1D CONVOLUTIONAL NEURAL NETWORKS AND MACHINE LEARNING ALGORITHMS FOR SPECTRAL DATA CLASSIFICATION WITH A CASE STUDY FOR COVID-19

Breno Aguiar Krohling
Nature-inspired Computing Lab
LABCIN - UFES
Federal University of Espirito Santo
Vitória, Brazil
brenokrohling@gmail.com

Renato A. Krohling
Nature-inspired Computing Lab
LABCIN - UFES
PPGI - UFES
Federal University of Espirito Santo
Vitória, Brazil
krohling.renato@gmail.com

Abstract

Machine and deep learning algorithms have increasingly been applied to solve problems in various areas of knowledge. Among these areas, Chemometrics has been benefited from the application of these algorithms in spectral data analysis. Commonly, algorithms such as Support Vector Machines and Partial Least Squares are applied to spectral datasets to perform classification and regression tasks. In this paper, we present a 1D convolutional neural networks (1D-CNN) to evaluate the effectiveness on spectral data obtained from spectroscopy. In most cases, the spectrum signals are noisy and present overlap among classes. Firstly, we perform extensive experiments including 1D-CNN compared to machine learning algorithms and standard algorithms used in Chemometrics on spectral data classification for the most known datasets available in the literature. Next, spectral samples of the SARS-COV2 virus, which causes the COVID-19, have recently been collected via spectroscopy was used as a case study. Experimental results indicate superior performance of 1D-CNN over machine learning algorithms and standard algorithms, obtaining an average accuracy of 96.5%, specificity of 98%, and sensitivity of 94%. The promising obtained results indicate the feasibility to use 1D-CNN in automated systems to diagnose COVID-19 and other viral diseases in the future.

Keywords Convolutional Neural Networks · Machine Learning · Spectroscopy · COVID-19.

1 Introduction

The use of spectroscopy has been showing an increasing research interest and practical applications with the advancement of spectral measurement techniques and data analysis studies in the field of chemometrics. In particular, the use of measurements in the infrared (IR) and near and infrared (NIR) spectrum, which provides information that is not visible by simple observation. The use of infrared spectroscopy is considered one of the most important tools in chemometrics (Stuart, 2000), as it has a low cost, is fast, and can analyze samples in their most diverse states, making it also a non-invasive approach. Besides that, the use of machine learning and deep learning, such as artificial neural networks, have been introduced in areas previously dominated by statistical analysis methods and/or classical artificial intelligence algorithms. With the growth of neural network applications in the last few years, such algorithms began to be explored in problems of different natures, including spectroscopy problems.

Data from spectroscopy presents some problems such as high dimensionality, and small datasets. Thus, It usually requires the use of pre processing techniques such as noise filtering and dimensionality reduction for proper classification. However, artificial neural networks (NN) are known to deal well with high dimensionality and have achieved remarkable results for different applications. However, NN need a significant amount of data to work properly. In such cases, it becomes necessary to analyze the trade-off between using standard methods or migrate to new approaches.

Convolutional neural networks can be applied to challenging real-world problems, such as the diagnosis of the Sars-Cov-2 virus. With the advance of the COVID-19 pandemic, testing the population in large scale for sanitary and public health reasons became necessary. Sars-cov2 virus testing presents two main barriers: 1) the price of testing, and 2) time to get the results. While the first affects mostly underdeveloped countries, the second influences containment plans around the world. By testing with information acquired via spectrometers and processing the data through a robust artificial intelligence model, information about the patient's diagnosis could be obtained quickly and at a low cost, making it accessible to public and private health services.

With the advances in machine learning techniques, there is a tendency to migrate from standard algorithms used in spectroscopy such as Support Vector Machines and Partial Least Squares, to recent approaches such as Convolutional Neural Networks (Zhang et al., 2019) and (Yuanyuan and Zhibin, 2018). However, CNNs commonly require a large amount of data to train the model. Although there are public databases (Kaewseekhao et al., 2020), (Chauvergne et al., 2020) and (Zyubin et al., 2020) there is still a lack of spectral datasets in public domain. So, it is necessary to investigate the performance of such algorithms when applied to problems with a small number of instances of training. Furthermore, spectral data present some characteristics, such as overlap of samples at certain wavelengths, which make the classification task difficult, as in the COVID-19 dataset provided by Yin et al. (2021).

The main contribution of this work are twofold:

- we applied a 1D-CNN to 7 available public spectral datasets in order to validate the results as compared to conventional machine learning algorithms.

- we present a case study involving spectral data of COVID-19 and show the feasibility of 1D-CNN.

This reminder of this paper is organized as follows. In section 2, we present related works. In section 3, the methodology involving pre-processing techniques and machine learning algorithms as well as a 1D convolutional neural network are described. The results are presented and a comparative analysis is carried out with discusions in section 4, and in section 5 the article is concluded with directions for future research.

2 Related Works

Sepctroscopy techniques are used in several areas of knowledge and can be applied to approach several real-world problems, from the classification *in vivo* of skin lesions (McIntosh et al., 2001) to determine the quality of agricultural products (Hayati et al., 2020). McIntosh et al. (2001) addressed the task of characterization skin lesions through statistical methods, specifically the use of the paired *t-test.* a classification of carcinogenic lesions was performed using Linear Discriminant Analysis (LDA), showing high accuracy and feasibility for discriminating skin lesions in a non-invasive way.

Gniadecka et al. (2004) combined spectroscopy techniques and Multilayer Perceptron neural networks for early diagnosis of skin cancer. The study aimed to differentiate melanoma, a skin cancer with a higher letal rate, from other skin lesions, such as pigmented nevus and basal cell carcinoma. The influence of chemical changes in cancer tissue and their importance for classification was investigated. The results obtained showed high sensitivity and specificity, important metrics in the detection of melanoma. Xiang et al. (2010) used infrared spectral data to diagnose endometrial cancer. The study contained 77 samples, including carcinogenic material and a control group. Noise reduction techniques were used as the Savitsky-Golay filter and data dimensionality reduction with PCA. Post-processed data were used for training artificial neural networks and the obtained results showed effectiveness for early diagnosis of the disease.

Wu et al. (2012) used spectroscopy data to non-invasively determine the presence of proteins and polysaccharides in powder samples of *coriolus versicolor*, a medicinal mushroom species. The authors compare artificial neural networks applied to original data and data with reduced dimensionality by applying PCA. The neural network was trained with the *Backpropagation* algorithm and presented better results using data with reduced dimensionality. Liu et al. (2017) used Deep Auto-Encoder (DAE) to extract and reduce features from the infrared spectral data. To evaluate its effectiveness, the algorithm was applied to the Cigarettes dataset and compared with methods established in the literature, such as PCA and in turn, the KNN algorithm was applied. The results showed that the model using DAE was able to extract features with better differentiation capacity as the concurrent approaches.

Peng et al. (2018) proposed an algorithm for extracting spectral information in the frequency domain. Principal components were determined by the entropic contribution of each component together with genetic algorithm. The selected components were used to build a regression model with Partial Least Square (PLS), showing a performance improvement when compared to the results obtained with the original data.

Cui and Fearn (2018) performed tests on three spectroscopy datasets to demonstrate the effectiveness of CNNs in relation to the PLS Regression method. The results obtained showed that the convolution layer of a CNN can act as a preprocessor of the spectrum, not requiring the application of signal processing techniques. In addition, results demonstrated the superiority of CNN in terms of the metrics used to validate the model, even using datasets with a low number of samples.

Yuanyuan and Zhibin (2018) applied an ensemble of CNNs to the Corn, Gasoline, and Mixed Gases datasets and the approach was compared with classical methods in the spectroscopy field, such as PLS, as well as neural networks trained with Backpropagation. Results showed that the proposed model obtained superior performance in all case studies investigated, showing its relevance for spectrum classification.

Lima et al. (2019) investigated the discrimination between non-melanoma cancer (basal cell carcinoma and squamous cell carcinoma) and actinic keratosis and normal skin. The study used spectral data collected from in-vivo and ex-vivo tissues. The t-test was used to find regions of the spectrum with significant differences between samples. The selected regions were used to discriminate the lesions using Euclidean and Mahalanobis distance. Results obtained show robustness when using ex-vivo tissue data.

Seoni et al. (2019) investigated hemoglobin concentration using infrared spectrum data on normal skin tissue and skin tissue with actinic keratosis, a benign skin lesion that, without treatment, can evolve into a cancerous lesion. The MANOVA statistical analysis was applied to assess possible differences between the healthy skin group and the one containing actinic keratosis. The results showed that, in general, it was possible to differentiate the two groups through the statistical approach.

Zhang et al. (2019) introduced a new approach based on a CNN with 1D convolutions and the adoption of Inception structure in two convolution layers. The method was evaluated in four public spectroscopy datasets: Corn, Tablet, Wheat, and Soil. The results obtained were superior to other CNN architectures and linear methods such as PLS and Support Vector Regression (SVR). In addition, pre-processing methods applied to the raw spectrum signal have also been applied in order to investigate the influence on the prediction results. The authors concluded that signal noise reduction is a necessary step to improve the performance of the algorithms.

Chen and Wang (2019) evaluated the ability to extract features from a CNN without applying dimensionality reduction on the spectrum. Experiments were carried out using the dataset Corn and compared with traditional chemometric methods such as PLS and neural network training algorithms such as Extreme Learning Machine (ELM) and Backpropagation. The results obtained in the tests performed show that: 1) If the methods/algorithms use the original spectrum as input, the CNN obtains better performance; and 2) If standard methods/algorithms are combined with dimensionality reduction algorithms and variable selection, the results obtained are similar to those obtained by CNN, with no statistical difference between the methods.

Chen et al. (2020) proposed the use of ensemble of artificial neural networks trained with the ELM algorithm. The proposed algorithm uses random initialization of weights of the

hidden layer to obtain diversity among models. The method was tested in 3 datasets, Tecator, Shootcut_2002, and Tablet. The approach was compared with PLS, and was superior as compared to those achieved by classic chemometric algorithm. Yan et al. (2020) collected infrared spectrum data in the hydrolysis process of materials used in traditional Chinese medicine. Regression models were developed using CNN and PLS with real-time data. The proposed CNN architecture presented better results over PLS, but with higher processing cost and computational time. Therefore, for real-time applications, it is necessary to analyze the maximum latency of the interval between results to decide which method should be used.

3 Methodology

3.1 Data pre-processing

Spectroscopy is a technique based on the vibrations of atoms in a molecule. The spectrum can be analyzed and classified by its wavelength, such as ultraviolet, visible, and infrared. Each wavelength provides unique information for analyzing a sample, so it is necessary to know the best wavelength range to use when collecting data. For example, the spectrum obtained at the infrared wavelength is widely used in applications that need to identify organic components of a sample (Stuart, 2000). The data generated by this technique, known as a spectrum, is obtained by emitting an amount of radiation onto a sample and determining which part of the radiation is absorbed or reflected by the sample at a given energy level (Stuart, 2000).

Data collected through spectrometers can suffer different types of interference, from low-quality equipment to external interference in uncontrolled environments, such as light scaterring, and noisy data, among others. Noisy components of the the spectrum signal is a challenging issue in spectral analysis, as the instability of the collected signal can lead to wrong conclusions. This instability can be reduced by performing more than one reading of the sample, obtaining the average spectra, but without eliminating all the existing noise. As a result, signal processing techniques become a crucial step in spectral analysis to remove, or reduce undesired noisy components of the spectral signal.

Selecting an appropriate technique to perform data pre-processing can directly influence the model performance and data interpretability (Lee et al., 2017). The most used pre-processing techniques can be divided into two categories: 1) Scatter-correction, and 2) Spectral derivatives (Rinnan et al., 2009). Scatter-correction algorithms are used to reduce the variability between samples. Spectral derivates algorithms use convolution operations, usually by fitting a polynomial to the points of the spectrum.

In this work, we use one of the most standard pre-processing filter in spectral analysis, the Savitzky-Golay (SG) filter (Savitzky and Golay, 1964), and evaluate its influence on the classification results. Savitzky and Golay (1964) proposed an algorithm to smooth data based on polynomial interpolation. They demonstrated that fitting a polynomial to a set of points in the spectrum and then evaluating it at the subsequent point respecting the point approximation interval would be equivalent to more complex convolution operations. Furthermore, they showed that the algorithm can remove noise from the data and keep

the shape and peaks in the spectrum wavelengths, preserving original characteristics in the smoothed data.

3.2 Machine Learning

One of the fastest growing areas in computing is Artificial Intelligence. Its applications are present in medicine, engineering, finance, among other areas (Pannu, 2015). One of the sub-areas of Artificial Intelligence is machine learning, which consists of algorithms acquiring knowledge from already known data and, subsequently, being able to infer solutions for new instances of the problem without being explicitly programmed for it (Alpaydin, 2020). Supervised Learning problems can be considered the most common among the existing ones. The algorithms look for patterns in order to learn how to associate an input $\mathbf{x}$ to its respective output y. In general, supervised learning algorithms learn through examples how to classify an output y given an input $\mathbf{x}$ by estimating $p(y|\mathbf{x})$ (Goodfellow et al., 2016). The machine learning algorithms used in this work are Partial Least Square - Discriminant Analysis, Support Vector Machine, k-Nearest Neighbors (Alpaydin, 2020), and Many-objective clustering based on Hill Climbing and minimum spanning tree (Esgario and Krohling, 2018).

3.3 1D Convolutional Neural Networks

Artificial neural networks (ANN) are structures composed of simple computing units, called artificial neurons, interconnected with each other (Schalkoff, 1997). The connections between neurons form complex layers for computing and propagating information. The algorithm for training ANN's is known as *Backpropagation*. Through it, it is possible to train ANNs for classification and regression/prediction tasks. Standard ANNs are characterized by three main layers, the data input layer, the hidden layers, and the output layer.

Convolutional neural networks (CNN), initially proposed by LeCun et al. (1989), are specialized networks for applications that present data in grid format such as images and temporal series, as they are able to capture spatial and temporal features (Goodfellow et al., 2016). The architecture gained prominence in 2012 when a CNN-based model won the LSVRC-12 competition in the ImageNet database (Krizhevsky et al., 2012). The main difference between CNN's and traditional ANN's is the presence of convolution and pooling layers in their architecture. CNNs use convolution operations instead of matrix operations in at least one of their layers (Goodfellow et al., 2016). The main components of a CNN and its work principle during the training process are described in the following.

Convolution Layer: Convolutions are linear operations on two functions, that generate a third one. In CNN's, convolutions are applied to extract features, where the first function of a convolution is represented by the input data of the network, and the second is represented by a *kernel*, a sliding filter used to extract features. By applying the convolution operation between these two functions, it is generated feature maps, which contain relevant information from the original input data. CNN's can have layers of convolutions of one, two, or three dimensions, depending on the problem domain and data characteristics. For example, image recognition applications use 2D convolutions, while applications whose data comes from time-series use 1D convolution.

After convolution operations, feature maps are applied to an activation function. The activation function aims to add non-linearity to the model (Goodfellow et al., 2016). When choosing an activation function, some characteristics must be observed, such as: The computational cost must be small, since the operation is performed frequently. Due to the training process of a neural network is based on gradient descent, the activation function must also be differentiable. The Rectified Linear Unit (ReLU) activation function is commonly used in neural networks because it presents such characteristics. The function returns zero for all negative values and the value itself for positive values. It is described by $ReLU(x) = max(0, x)$.

Pooling Layer: The pooling layer, normally applied after the convolution operation, consists of reducing the dimensionality of the feature map in order to eliminate redundant information and keep the main features (Goodfellow et al., 2016). Similar to convolution, a sliding window moves over the feature map generating a new output. However, pooling operations are deterministic, typically calculating the maximum or average value of the elements in the sliding window (Zhang et al., 2020). When we apply the MaxPooling technique, the maximum value present in the analyzed data group is propagated to the next layers, while the AvgPooling technique propagates the average value of the data.

Fully connected layer: The combination of convolution and pooling layers alone is not enough to perform the classification/regression task, its main function in the architecture is to extract features from the given data. The extracted features, therefore, need a classifier to complete the task, which we call a fully connected layer. Typically, the fully connected layer of a CNN is represented by an ANN known as a Multilayer Perceptron (MLP) (Zhang et al., 2020), a feedforward neural network composed of an input layer, a set of hidden layers, and the output layer.

A MLP maps a $f : x \to y$ function, where x are the features extracted by the convolution and pooling layers, and y is the desired network output. This process, known as training an ANN, is an optimization problem that seeks to minimize a loss function (Goodfellow et al., 2016). The loss function calculates the error between the desired outputs and those obtained by the neural network. The error obtained is used to recalculate the weights of the network (Zhang et al., 2020). The Cross-Entropy loss function commonly used by CNN's in classification problems is described by:

$$CrossEntropy(pa, pb) = -\sum_{i}^{k} pa_i \log pb_i \tag{1}$$

where pa_i and pb_i represent the real and predicted probability for the class i, respectively and k represents the number of classes. For unbalanced datasets, loss functions that consider the frequency of each class in its calculation, such as the Weighted Cross Entropy are often used.

Along with the parameter adjustment algorithm, weight regularization techniques can be included in order to avoid problems such as overfitting, when a model presents good results in the training process but it is not able to work properly to unknown data. In this work, the dropout regularization technique was applied. This technique consists of removing some of the neuron connections with a certain probability, in order to improve the network's generalization to unknown data (Zhang et al., 2020).

By combining multiple convolution and pooling layers with a fully connected layer is generated CNN-like architectures. Figure 1 shows the 1D-CNN architecture used to classify spectral data in this work.
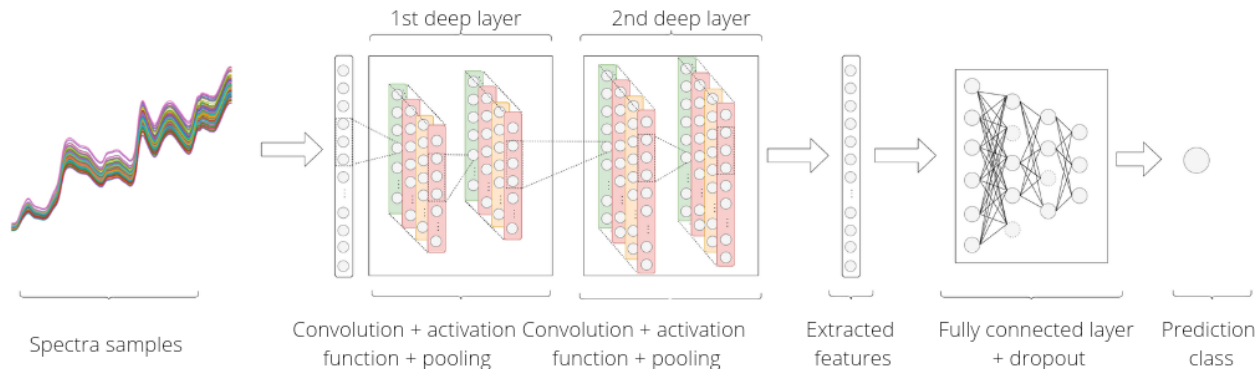

Figure 1: 1D-CNN architecture made up of convolution, pooling, and fully connected layers.

4   Experiments and Results

This section presents the results with discussions. First, the datasets used in the experiments are presented. Second, the algorithms are applied to the datasets in order to compare with machine learning and standard algorithms of chemometrics. Finally, a case study with spectral data from samples of the Sars-Cov-2 virus is presented.

4.1   Datasets

Next, we present a brief description of the datasets. Kosmowski and Worku (2018) collected infrared spectral data from barley, chickpea, and sorghum crops. Each type of grain originated a dataset. The Barley dataset is composed of 1200 barley grain samples, distributed in 24 classes of grain variants. The Chickpea dataset has 950 chickpea samples distributed in 19 classes, and finally, the Sorghum dataset has 500 sorghum samples distributed in 10 classes. Despite a large number of samples in the dataset, it is clear that they are distributed in a large number of classes. Consequently, there is a small frequency of samples per class. All three datasets are well balanced. Zheng et al. (2019) presented an approach that uses Extreme Learning Machine for spectra classification and made the data available. Four datasets were used in this study, Coffee, Meat, Oil, and Fruits. The Coffee dataset presents samples divided

Dataset	Samples	N. classes
Barley	1200	24
Chickpea	950	19
Sorghum	500	10
Coffee	56	2
Meat	120	3
Oil	120	4
Fruits	983	2

Table 1: Number of samples and classes present in each dataset.

into Arabica and Conilon coffee beans uniformly distributed in 56 samples. The Meat dataset is made up of 60 samples of chicken, pork, and turkey, also evenly distributed. The Oil dataset contains 60 extra virgin olive oil spectrum samples from 4 different countries: Greece, Italy, Portugal, and Spain. The Fruits dataset consists of 983 samples of fruit purees divided into two classes, strawberry, and non-strawberry. Table 1 lists the number of samples and classes present in each dataset.

4.2 Experimental Setting

The training of 1-D CNN and machine learning algorithms investigated in this work were performed using cross-validation. This method is used to decrease the bias and avoid over-estimated models. The learning process consists of dividing the dataset into k *folds* and use each one separately as a test set of a model. At the end, one calculates the average accuracy and standard deviation of the algorithm for the dataset. The process of training and evaluating an algorithm using *cross-validation* is described in the Algorithm 1. For all experiments a number of $k = 5$ folds was adopted, in other words, the dataset was divided into 80% for training and validation and 20% was used exclusively for testing.

Algorithm 1: Cross-validation evaluation

---
1 - Shuffle the samples.
2 - Split it into k folds.
for *f in folds* **do**
    1 - Keep f as a test set.
    2 - Train a new model with the remaining folds
    3 - Evaluate the model with *fold f*.
    4 - Save the model accuracy.
end
3 - Get the cross-validation mean accuracy and standard deviation.
4 - Return the model within the closest accuracy to the cross-validation mean accuracy.

The algorithms PLS-DA, KNN, SVM, and CNN need hyperparameter adjustments. So, one can select the best configuration for each dataset and maximize its performance.

The PLS-DA algorithm is based on components construction in order to maximize the variance of the variables. Each component explains the degree of variance it represents. For

each dataset, c components were selected that, together, represent 95% of the total variance of the data.

For the KNN and SVM algorithms, a variation of the algorithm 1 was used, in order to find the best set of hyperparameters for each cross-validation round. The data used to train the model undergoes a new division into a training set and validation set, in order to find the best combination of hyperparameters. Thus, for each combination of training and testing folds, one finds a set of optimal hyperparameters for the model. At the end of the process, one calculates the mean and standard deviation obtained by the algorithm and the set of hyperparameters that generated the model with the closest accuracy to the average accuracy, avoiding an overestimated model. The process for finding the best hyperparameters and average accuracy is described in the algorithm 2.

Table 2 presents the search space used for the KNN and SVM algorithms. Table 3 presents the number of components c needed to obtain 95% of the data variance in the PLS-DA algorithm, in addition to the amount of k-neighbors and the kernel function that provided the closest result to the average result obtained when executing the algorithm 2.

Algorithm	Hyperparameter	Search Space
KNN	k-neighbors	[2, 3, 5, 10, 15, 20, 24]
SVM	Kernel Function	[Linear, Polinomial, Sigmoid, RBF]

Table 2: Search space for KNN and SVM hyperparameters.

Algorithm 2: Cross-validation evaluation - variation to find best set of hyperparameters

1 - Shuffle the samples.
2 - Split it into k folds.
for *f in folds* **do**
    1 - Keep f as a test set.
    for *each set λ of hyperparameters* **do**
        for *v in (folds - f)* **do**
            1 - Keep v as a validation set.
            2 - Train a new model with the remaining folds.
            3 - Evaluate the model with fold v.
            4 - Save the model accuracy.
        end
        1 - Save the mean accuracy and standard deviation for the λ set.
    end
    2 - Train a new model with the remaining folds with the best λ set.
    3 - Evaluate the model with fold f.
    4 - Save the accuracy and the model
end
3 - Get the cross-validation mean accuracy and standard deviation.
4 - Return the model and λ set within the closest accuracy to the cross-validation mean accuracy.

Datasets	N. of components	N. k-neighbors	Kernel Function
Barley	2	15	Sigmoid
Chickpea	2	10	Sigmoid
Sorghum	2	10	Sigmoid
Coffee	2	3	RBF
Meat	3	3	RBF
Oil	6	5	RBF
Fruits	4	3	RBF

Table 3: Hyperparameters that obtained the best results for the PLS-DA, KNN, and SVM.

Finally, the 1D-CNN, as a recent approach in the field of Chemometrics, it does not have pre-defined and pre-trained architectures for applications that involve spectral data classification. To choose the architecture and hyperparameters used in this work, the software Optuna[1] was used, a hyperparameter optimizer developed to be used in machine learning problems that search for combinations of hyperparameters and architectures that maximize model performance. Similar to the algorithm 2, it searches for the best model in the training and validation dataset and, after finding a set of suboptimal hyperparameters, tests the model in the test set.

Table 4 lists the selected hyperparameters for each dataset involving the training and testing step. All experiments and algorithms were programmed using utilities from the scikit-learn[2] and PyTorch[3] libraries from Python[4]. The source code in the experiments is available from authors upon request.

Hyperparameters	Dataset						
	Barley	Chickpea	Sorghum	Coffee	Meat	Oil	Fuit
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.0001	0.001	0.001	0.001
Epochs	5000	5000	5000	100	700	1000	1000
Dropout Rate	0.4	0.4	0.3	0.1	0.1	0.2	0.1
Batch Size	50	50	50	6	10	10	50
Loss Function	Cross entropy	Cross entropy	Cross entropy	Cross entropy	Cross entropy	Cross entropy	Cross entropy

Table 4: Hyperparameters that obtained the best results for 1D-CNN.

---

[1]https://optuna.readthedocs.io/

[2]https://scikit-learn.org/

[3]https://pytorch.org/

[4]https: //www.python.org/

4.3   Experimental Results

This subsection presents the results obtained by each algorithm when applied to the datasets in this work. The experiments were split into two parts: 1) Using original spectral data, and 2) Applying the Savitzky–Golay (SG) filter as a pre-processing step. The division of the experiments into two stages seeks to analyze the influence of data pre-processing on the model results. All experiments were performed on a machine with Windows 10, equipped with an Intel Core I5-8265U processor and 8GB of RAM memory.

Original Data: The results obtained for the first configuration of experiments, data without pre-processing, are presented in table 5. The investigated algorithms (MOCHM, PLS-DA, SVM, KNN and, 1D-CNN) were applied to each dataset. Best results are shown in bold.

Dataset	MOCHM	PLS-DA	SVM	KNN	1D-CNN
Barley	-	10.66 ± 0.62	7.75 ± 0.54	16.08 ± 1.38	$\mathbf{47.4 \pm 4.7}$
Chickpea	-	14.73 ± 1.12	23.58 ± 1.07	27.16 ± 4.48	$\mathbf{54.2 \pm 3.3}$
Sorghum	-	18.79 ± 3.31	15.79 ± 1.94	34.79 ± 2.99	$\mathbf{55.8 \pm 3.3}$
Coffee	51 ± 2.1	85.7 ± 3.4	42.7 ± 3.4	46.3 ± 1.37	$\mathbf{100 \pm 0}$
Meat	36 ± 0.2	70.83 ± 6.56	41.6 ± 3.1	87.5 ± 7	$\mathbf{95 \pm 3.3}$
Oil	44 ± 0.7	75 ± 2.04	41.6 ± 1.02	73.3 ± 3.11	$\mathbf{88.8 \pm 5.3}$
Fruit	95 ± 1.9	91.04 ± 1.02	88.3 ± 2.96	88.9 ± 1.75	$\mathbf{96 \pm 0.6}$

Table 5: Average accuracy obtained by the algorithms when applied to data without pre-processing.

The obtained results with the 1D-CNN were superior for all datasets using data without pre-processing. It is observed, however, that the datasets *Barley*, *Chickpea* and *Sorghum* achieved an average accuracy lower than other datasets. It turns out that due to the high number of classes present in the data and the small number of samples available for each class in the training step, the model was not able to generalize effectively. Nevertheless, we notice a difference when comparing the results obtained by 1D-CNN with those obtained by the standard algorithms used in Chemometrics, i.e., PLS-DA, and SVM. The KNN algorithm also obtained better results than the PLS-DA and SVM algorithms, reaching an average accuracy of 20% for the *Sorghum* dataset compared to SVM, the worst algorithm performance at the dataset. In the other datasets, the PLS-DA algorithm obtained, in general, the best results after 1D-CNN. The MOCHM algorithm was stoped before finishing its execution for the *Barley*, *Sorghum* and *Chickpea* datasets because it demanded a computation time over 60 minutes to finish the process.

Pre-processed data: The second configuration of experiments was performed with the same algorithms and datasets, but applying the SG filter in the pre-processing step. The SG filter was performed with a standard hyperparameter configuration: the window size was set to 11, using a polynomial of degree 3 and second-order derivative. In the table 6 the results obtained in the experiment are presented.

We can observe that the 1D-CNN did not obtain the hegemony of the results as reported in the previous case. For the *Barley*, *Chickpea* and *Sorghum* datasets, KNN was superior

Dataset	MOCHM	PLS-DA	SVM	KNN	1D-CNN
Barley	-	11.83 ± 2.77	14.25 ± 2.3	**64.91 ± 1.3**	46.16 ± 0.97
Chickpea	-	15.78 ± 1.07	20.4 ± 4.28	**71.36 ± 2.11**	55.2 ± 1.1
Sorghum	-	25.99 ± 2.63	28.38 ± 5.87	**71.39 ± 3.3**	59.6 ± 2.1
Coffee	83 ± 1.07	87.4 ± 2.1	92.69 ± 6.1	80.11 ± 4.9	**100 ± 0**
Meat	81 ± 0.8	71.66 ± 4.24	58.33 ± 3.1	92.49 ± 4.08	**96.6 ± 2.07**
Oil	**91 ± 1.1**	75.8 ± 3.11	41.6 ± 1.02	79.1 ± 1.16	89.4 ± 1.12
Fruit	92 ± 2.2	79.2 ± 1.02	93.49 ± 0.86	91.86 ± 0.38	**94.2 ± 0.3**

Table 6: Average accuracy obtained by the algorithms when applied to data with pre-processing.

to CNN. For the other datasets, CNN maintained its superiority except for the *Oil* dataset, where the MOCHM obtained the best accuracy.

In general, the performance gain of the algorithms when using pre-processed data is noticeable. We can evaluate the gain from two aspects: 1) The increase in the average accuracy of the algorithms for the datasets, with emphasis on the MOCHM and KNN algorithms, and 2) The reduction in standard deviation, as observed on 1D-CNN.

The MOCHM achieved a gain of approximately 50% in accuracy in the *Oil* dataset, achieving better overall performance. KNN also achieved significant gains in accuracy surpassing CNN in the *Barley*, *Chickpea* and *Sorghum* datasets. Regarding the decrease in standard deviation, CNN presents this characteristic for all datasets. Despite not presenting a high gain in accuracy as in other algorithms, the reduction in the standard deviation indicates greater stability, generating models with a higher degree of reliability. This fact is explicit in the results obtained by the 1D-CNN in the *Barley* dataset, where it reached an accuracy of 47.4 ± 4.7 in the experiments without preprocessing against 46.16 ± 0.97 in the experiments with preprocessing.

Despite the superior results obtained by 1D-CNN in terms of accuracy for the experiments performed, it is necessary to analyze the trade-off in the results. As 1D-CNN is an algorithm which demands a higher computational cost compared to other algorithms, then it is necessary to analyze in which cases its use is preferred. Table 7 presents the training time required for a model of each algorithm per dataset. It is clear that the larger the dataset, the longer the time needed for training the 1D-CNN. Furthermore, it was the only algorithm that needs, in most cases, minutes to train the model. In problems with a large number of samples and high dimensionality, the efficient use of pre-processing algorithms combined with simple algorithms, such as KNN, might be a viable alternative instead of using 1D-CNN.

Data visualization: In order to get a 2D visualization of the data pre-processing influence, the t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) was applied to the cases without and with pre-processing.

Figure 2 shows the original data and with pre-processing for the *Coffee* dataset. It is not possible to observe in the original data clear groups that distinguish the two classes present in the dataset so that a simple classifier had difficulties to correctly separate the samples through

Dataset	MOCHM	PLS-DA	SVM	KNN	1D-CNN
Barley	-	0.1328	1.7173	0.3353	603.12
Chickpea	-	0.0855	0.8775	0.2436	485.30
Sorghum	-	0.0255	0.2826	0.1020	443.05
Coffee	22.10	0.1196	0.0083	0.0795	4.22
Meat	18.30	0.0377	0.0136	0.0277	45.27
Oil	31.55	0.0152	0.0194	0.0231	106.07
Fruit	32.26	0.0377	0.3920	0.2325	81.13

Table 7: Model Training time for each algorithm in seconds.

a linear function. However, looking at the pre-processed data, we see the data grouped in a
concise way, showing that the SG filter helps the model to discriminate the data.

(a)   (b)
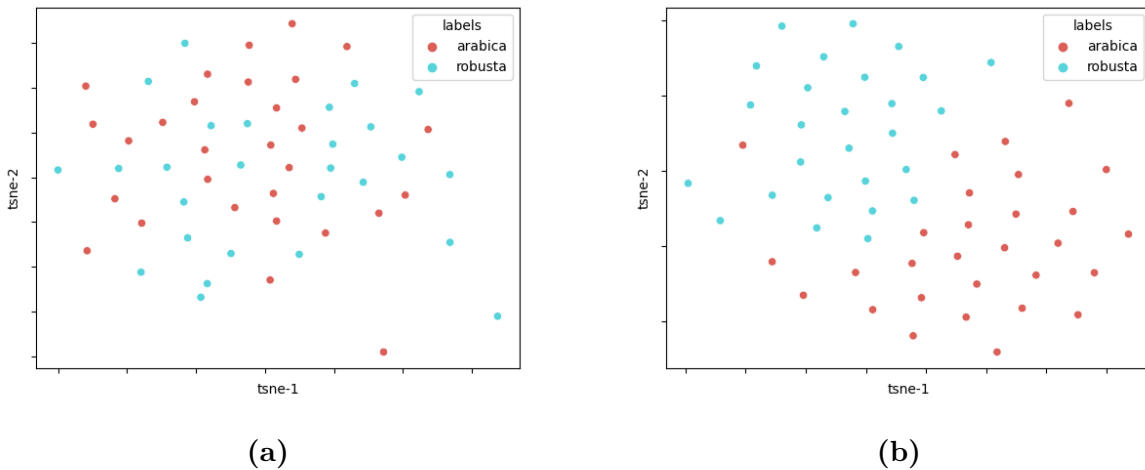
Figure 2: Visualization of the spatial arrangement of (a) original data and (b) data after
pre-processing using t-SNE for the Coffee dataset.

The results show the importance of applying pre-processing algorithms to spectral data
for noise reduction and how this positively impacts the results obtained by each algorithm,
whether in terms of accuracy gain or model stability. Convolutional Neural Networks are
promising in the analysis of spectral data, since the results obtained indicate superior
performance over traditional algorithms. It is worth to mention the ability of 1D-CNN to
generalize well for small datasets. Furthermore, algorithms such as KNN prove to be a
viable and simple alternative to use in most cases when combined with a data pre-processing
algorithm, since it obtains good results and does not demand high computational costs.
Based on the results obtained, we extend our work to a case study using 1D-CNN, PLS-DA,
and KNN in the next section, since these algorithms obtained the best average performance
in the experiments previously investigated.

4.4 Covid-19 case study

The Sars-Cov-2 virus (COVID-19) pandemic has emerged as the biggest and challenging health problem facing the world in recent decades. With the first case confirmed in December, 2019 in Wuhan District, China (Zhu et al., 2020), the virus has spread rapidly in China and to other countries, and in January 2020 the World Health Organization (WHO) declared a state of global emergency (OMS, 2020).

Due to its rapid transmission characteristics and high transmission rate, in a few months since the first confirmed case, the virus has been reported in 144 countries around the world (OMS, 2020). In order to carry out sanitary restrictions, assess and report on the spread of the virus, mass testing was started in several territories that had cases of COVID-19. However, the available methods for testing have undesirable characteristics that can influence the decisions taken to contain the virus. First, the high cost of available tests, causing the unavailability of testing in underdeveloped countries. Second, the time needed to obtain definitive results from the presence of the virus in the patient. Third, the high rate of false-negative tests. Among the mentioned problems, the high rate of false negatives is a crucial factor that has a direct impact on the virus containment. Individuals tested as false negatives can relax the sanitary measures imposed by healthy organizations and increase the spread of the virus (West et al., 2020). Recently, there is a growing research interest to automated diagnose of COVID-19 using artificial intelligence (Khana et al., 2021).

In this context, automated diagnostic applications using data from spectrometers may become an alternative to traditional tests. Since spectrometers are portable and easy-to-use device, the data collection performed by such devices in conjunction with computational classification models might provide a quick diagnosis in minutes. Such characteristics would affect positively the impact of two of the undesirable characteristics on current testing, high cost, and time required for the result. Finally, in order to reduce false-negative results, the third problem found in current testing, two additional metrics will be included in the analysis of results along with accuracy (ACC) i.e., specificity (ESPEC) and sensitivity (SE), common metrics in biomedical applications.

The specificity is calculated by:

$$ESPEC = \frac{TN}{TN + FP} \tag{2}$$

where TN represents the true negative values, i.e, the samples that belonged to the control group and the classifier classified them correctly, and FP the false-positive values, representing the control group samples that were classified positively for Covid-19.

The sensitivity metric (also known as recall) is calculated by:

$$SE = \frac{TP}{TP + FN} \tag{3}$$

where TP represents the true-positive values, i.e, the virus samples that were correctly classified, and FN represents the false negative values, i.e, the samples that belonged to the Covid-19 group but were classified as a control group.

By using such metrics, we can answer two important questions:

- Among those samples labeled as a control group, in this case, samples that did not carry COVID-19, how many did the model correctly classify?

- For all the samples that are positive for COVID-19, how many did the model classify as carrying the virus?

Especially in case of COVID-19, due to its high dispersion and lethality characteristics, we are interested in answering the second question, as this information is directly linked to the sensitivity metric. By obtaining a high sensitivity rate in the model, we are ensuring that cases in which COVID-19 is present in the sample are classified as positive for the virus, i.e, it reduces the number of false-negative diagnoses, a current problem faced in the test protocols used for detecting the virus.

Yin et al. (2021) collected spectral samples from the control group and from patients infected with COVID-19 and made them available for public use. The dataset provided includes 309 samples, of which 150 belong to the control group and 159 to the group of patients infected by the virus. The collected spectra is shown in Figure 3.

Table 8 presents accuracy, specificity, and sensitivity results obtained for each algorithm when applied to the *Covid* dataset (Yin et al., 2021), with original data and pre-processed by the SG filter. Similar to the results obtained for other datasets reported in the previous section, the 1D-CNN obtained the best results with an average accuracy of 96.5 ± 1.1 when using the SG filter in the pre-processing step. Furthermore, it is important to highlight the high values of specificity and sensitivity, with an average of 98.06 ± 1.86 and 94.06 ± 2.43, respectively. As explained before, low sensitivity values indicate a high number of false negatives, considered the worst scenario in the diagnosis of COVID-19. Such behavior can be observed in the results obtained by the KNN, which obtained, on average, a rate higher than 25% of false negatives for data without pre-processing and 12% for pre-processed data. This value indicates the infeasibility of the algorithm for real applications in the context of COVID-19. The PLS-DA algorithm presented, in general, promising results for all the analyzed metrics, however inferior to those presented by the 1D-CNN.

COVID-19	ACC	ESPEC	SE
1D-CNN	94.19 ± 1.6	96.82 ± 2.3	94.72 ± 3.39
SG + 1D-CNN	$\mathbf{96.5 \pm 1.1}$	$\mathbf{98.06 \pm 1.86}$	$\mathbf{94.06 \pm 2.43}$
PLS-DA	93.7 ± 1.81	92.85 ± 1.8	90.3 ± 2.1
SG + PLS-DA	94.17 ± 0.7	95.12 ± 1.15	92.1 ± 1.9
KNN	83.17 ± 4.84	92.78 ± 7.1	73.59 ± 10.1
SG + KNN	93.2 ± 1.58	97.23 ± 0.9	88.71 ± 1.1

Table 8: Results obtained for the Covid dataset.

Figure 4 shows the confusion matrices for each algorithm using pre-processing. Each matrix is equivalent to the model that obtained the closest accuracy to the average accuracy obtained through the *cross-validation* technique. The confusion matrix referring to 1D-CNN presents
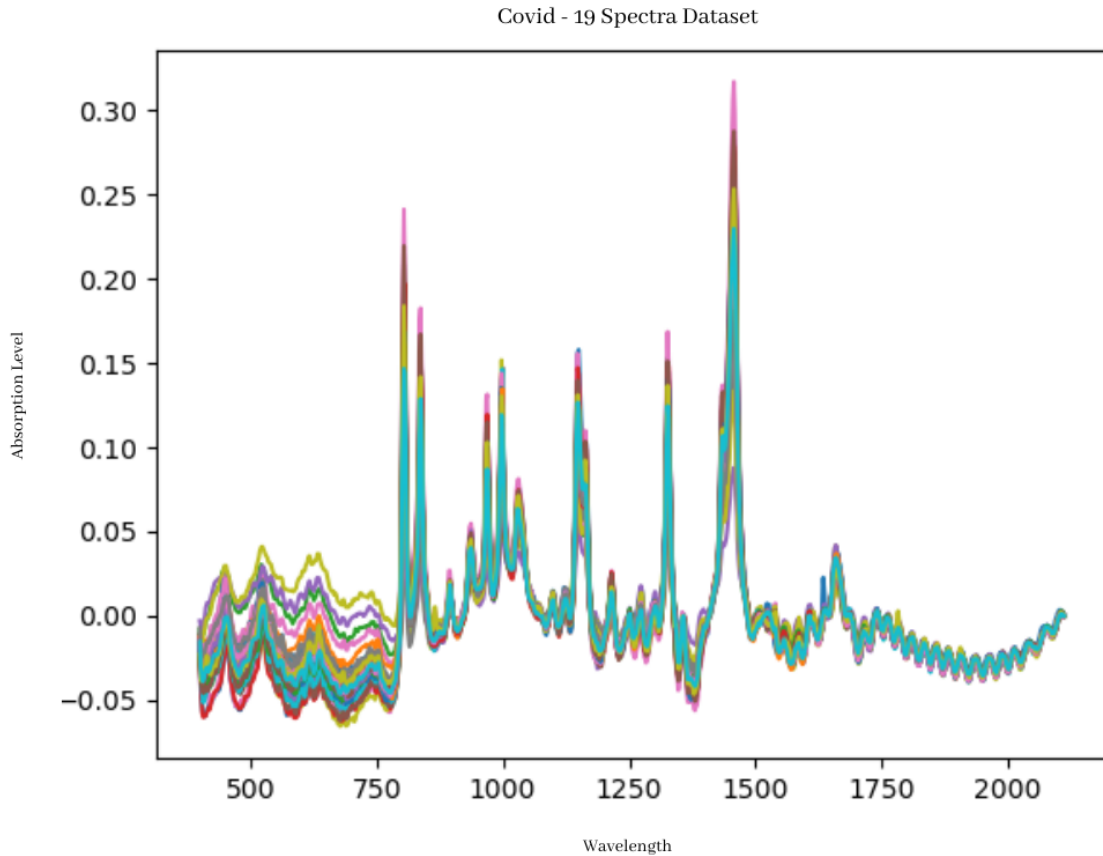
Figure 3: Absorption levels for each wavelength in nanometers for the spectrum collected from the Covid-19 dataset (Yin et al., 2021).

a high hit rate and only two false-negative, which confirms the results presented previously. The KNN confusion matrix presents a high amount of false negatives, in agreement with the results presented in table 8.

The spatial arrangement of the data is shown in Figure 5. Figure 5a shows the visualization of the original data, Figure 5b the visualization of the data after pre-processing, and in Figures 5c and 5d the features extracted by 1D-CNN for original and pre-processed data, respectively. Figure 5a shows that the original data presents overlapping and are not well grouped between the two classes. In Figure 5b, although the application of the SG filter

(a) SG + 1D-CNN - Confusion matrix

(b) SG + PLS-DA - - Confusion matrix

(c) SG + KNN - - Confusion matrix
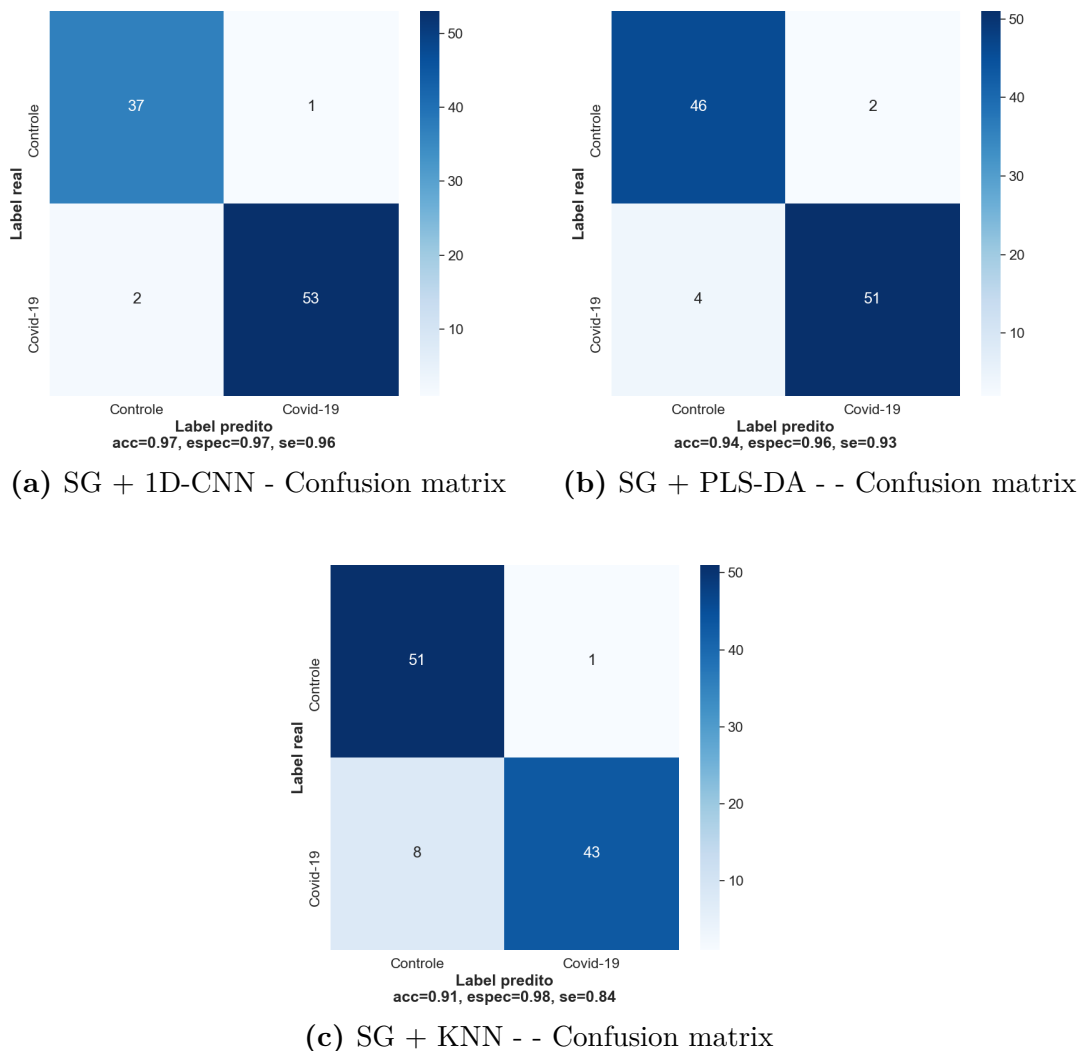
Figure 4: Confusion matrices obtained by the models.

improves the separation of the data into better defined groups, there is still an overlap of data in one of the groups, which makes the classification task difficult for algorithms that do not have feature extraction mechanisms like KNN. This problem can be solved by feature extractors, present on 1D-CNN. The visualization of the features extracted by 1D-CNN can be observed in Figures 5c and 5d, where we notice a better arrangement of groups than in those in Figures 5a and 5b. Especially, in Figure 5c, we note that 1D-CNN was able to group the data even without using the pre-processing technique.

Recent studies have applied spectroscopy techniques in conjunction with classification algorithms to analyze data collected from COVID-19 samples. Yin et al. (2021) applied the SVM algorithm to their dataset to perform the classification between the control group and the virus-infected group. Barauna et al. (2020) used ATR-FTIR spectroscopy to collect data from COVID-19 samples. In their study, it was used 111 samples for the control group and 70 positives for COVID-19. The authors applied a Genetic Algorithm with Linear
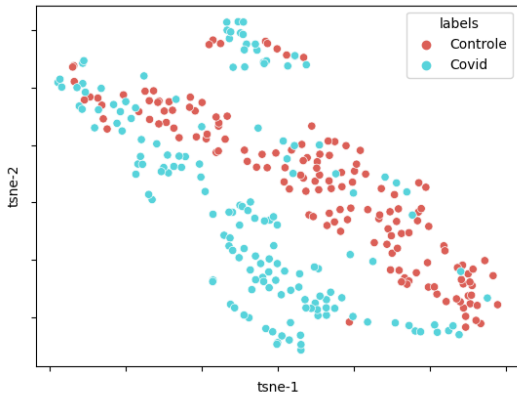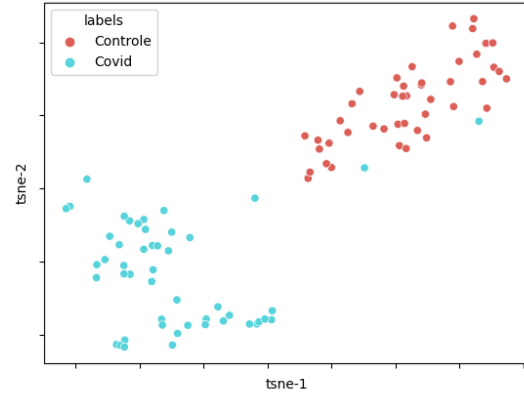
(a) Spatial arrangement of original data.
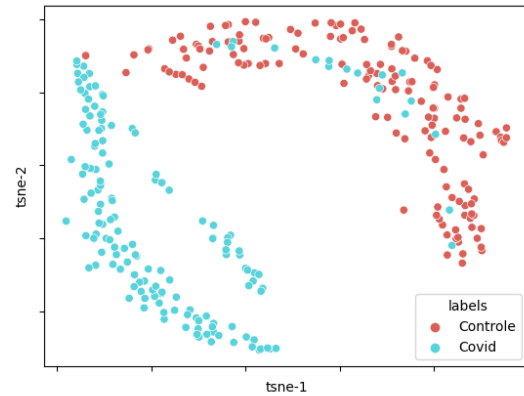
(b) Spatial arrangement of pre-processed data.
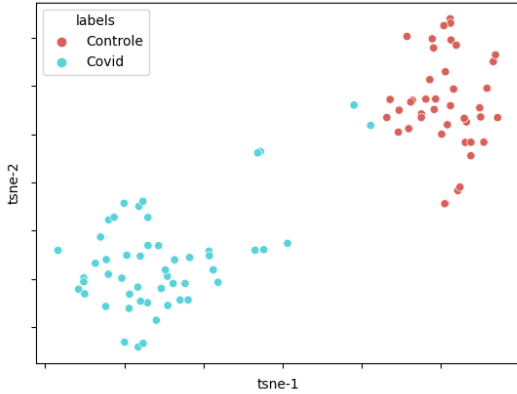
(c) Spatial arrangement of features extracted by 1D-CNN from original data.

(d) Spatial arrangement of features extracted by 1D-CNN from pre-processed data.

Figure 5: Data spatial arrangement visualization using t-SNE for the Covid-19 dataset.

Discriminant Analysis (GA-LDA) to the data pre-processed by the SG filter to perform the classification task. Carlomagno et al. (2021) also proposed the use of CNN-type networks for the classification of spectral samples from COVID-19, with a dataset containing 33 control samples and 30 positive samples for the virus. Table 9 presents the results obtained in this work and the works previously mentioned in the literature.

Although the dataset used was not the same for all studies, the approaches applied in this work and in the work of Carlomagno et al. (2021) using 1D-CNN obtained superior results in respect to the metrics used. As noted earlier in the comparative study, steps such as data pre-processing can significantly influence the final classification results. It is worth mentioning the superiority of 1D-CNN in relation to SVM presented by Yin et al. (2021) using the same the Covid-19 dataset. The results presented by Barauna et al. (2020) using the GA-LDA algorithm show promising sensitivity values, surpassing those obtained in this work by 1%. However, they do not present standard deviation values. The results presented by

Algoritmo	ACC	ESPEC	SE
SG + 1D-CNN*	96.5 ± 1.1	98.06 ± 1.86	94.06 ± 2.43
Yin et al. (2021)*	91 ± 4	93 ± 6	89 ± 7
Barauna et al. (2020)**	90 ± 0	89 ± 0	95 ± 0
Carlomagno et al. (2021)**	97.8 ± 0	98 ± 0	97.5 ± 0

* Tests performed using the same database.

** Tests reporting author's own dataset and not yet publicly available.

Table 9: Positioning of our results in respect to other studies in the literature for different spectral datasets obtained from Sars-Cov2 spectroscopy samples.

Carlomagno et al. (2021) and Barauna et al. (2020) may suffer changes when going through methods such as *cross-validation*. In general, the results obtained using data pre-processed by the SG filter and 1D-CNN show promising results in terms of accuracy and sensitivity metrics for the dataset investigated, indicating the feasibility of automated systems to aid in the diagnosis of the COVID-19 using spectroscopy data.

5 Conclusion

In this work, we present a CNN-1D to tackle classification of spectral data. Firstly, we carried out a comparative study between 1D-CNN and machine learning algorithms as SVM, PLS-DA, KNN and also MOCHM with the objective of evaluating the performance of these algorithms in chemometrics problems. We investigate the impact of data pre-processing and analyzed how this technique influences the performance of each algorithm. The results indicate that the 1D-CNN obtained an average performance superior to the other algorithms investigated when applied to spectroscopy datasets. Furthermore, data pre-processing proved to be a crucial step in the problem modeling process, influencing the gain on average accuracy and reduction of standard deviation, which impacts the reliability of the results. Besides that, the reduced amount of samples present in the datasets did not appear to be an obstacle in generalization of the models for the investigated datasets. Due to its characteristics that favor the use of spectroscopy techniques in various everyday problems, a case study with data obtained from samples of patients with COVID-19 was carried out using 1D-CNN, PLS-DA and KNN for comparison. The algorithms were applied to a public dataset referring to samples of patients from the control group and patients infected with COVID-19. 1D-CNN together with the Savitzky–Golay filter obtained the best results and was used for comparison with other studies involving the diagnosis of COVID-19 from spectral data. The results presented show that approaches involving 1D-CNN's provided the best performance in terms of accuracy, specificity and sensitivity. In particular, in problems such as COVID-19 detection, a high sensitivity rate is very desired and needed since false negatives represent the worst scenario. In this work, the obtained results were, on average, 96% accuracy, 98% specificity and 94% sensitivity, indicating the feasibility of solutions for virus diagnosis using data obtained through spectroscopy with 1D-CNN as classifier. In future works, the influence of signal processing algorithms used in the data pre-processing step can be further investigated. Also, many problems present information beyond that the spectrum provides and may be utilized.

Acknowledgments

References

Ethem Alpaydin. *Introduction to Machine Learning*. MIT press, 2020.

Valério G. Barauna, Maneesh N. Singh, Leonardo Leal Barbosa, Wena Dantas Marcarini, Paula Frizera Vassallo, Jose Geraldo Mill, Rodrigo Ribeiro-Rodrigues, Patrick H. Warnke, and Francis L Martin. Ultra-rapid on-site detection of sars-cov-2 infection using simple ATR-FTIR spectroscopy and analysis algorithm: high sensitivity and specificity. *medRxiv*, 2020. doi:10.1101/2020.11.02.20223560.

C Carlomagno, D Bertazioli, A Gualerzi, S Picciolini, PI Banfi, A Lax, E Messina, J Navarro, L Bianchi, A Caronni, et al. Covid-19 salivary Raman fingerprint: innovative approach for the detection of current and past sars-cov-2 infections. *Scientific Reports*, 11(1):1–13, 2021.

Céline Chauvergne, Laurent Bonnal, Denis Bastianelli, Hélène Carrère, Yves Griveau, Marie-Pierre Jacquemot, Matthieu Reymond, Valérie Méchin, Virginie Rossard, and Éric Latrille. Dataset of organic sample near infrared spectra acquired on different spectrometers. *Data in Brief*, 32:106264, 2020.

Hui Chen, Chao Tan, and Zan Lin. Ensemble of extreme learning machines for multivariate calibration of near-infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 229:117982, 2020.

Yuan-Yuan Chen and Zhi-Bin Wang. End-to-end quantitative analysis modeling of near-infrared spectroscopy based on convolutional neural network. *Journal of Chemometrics*, 33(5):e3122, 2019.

Chenhao Cui and Tom Fearn. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems*, 182:9–20, 2018.

José GM Esgario and Renato A Krohling. Clustering with minimum spanning tree using TOPSIS with multi-criteria information. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2018.

Monika Gniadecka, Peter Alshede Philipsen, Sonja Wessel, Robert Gniadecki, Hans Christian Wulf, Sigurdur Sigurdsson, Ole Faurskov Nielsen, Daniel Højgaard Christensen, Jana Hercogova, Kristian Rossen, et al. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. *Journal of Investigative Dermatology*, 122(2):443–449, 2004.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT Press Cambridge, 2016.

Rita Hayati, Agus Arip Munawar, and F Fachruddin. Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango. *Data in Brief*, page 105571, 2020.

Benjawan Kaewseekhao, Noppadon Nuntawong, Pitak Eiamchai, Sittiruk Roytrakul, Wipa Reechaipichitkul, and Kiatichai Faksri. Dataset of serum proteomic spectra from tuberculosis patients detected by Raman spectroscopy and surface-enhanced Raman spectroscopy. *Data in Brief*, 28:104891, 2020.

Muzammil Khana, Muhammad Taqi Mehrana, Zeeshan Ul Haqa, Zahid Ullaha, Salman Raza, Naqvia Mehreen, and Ihsanb Haider Abbass. Applications of artificial intelligence in covid-19 pandemic: A comprehensive review. *Expert Systems with Application*, 185(1): 115695, 2021.

Frédéric Kosmowski and Tigist Worku. Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. *PloS one*, 13 (3):e0193620, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25: 1097–1105, 2012.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Loong Chuen Lee, Choong-Yeun Liong, and Abdul Aziz Jemain. A contemporary review on data preprocessing (dp) practice strategy in ATR-FTIR spectrum. *Chemometrics and Intelligent Laboratory Systems*, 163:64–75, 2017.

Ana Mara Ferreira Lima, Camila Ribeiro Daniel, Ricardo Scarparo Navarro, Benito Bodanese, Carlos Augusto Pasqualucci, Marcos Tadeu Tavares Pacheco, Renato Amaro Zângaro, and Landulfo Silveira Jr. Discrimination of non-melanoma skin cancer and keratosis from normal skin tissue in vivo and ex vivo by Raman spectroscopy. *Vibrational Spectroscopy*, 100:131–141, 2019.

Ting Liu, Zhongren Li, Chunxia Yu, and Yuhua Qin. NIRS feature extraction based on deep auto-encoder neural network. *Infrared Physics & Technology*, 87:124–128, 2017.

Laura M McIntosh, Randy Summers, Michael Jackson, Henry H Mantsch, James R Mansfield, Marilyn Howlett, A Neil Crowson, and John WP Toole. Towards non-invasive screening of skin lesions by near-infrared spectroscopy. *Journal of Investigative Dermatology*, 116(1): 175–181, 2001.

OMS. Archived: who timeline - covid-19, 2020. URL https://www.who.int/news/item/27-04-2020-who-timeline---covid-19.

OMS. World health organization coronavirus disease 2019 (covid-19) situation report, 2020.

Avneet Pannu. Artificial intelligence and its application in different areas. *Artificial Intelligence*, 4(10):79–84, 2015.

Dan Peng, Linqing Li, He Guo, Yanlan Bi, and Guolong Yang. Information extraction in frequency domain based on entropy theory and genetic algorithm in near-infrared spectra. In *5th International Conference on Information Science and Control Engineering (ICISCE)*, pages 1143–1147, 2018.

Åsmund Rinnan, Frans Van Den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, 28(10):1201–1222, 2009.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

Robert J Schalkoff. *Artificial Neural Networks*. McGraw-Hill Higher Education, 1997.

Silvia Seoni, Federica Veronese, Vanessa Tarantino, Elisa Zavattaro, Massimo Salvi, Nicola Michielli, Bruno De Santi, Filippo Molinari, Paola Savoia, and Kristen M Meiburger. Non-invasive analysis of actinic keratosis using a cold stimulation and near-infrared spectroscopy. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 467–470, 2019.

Barbara Stuart. Infrared spectroscopy. *Kirk-Othmer Encyclopedia of Chemical Technology*, 2000.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Colin P West, Victor M Montori, and Priya Sampathkumar. Covid-19 testing: the threat of false-negative results. In *Mayo Clinic Proceedings*, volume 95, page 1127, 2020.

Li-yan Wu, Dong-sheng Yang, Ming-zhi Zhao, and Fan-xin Meng. Application of principal component analysis-artificial neural network in near infrared spectroscopy for non-destructive determination of coriolus versicolor. In *2012 Fifth International Conference on Intelligent Computation Technology and Automation*, pages 106–109. IEEE, 2012.

Yuhong Xiang, Jing Tian, Zhuoyong Zhang, and Yinmei Dai. Diagnosis of endometrial cancer based on near infrared spectroscopy and general regression neural network. In *2010 Sixth International Conference on Natural Computation*, volume 3, pages 1228–1231. IEEE, 2010.

Xu Yan, Sheng Zhang, Hao Fu, and Haibin Qu. Combining convolutional neural networks and on-line Raman spectroscopy for monitoring the cornu caprae hircus hydrolysis process. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 226:117589, 2020.

Gang Yin, Lintao Li, Shun Lu, Yu Yin, Yuanzhang Su, Yilan Zeng, Mei Luo, Maohua Ma, Hongyan Zhou, Lucia Orlandini, et al. An efficient primary screening of covid-19 by serum Raman spectroscopy. *Journal of Raman Spectroscopy*, pages 1–10, 2021.

Chen Yuanyuan and Wang Zhibin. Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks. *Chemometrics and Intelligent Laboratory Systems*, 181:1–10, 2018.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. 2020. https://d2l.ai.

Xiaolei Zhang, Tao Lin, Jinfan Xu, Xuan Luo, and Yibin Ying. Deepspectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, 1058: 48–57, 2019.

Wenbin Zheng, Hongping Shu, Hong Tang, and Haiqing Zhang. Spectra data classification with kernel extreme learning machine. *Chemometrics and Intelligent Laboratory Systems*, 192:103815, 2019.

Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733, 2020.

Andrey Zyubin, Vladimir Rafalskiy, Anna Tcibulnikova, Karina Matveeva, Ekaterina Moiseeva, Alina Tsapkova, Ilia Samusev, Valery Bryukhanov, and Maksim Demin. Dataset of human platelets in healthy and individuals with cardiovascular pathology obtained by surface-enhanced Raman spectroscopy. *Data in Brief*, 29:105145, 2020.