseases, Raman spectroscopy has gained more attention because of its ability to reflect the rotation energy level of molecules and their internal vibrations [11]. Raman spectroscopy achieves this capability by analysing molecular patterns from scattering spectra of chemical structures, and it is used for the identification of functional groups in molecules with non-destructive and accuracy [12]. Hence, the method is ideal for identifying material composition.

To date, accurate detection of SARS-CoV-2 for large-scale diagnosis is still unavailable [2]. As such, the use of traditional machine learning and deep learning algorithms has gained more attention in detecting the virus. Machine learning is a subset of artificial intelligence (AI) that involves developing computer models that can learn and make independent predictions based on the data provided. These models continually improve their accuracy through learned data [13]. Sanksham Jain et al. used machine learning-based approaches to improve the speed of screening SARS_CoV_2 using chest X-rays [14]. The study produced results with 86.93% accuracy. Yang Yanjun et al. used a deep learning model based on recurrent neural networks (RNN) to detect SARS-CoV-2 RNA in human nasopharyngeal specimens with 98.9% accuracy [15]. Furthermore, David Chen [16] coupled a stacked subssemble classifier model with an iteratively validated and automated feature selection and engineering workflow to predict COVID-19 infection status from Raman spectra of 250 human serum samples. The method achieved an accuracy of 98% from 10-fold cross-validation classification [16]. Lastly, experimental work has demonstrated the reliability of combining extreme gradient boosting (XGBoost) and Raman spectroscopy data in the detection of SARS-CoV-2 by 93.55% [17]. The current study continues with the need for the development of a rapid, sensitive, and reproducible method to diagnose SARS-CoV-2. Rapid and sensible detection of coronavirus is crucial for the control and prevention of the disease.

In this study, Raman spectrograms are classified into healthy, and SARS-CoV-2 infected using machine learning algorithms (i.e., artificial neural network, support vector machine, and decision tree). The choice of support vector machine in this work was directed by the ability of the model to detect complex relationships between linear and non-linear data without requiring a user to perform a lot of data transformation [12]. On the other hand, the decision tree is employed to break down complex data generated from Raman spectroscopy into manageable parts for evaluation. Lastly, the artificial neural network with fully connected layers is developed to improve the decreasing accuracy caused by the vanishing gradient in high-level neural networks. Previous work has shown that integrating ML with Raman spectroscopy reduces the false positive rate and reduces the detection cycle [17]. This paper demonstrates the reliability of machine learning algorithms combined with Raman spectroscopy in detecting SARS-CoV-2.

2 Data collection and methodology

The aim of this work is to develop and demonstrate the ability of machine learning algorithms to automate the detection of SARS-CoV-2 from Raman spectroscopy data. In this study, a
supervised learning approach is used to classify Raman spectrograms as healthy, and SARS-CoV-2 infected. In this study, labelled Raman spectroscopy data was used as input for the machine learning models (artificial neural network, support vector machine, and decision tree) that were used to train the models and produce the output. The algorithms were developed using Python programming language and executed on Google Colab [18]. The study follows these steps: (1) data transformation and splitting, (2) machine learning model training, (3) model evaluation, (4) comparison of the developed algorithms and (5) comparison of the algorithms in the study and other relative studies. The research methodology is shown in Figure 1.

![Research methodology of the current work.](image)

**Fig. 1.** Research methodology of the current work.

### 2.1 Dataset

The Raman spectroscopy data used in this work were downloaded from the publicly available SARS-CoV-2 dataset in the Kaggle database [19]. The study used 302 collected samples, 150 Raman spectral data from healthy people, and 152 Raman spectral data from SARS-CoV-2 infected people. Figure 2 shows an example of a Raman spectrogram for the two classes. The data labels in the two classes were confirmed by polymerase chain reaction (PCR). This study used unprocessed spectrograms produced by Raman spectroscopy as shown in Figure 2 [a]. The spectrograms were transformed from Figure 2 [a] to Figure 2 [b] by transposing the original data. The spectrograms were split into 80% training and 20% testing of 302 samples. The algorithms in the current work were evaluated from 62 spectral data, 34 healthy and 28 SARS-CoV-2. Although the data is small, this study used 97 more samples than the work of David Chan, which used Raman spectra of 250 human serum samples [16]. Moreover, the support vector machine has been reported to produce better results with smaller datasets [12]. This work validates the performance of the model by employing the grid search on another traditional model (decision tree) to develop an optimized model that can classify the small dataset with high efficiency. Moreover, the optimization technique is employed in the deep learning model (ANN) to determine a lightweight model to deal with the small dataset. To ensure that each feature in the dataset contributes equally to the learning process, the standard scaler method was used to standardize the independent features in a fixed range. The method removes the mean and scales each feature to a unit variance [12].
Fig. 2. A graphical representation of the Raman spectral data [a] before transformation, [b] after transformation, and [c] healthy and SARS-CoV-2 infected as confirmed by polymerase chain reaction (PCR).

2.2 Artificial neural network (ANN)

Artificial neural networks (ANNs) are a computational network inspired by biological neurons; the model consists of connected units (artificial neurons) that receive inputs and produce outputs based on predefined activation functions [20]. The artificial neural network in this work was initialized as sequential. All independent variables from the data set were grouped into a matrix of features. The feature scaling was performed from the standardization technique to convert the values in the dataset to range from -3 to +3. The only hidden layer of the model was created from a dense class with 2 neurons in the respective layer and rectified linear unit (ReLU) activation function. The output layer had only one unit with a sigmoid activation function. The stochastic gradient descent in the model was performed from the Adam optimizer. The model had a binary cross-entropy loss function, and the performance was computed by monitoring the accuracy from the matrices. The model was trained with 32 spectral data per patch for 100 epochs.
2.3 Support vector machine (SVM)

The support vector machine is a supervised learning algorithm that performs the separation of data points with dimension by determining the \((p - 1)\) dimensional hyperplane that can maximize the distance from the determined hyperplane to the nearest data point on each class \([12]\). This work used default parameters for the SVM classifier to determine the biggest margin from the hyperplane to the nearest data points in each class.

2.4 Decision tree

A decision tree is a type of supervised learning algorithm that divides tree structures into subtrees while performing the classification of the input data \([21]\). The quality of splitting of spectral data in this work was performed by the entropy function. The function considers a product of the probability of each class and \(\log_2\) of the probability of each class. The equation of the entropy function employed in the decision classifier of the current work is as follows:

\[
\text{Entropy} = \sum_{i=2}^{c} -p_i \times \log_2(p_i)
\]

where \(C\) is a representation of class number and \(p\) is the probability of sample data in each class. Furthermore, the decision tree classifier in this work used 8 minimum number of spectral data to split its internal node to a maximum tree depth of 17.

2.5 Confusion matrices

The performance of the algorithms in the current study were evaluated using performance metrics, where \(TP\): true positive, \(TN\): true negative, \(FP\): false positive and \(FN\): false negative, represents number of SARS-CoV-2 Raman spectral data that are predicted as SARS-CoV-2, the number of healthy Raman spectral data that are predicted as healthy, the number of healthy Raman spectral data that are predicted as SARS-CoV-2 and the number of SARS-CoV-2 Raman spectral data that are predicted as healthy, respectively.

2.5.1 Accuracy

The ability of the algorithms to correctly predict the Raman spectral data was measured from the accuracy formula as follows:

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

2.5.2 Sensitivity

The ability of the algorithms to differentiate spectral peaks was measured from the sensitivity formula as follows:

\[
\text{Sensitivity} = \frac{TP}{FN + TP}
\]

2.5.3 Specificity

The ability of the algorithms to identify several true negatives that were correctly identified in spectral data was calculated from the specificity formula as follows:
3 Results and discussion

3.1 Performance metrics

The performance of the algorithms in classifying raw Raman spectroscopy data into healthy and SARS-CoV-2 was evaluated from the performance matrices as shown in Figure 3. The ANN algorithm correctly classified 28 of the 28 spectral data as infected with SARS-CoV-2 and 30 of the 34 spectral data as healthy. The algorithm incorrectly classified 4 spectral data as healthy while SARS-CoV-2 infected. On the other hand, the SVM correctly predicted 26 of 28 spectral data as SARS-CoV-2 infected and 30 of 34 spectral data as healthy. The algorithm incorrectly classified 2 of 28 spectral data as infected with SARS-CoV-2 while healthy and misclassified 4 of 30 spectral data as healthy while infected with SARS-CoV-2. Lastly, the decision tree algorithm correctly classified 26 of 30 spectral data as SARS-CoV-2 infected and 28 of 34 spectral data as healthy. The algorithm misclassified 2 of 28 spectral data as infected with SARS-CoV-2 while it was healthy and 4 of 34 spectral data as healthy while SARS-CoV-2 infected. Amongst the algorithms, the artificial neural network outperforms all the algorithms in terms of accuracy, by being able to detect the SARS-CoV-2 infected spectral data without any false detections. However, it misclassifies healthy spectral data by 11.76%. This misclassification is identified in the three algorithms of the current work, with the support vector machine having the same percentage and the decision tree with a higher percentage (17.67%). The major downfall of both SVM and decision tree is noticed in the misclassification of the SARS-CoV-2 infected spectral data by 7.14% each.

\[ Specificity = \frac{TN}{FP + TN} \]  

**Fig. 3.** Confusion matrices of the algorithms on the raw Raman spectroscopy data [a] artificial neural network, [b] support vector machine, and [c] decision tree.
3.2 Clinical indicators of the models

In the current work, the models were evaluated from three evaluation metrics. The performance of the three algorithms in the Raman spectroscopy data in both cases is shown in Table 1. The results are determined from the formulas (2)-(4). The artificial neural network can accurately predict the Raman spectrogram data by 93.55%. This high percentage follows from the ability of the model to differentiate the spectral peaks by 100%. However, the model falls short of identifying several true negatives that were correctly identified in the spectral data by 11.76%. On the other hand, the support vector machine algorithm accurately predicts the raw Raman spectroscopy data by 90.32%, while differentiating the spectral peaks by 92.68%. Similarly, the support vector machine can identify several true negatives in spectral data by 88.24%. Lastly, the decision tree has the lowest accuracy of 87.10%, with 92.86% in differentiating the spectral peaks and 82.35% in identifying true negatives in the spectral data. The performance of each algorithm depends mainly on the quality of the input spectral data [22]. The high accuracy of the ANN is due to the ability of the model to deal with the distribution of the dataset. Furthermore, the default value of the learning rate in the Adam optimizer facilitated a lower rate at which the optimal weights of the model were determined. This implies that there were smaller changes in the model in response to errors. The model contained only one hidden layer, which increased the generalizability of the model in the test data to avoid overfitting. Lastly, the nature of the input dataset resembles balanced data, that is, the proportion of the original false negatives was low compared to the negative original true negatives. As a result, the ANN algorithm reduced the number of false negatives, providing a 7.14% increase in sensitivity compared to other algorithms in the current work. Conversely, the nature of the spectral data produced a decision boundary with maximum margins in both populations. As such, the default support vector machine classifier obtained the best hyperplane to classify the test data with 90.32%. The good performance of the models is validated by the higher F1 scores.

Table 1. Detection results of the algorithms on Raman spectroscopy data from SARS-CoV-2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>F1-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>93.548</td>
<td>100.000</td>
<td>88.235</td>
<td>93.333</td>
</tr>
<tr>
<td>SVM</td>
<td>90.323</td>
<td>92.857</td>
<td>88.235</td>
<td>89.655</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>87.097</td>
<td>92.857</td>
<td>82.353</td>
<td>86.667</td>
</tr>
</tbody>
</table>

3.3 Model comparison

Wandan Zeng et al. [1] experimentally demonstrated the reliability of combining extreme gradient boosting (XGBoost) and Raman spectroscopy data in the detection of SARS-CoV-2. The result of their work is presented with a comparison of this work in Table 2. Their experimental work achieved an accuracy of 93.55% after performing data processing (normalization, smooth denoising) on the raw data [1]. Similarly, the artificial neural network of this work achieved the same accuracy (93.55%) in two decimal places, working with the Raman data. Although both works achieve the same accuracy, the current work offers a 10% increase in sensitivity. This increase was due to a low proportion of original false negatives in the dataset as compared to the negative original true negatives. This proportionality allowed the artificial neural network of this work to learn effectively. Table 3 shows the comparison of deep learning models to detect SARS-CoV-2 from various dataset. Sanksham Jain et al. [14] used a machine learning-based approach to improve the speed of screening SARS_CoV_2 using chest X-rays. The approach achieved an accuracy of 86.93%. In the other work, Yang Yanjun et al. [15] developed a recurrent neural network (RNN) - based deep learning model to detect SARS-CoV-2 RNA in human nasopharyngeal specimens. The
approach achieved an overall accuracy of 98.9%. The artificial neural network in this work achieves an accuracy within the range of the two models, which demonstrates the reliability of using the ANN with Raman spectroscopy to detect SARS-CoV-2.

The applications of machine learning and artificial neural network models were extended in the work of David Chen [16] to assess the improvement of the ensemble methods from baseline machine learning models. Ensemble methods developed in the work achieved 98% classification accuracy to predict the infection status of COVID-19 from Raman spectra of 250 human serum samples [16].

**Table 2.** Comparison of algorithms to detect SARS-CoV-2 from Raman spectroscopy data and the algorithms developed in this work.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBDT [1]</td>
<td>87.097</td>
<td>89.655</td>
<td>84.848</td>
</tr>
<tr>
<td>XGBoost [1]</td>
<td>88.710</td>
<td>84.848</td>
<td>93.103</td>
</tr>
<tr>
<td>PCA+XGBoost [1]</td>
<td>91.935</td>
<td>89.655</td>
<td>93.939</td>
</tr>
<tr>
<td>RFE+XGBBoost [1]</td>
<td>93.548</td>
<td>87.500</td>
<td>96.774</td>
</tr>
<tr>
<td>Stacked subsemble classifier [16]</td>
<td>98.000</td>
<td>98.500</td>
<td>95.550</td>
</tr>
<tr>
<td><strong>This work</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>93.548</td>
<td>100.000</td>
<td>88.235</td>
</tr>
<tr>
<td>SVM</td>
<td>90.323</td>
<td>92.857</td>
<td>88.235</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>87.097</td>
<td>92.857</td>
<td>82.353</td>
</tr>
</tbody>
</table>

**Table 3** Comparison of deep learning models to detect SARS-CoV-2 from different data types.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN [14]</td>
<td>86.930</td>
<td>96.296</td>
<td>91.793</td>
</tr>
<tr>
<td><strong>This work</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>93.548</td>
<td>100.000</td>
<td>93.333</td>
</tr>
</tbody>
</table>

**4 Conclusions**

This paper demonstrates the reliability of machine learning algorithms (artificial neural network, support vector machine, and decision tree) to classify the Raman spectroscopy data as healthy and SARS-CoV-2 infected. The ANN, SVM and decision tree were able to correctly predict the Raman spectral data by 93.55%, 90.32%, and 87.10%, respectively. The standardization technique and Adam optimizer used in the ANN algorithm were able to deal with the distribution of the data set and facilitated smaller changes in the model in response to the errors. This increased the generalizability of the model on the test data to avoid overfitting. As such, the model can be robust. Therefore, the integration of Raman spectroscopy with machine learning has the potential to serve as an alternative diagnostic tool in point-of-care settings and may be used in addition to more traditional methods of disease diagnosis by medical professionals.

The authors acknowledge the Department of Science and Innovation (DSI) for funding this research and the CSIR.
The authors tool in point spectroscopy with machine learning has the potential to serve as an alternative diagnostic. This increased the general with the distribution of the data set and facilitated smaller changes in the model in response to the errors. This increased the generalizability of the model on the test data to avoid overfitting. As such, the model can be robust.

Therefore, the model can be robust.

CoV, which demonstrates the reliability of machine learning algorithms (artificial neural network models) to detect SARS-CoV-2 from Raman spectra of human serum samples. This work demonstrates the reliability of machine learning algorithms (artificial neural network models) to detect SARS-CoV-2 from Raman spectra of human serum samples. Therefore, the model can be robust.

**Comparison of algorithms**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFE+XGBBoost</td>
<td>91.935</td>
<td>89.655</td>
<td>95.550</td>
</tr>
<tr>
<td>XGBoost</td>
<td>88.710</td>
<td>89.655</td>
<td>96.774</td>
</tr>
<tr>
<td>ANN</td>
<td>87.097</td>
<td>89.655</td>
<td>93.939</td>
</tr>
<tr>
<td>RNN</td>
<td>84.848</td>
<td>89.655</td>
<td>93.939</td>
</tr>
<tr>
<td>SVM</td>
<td>82.353</td>
<td>89.655</td>
<td>93.939</td>
</tr>
<tr>
<td>SVM</td>
<td>88.235</td>
<td>89.655</td>
<td>96.774</td>
</tr>
<tr>
<td>SVM</td>
<td>88.235</td>
<td>89.655</td>
<td>96.774</td>
</tr>
</tbody>
</table>

This work achieved 98% accuracy in the work of David Chen. Ensemble methods developed in the work achieved 98% accuracy in the work of David Chen. Therefore, the model can be robust.

**Applications**

The applications of machine learning and artificial neural network models were extended to detect SARS-CoV-2. This work demonstrated the reliability of machine learning algorithms (artificial neural network models) to detect SARS-CoV-2 from Raman spectra of human serum samples. Therefore, the model can be robust.

**References**

14. J. Saksham, M. Kumar, N. Sindhwani, and P. Singh, (IEEE, 2021)
18. E. Bisong and E. Bisong, (Google colaboratory, 2019).
21. J. Yang, and J. Guan, Info., **13**, 475 (2022)