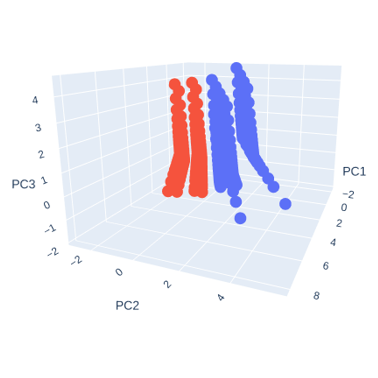
PCA statistical method for classification of HIV HCV sensors

**ABSTRACT:** Many of the global population suffer from infectious diseases, so studies in the health area aimed at identifying these diseases are critical. Rapid detection tests are essential for disease control and eradication. Some diseases have a long immunological window, where antibodies take a long time to be identified. A possible identification and classification method uses the statistical analysis performed by the PCA (Principal Components Analysis), through which we can reduce the number of variables and identify the presence of these antibodies. This article aims to classify immunosensors according to the antibody detected, analyzing their responses in relation to impedance and frequency using the PCA statistical method. An interactive laboratory was adopted with Jupyter Notebook, Python for the PCA statistical method, using well-known libraries such as Pandas and Scikit-learn. The study was based on data collected from two immunosensors, HCV sensor and HIV sensor, analyzing their response as a function of antibody concentration. This article analyzed several data and variables from the dataset of both sensors to build models with the PCA statistical method; it was possible to separate and classify the HIV and HCV sensors at specific concentrations. It was possible to identify the dependence of the response of both sensors with the adopted immobilization matrix. The PCA analysis for the selected datasets showed a relevant classification using PC1, frequency and PC2, concentration, being the variable PC2 the main responsible for separating and identifying the sensors.

*Keywords*

1. PCA; 2. HIV; 3. HCV; 4. sensor; 5. antigen; 6. antibody



1. **Introduction**

The World Health Organization (WHO) identifies the most serious diseases for the population and some of them are prioritized for research and development with the aim of being eradicated, such as malaria, tuberculosis, human immunodeficiency virus (HIV), infection and human papillomavirus (HPV), among others. Methods that allow rapid detection of diseases, rapid detection test (RDT), are important resources[1].

Electrochemical biosensors have advantages for detecting some diseases because they are endowed with simplicity, low cost, fast response, good selectivity, and sensitivity[1], [2].

Currently, an excellent method to detect HIV in the early stages of the disease is the method that uses the detection of HIV-p24 antigen[1]. One of the major problems in HIV detection is co-infection with other diseases, such as hepatitis C (HCV)[1], a chronic infection that currently affects 200 million people worldwide[3]. An HIV/HCV co-infection is a very serious health issue and in recent years, it has been among the leading causes of morbidity and mortality in the world. These diseases have similar means of transmission and approximately a quarter of HIV seropositive people have HCV. In addition, patients with HIV, when exposed to HCV, have difficulty controlling the infection, and patients with HCV, when exposed to HIV, significantly increase the risk of disease transmission[4], so methods for multiple detections of these diseases are of great interest[1].

In recent years, several studies have been carried out using a statistical method to extract and classify data in the health care field. This method is called Principal Component Analysis (PCA)[5]. PCA is a statistical method widely used in analyzing multivariable data and dimension reduction techniques for many parameters or measurements, identifying the dominant patterns in these data[6]–[9]. It is a technique applied to extract the most useful information, principal components (PCs), and to verify the relationship between objects and variables[6], [8]. After applying the method, the principal components (PC) used in the PCA method will have a good representation of the data since the data with the most significant variables are used or the data most impacted during the measurement process[9]. Each analyzed substance leaves a unique footprint on the sensors; that footprint will be analyzed by the algorithm created after applying the PCA and machine learning algorithms, where the substances can be identified. Once the detection method and the sensitive element to the sample have been decided and applied, the data can be interpreted[6].

This article presents the step-by-step use of the PCA statistical method to analyze and classify data from two sensors used to identify HIV and HCV antigens and compare them.

This paper is organized as follows. In section 2, we will have the PCA projection techniques and an explanation of the statistical method. In session 3 we will present the methodology and how the sensors were prepared, in session 4 we will present the analysis of the obtained data and results. Then the research conclusions are presented in section 5.

* 1. *PCA mathematical foundations*

The objective of this section is present the concept behind the statistical method and how its projection technique works.

PCA is an unsupervised mathematical method that uses data matrices, where X can be the samples and Y the sample variables, to convert a large amount of possibly related data into a small amount of unrelated data[8]. Orthogonal vectors, called principal components (PC), represent the maximum variation of the original data [6], [9]. Although there is this loss of less sensitive information in the application of the method, it is advantageous in cases of high correlation between the substances to be detected[9].

The analysis of multivariable data is applied to obtain statistical data from a large number of parameters or measurements, identifying the dominant patterns in these data[6]. The PCA projection allows for observing all the information from the database in a point graph with the linear combinations of the original response signal from the sensors where the axes will represent the principal components (PCs)[6].

Considering a matrix of X of variables with the following representation:

The first step of the analysis is the preprocessing of the dataset. To perform the PCA analysis, it is necessary to subtract the mean of each variable from the data, the feature scaling or standard scaler, producing a dataset with the mean centered on zero (0). This ensures that the data is centered at the origin of the principal components and maintains the spatial relationship of the data and the value of the variables[10]–[12].

The X̅ is the mean value that is equal to the sum of all X variables divided by the number of samples n of X variable. Now that we have the mean of the variable X, we need to know the scale of the variable and how much our data is spread on this scale. For this, we need to calculate the Standard Deviation (SD)[12].

Standard Deviation (SD) represents the average distance from the calculated mean to a point in the dataset. It represents the square root of the sum of each point of variable X minus its mean value X̅ squared, divided by the number of samples of that variable X minus 1[12].

In this article, the samples are represented by the row vector, each sample has a representative line and a representative point line regarding its frequency value, and each variable is represented by the column vector[11].

Principal Component Analysis (PCA) is then defined mathematically as the decomposition of the initial matrix into two matrices, the first one representing the T scores and the second one representing the L loadings[11].

The scores (T) indicate the relationship between the samples and the loadings (L) indicate the relationships between the variables. PCA is a statistical method of data projection. When the initial matrix decomposition is performed from related data to unrelated data, using only the scores and loadings, each of the J columns of the loading matrix (L) starts to define the direction of the axis of a PC (Principal Component), projecting a new score matrix [10], [11].

The R-value is the transformation matrix from the original to the principal components (PC)[11].

The scores and loadings can have positive and negative values, having no relation to the original values. When we have positive loadings, this indicates that the variable is positively correlated with the PC. That is, if the value of the variable increases, it will increase the point value in the PC and then, when we have negative loadings, this indicates that the variable is negatively correlated with PC, and if there is an indication of null value, zero, it means that there is no correlation. If the loading is positively or negatively large, it indicates a great dependence of the Principal Component on this variable[10], [12], [13].

* 1. *Covariance matrix*

Starting the concept with variance (s2), the variance calculation resembles the Standard Deviation (SD) calculation. The variance represents a Principal Component (PC) 's variation in the dataset[12].

To understand the concept of covariance matrix, we need to understand the concept of covariance. The variance is a calculation made in a single dimension where it is possible to calculate the variation in that dimension. Still, if there is more than one dimension, it is possible to calculate the variation between the mean value calculated in each of these dimensions; this calculation is called covariance (cov) [12].

Since we could have several variables (var) in different dimensions (X, Y and Z), we could calculate all the covariance data in a matrix, finally getting the covariance matrix. Based on this information, we can represent a covariance matrix (C) as follows:

The covariance matrix will be a square matrix, and it will represent the covariance between the variables in the dataset[12]. Eigenvectors of a matrix X is a vector μ, which can be represented as:

The λ value is the scalar called the eigenvalue. Considering the different variables and their dimensions in space, and if there is a correlation between them, the PCA will identify this correlation and trace a direction in order to minimize the sum of squared deviations, this vector created in this direction will form the first Principal Component (PC) of the data, or eigenvector 1. This vector represents the largest linear variation in the plane, and the second largest linear orthogonal variation to this vector would then be PC2, and so forth[14].

* 1. *Eigenvectors and eigenvalues of covariance matrix*

Eigenvectors and eigenvalues are values that provide the eigen-decomposition of a square matrix. In the case of PCA, the eigen-decomposition of a covariance matrix will provide the minimum square estimate of the original matrix data[15]. Eigenvectors are the vectors, and eigenvalues are scalar values. That is, they are numerical values associated with a unit of measure. Each eigenvector will be a Principal Component (PC). That is, the directions of the PCA plane and the eigenvalues are the length, the greatest distance, and the values with the greatest magnitude, PC1 for example, would have the greatest eigenvalues and, therefore, the greatest variance[16].

1. **Experimental**

This session will describe the step-by-step development of the analysis performed in the Jupyter Notebook virtual laboratory, using Python and statistical libraries. It will also briefly explain how the sensors were made and their films and measurements performed.

* 1. *Selecting dataset*

The dataset contains background information about sensor measurements (HIV/ HCV). Preparing the data and removing the variables with no significant values was necessary. For our analysis, the variables used were the sensor type, the electrode adopted, the frequency (Hertz), the real impedance (Ohms), imaginary impedance (Ohms), HCV antibody concentration (μg/ml), HIV antibody concentration (μg/ml).

For this step, we imported the main data analysis library, which makes it possible to read the data in csv format, and then the data was loaded into memory. We will adopt the df variable name for the data frame.

Once the data was loaded, all the columns with the data to be treated were selected, except for the column with the target data of the analysis. After this selection, the to\_numpy function was used to transform this data into a list, an array of data.

In the next step, the target column was selected, and the to\_numpy function was applied to return an array with all the values of this column.

* 1. *PCA analysis*

In the previous steps, the data were loaded and selected. In this step, the data were preprocessed and prepared for use in the PCA calculation.

As mentioned earlier in this section, it is imperative to standardize the data because PCA is strongly affected by scale, so we needed to scale the data's features to optimize the performance of the algorithms.

The first step in applying the PCA function was to decompose the variables in the orthogonal components, eigenvalues and eigenvectors according to the maximum variance.

To do this, we applied the function below:

In this function, we choose the number of desired components n\_components=2.

The next step was to build the model; for that, we applied the fit function.

The input data is essentially the x variable because PCA is an unsupervised algorithm, and it does not need the y-plan to learn.

The fit method is a function used to train the algorithm based on the data, using the mean and standard Deviation, based on the machine learning model[17]. Model training in data science is the stage where we design a model in order to minimize data loss while designing the prediction outcome values. After training the model, we applied the transform function to project data based on the trained model. The new data projected represents the scores, that is, the coordinates of the samples in the new axes system.

The following steps were taken to improve the reading and interpretation of the data. First, we nominate the Principal Component vectors as PC1, PC2 and PC3 if necessary. Then the target column was replaced by the respective name of the electrode sensors. The Principal Components data were then concatenated with the target column data. Finally, we retrieve the loading values and the variable's relationship regarding the Principal Components.

To access the loading values, we applied the components\_.T method.

After accessing the stored loadings, we created a new data frame with the Principal Components (PC), where it was possible to analyze the importance of each original variable in the new axis system of the PC.

In this step, the percent variance corresponding to each Principal Component was calculated, and from this point, the need to include or not other Principal Components for the statistical model analysis was considered in order to increase the variance and obtain a better projection of the data.

* 1. *Scree plot*

The scree plot will indicate the relationship between the magnitude of the eigenvalues for each eigenvector, or each Principal Component (PC), explaining the percentage of the variance of that PC in relation to the total data as a PC selection criterion; PC1 will have the highest percentage of variance and will be the most significant, PC2 will have the second highest percentage of variance and so forth[18]–[20]. The greater the percentage of variance desired, the greater the number of Principal Components to be analyzed, as the variance of the Principal Components is added up to 100% variance in the data. The number of Principal Components to be considered will depend on the objective percentage for data variance[20].

* 1. *Scores plot*

The scores plot indicates the relationship between the samples helping to identify which are the main components that explain the variations of the original data. This relationship is highly influenced by the scaling and pre-treatment of the original data, so these steps must be performed properly to reduce significant value effects that can dominate comparisons in a dataset[21], [22].

* 1. *Loading plot*

In loading plot visualization, we can understand the reason that the data were grouped and the influence of their positions by studying the relationship between the variables. Variables positioned in the same direction from the center have a positive correlation, and variables positioned on opposite sides of the center have a negative correlation[22]. While the scores plot will show the classification of the samples according to the Principal Components (PC), the loading plot will contribute to the identification of the variables responsible for this classification[21].

* 1. *HIV Sensors*

For the HIV sensors, Lignin was used in the immobilization matrix. This Lignin was extracted from sugar cane through an organosolv process, a pulping technique that uses an organic solvent. The p17-1 peptide sequence H2N-LSGGELDRWEKIRLRPGG-OH, Anti-p17 antibody and anti-HCV were provided by external companies. The peptide solution was prepared with 0.5mg of it to 1ml phosphate buffered saline solution (PBS). Different concentrations of antibody were prepared for the stock solutions. Lignin and peptide film layers were produced by immersing the substrate in the Lignin solution and then in the peptide solution, washing each immersion with Milli-Q water and drying with nitrogen. Interdigitated gold electrodes were used for impedance spectroscopy analysis and measurements were performed using the Solartron 1260A impedance analyzer in the range from 1Hz to 106Hz[2], [23].

* 1. *HCV Sensors*

For the HCV sensor, Silk Fibroin (SF) was used as an immobilization matrix, this Silk Fibroin was extracted from the Bombyx mori silkworm cocoon, The Silk Fibroin, the peptide PPLLESWKDPDYVPPWHG (NS5A-1), the anti-HCV antibody were provided by external companies. The peptide solution was prepared with 0.5mg of it to 1ml Milli-Q water. Different concentrations of antibody were prepared for the stock solutions. Silk Fibroin and peptide film layers were produced by immersing the substrate in the Silk Fibroin solution and then in the peptide solution, washing each immersion with Milli-Q water to remove loosely adsorbed molecules. Interdigitated gold electrodes were used for impedance spectroscopy analysis and measurements were performed using the Solartron 1260A impedance analyzer in the range from 1Hz to 106Hz[3].

1. **Results and discussion**

This section presents the results of the PCA statistical analysis and the explanation of fundamental concepts adopted, materials and discussions.

* 1. *HIV Sensors*

For the HIV sensor dataset, the variables are frequency range (Freq(Hz)), anti-HIV antibody concentration, electrode adopted, parameters such as series capacitance, parallel capacitance, parallel resistance, real impedance (Z'), imaginary impedance (Z'') and parallel loss.

he antibody used is the anti-p17 which detects HIV-1. The specific sensors for detecting HIV in the presence of anti-p17 antibody are electrode 1, electrode 2, electrode 3, and electrode 4. The specific sensor for detecting HIV in the presence of anti-HCV is the antiHCV Electrode 1, and the non-specific sensor for detecting HIV in the presence of the antibody is the Lignin sensor.

The antibody concentrations used to measure the sensors were 0.0001 μg/mL, 0.001 μg/mL, 0.01 μg/mL, 0.1 μg/mL and 1 μg/mL. For analysis purposes, data in PBS solution, the phosphate buffered saline with zero antibody concentration (PBS) was considered.

Table 1 presents the values of the Principal Components PC1, PC2, PC3 and the target column selected for analysis.

Table 1 – Target column and Principal Components HIV sensor

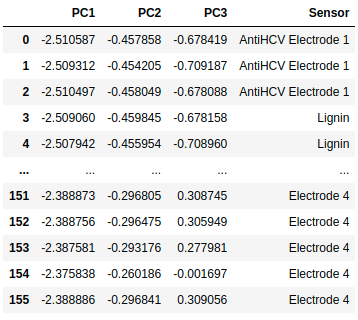


Figure 1 presents the scree plot considering PC1, PC2 and PC3. The PC1, PC2 and PC3 added together reached a variance of 93.3% (Table 2).

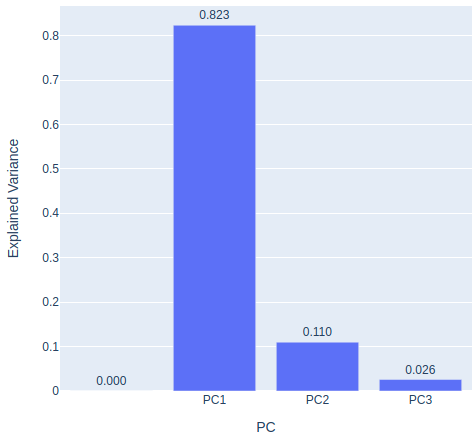
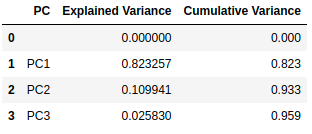


Figure 1 – Scree plot HIV sensor

Table 2 – Explained variance and cumulative variance HIV sensor



In the scores plot (Figure 2), among the specific electrodes for antibody detection, it is possible to notice that electrode 3 has the most significant frequency variation and detection sensitivity. Using the PCA statistical method, there was not a good separation of the sensors at low antibody concentrations. Still, we can observe that the specific sensor to detect HIV in the presence of the HCV antibody, antiHCV electrode 1, separated significantly from the other samples at a higher concentration.

An interesting fact in this analysis is that the non-specific sensor to detect HIV in the presence of the HIV antibody, Lignin, did not have good separation from the other electrodes in the PCA analysis with low concentration. A possible cause would be the application of Lignin to all specific sensors as the antibody immobilization matrix.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

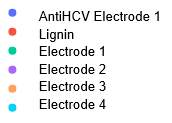


Figure 2 – Scores plot HIV sensor

Analyzing the loading plot (Figure 3) and the magnitude and influence of each PC, we observed that in PC1 the most important variables are the real impedance and imaginary impedance, in PC2 the predominance is of the concentrations of HIV antibody and HCV antibody, and that the PC3 is almost completely influenced by frequency (Table 3).

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Figure 3 – Loading plot HIV sensor

Table 3 – Influence of variables for each PC HIV sensor

Texto

Descrição gerada automaticamente

* 1. *HCV Sensors*

For the HCV sensor data set, the variables frequency (Freq(Hz)), real impedance (Z'), imaginary impedance (Z''), anti-HIV antibody concentration (antiHIVmi-crog/ml) and anti-HCV antibody concentration were considered (antiHCVmicrog/ml).

The antibody used is the NS5A‑1 peptide that detects HCV-1. The specific sensors for detecting HCV in the presence of NS5A‑1 antibody are SF\_NS5A\_1bic and SF\_NS5A\_5bic and the non-specific sensor for detecting HCV in the presence of the anti-p14 HIV antibody is the SF\_NS5A\_5bicp24 sensor.

The antibody concentrations used to measure the sensors were 0.002 µg/mL, 0.01 µg/mL, 0.02 µg/mL, 0.1 µg/mL, 0.2 µg/mL and 0.1 µg/mL. For analysis purposes, data in PBS solution, phosphate buffered saline with zero antibody concentration, were not considered.

Table 4 presents the values of the main components PC1, PC2 and PC3 and the target column selected for analysis.

Table 4 – Target column and Principal Components HCV sensor

Tabela

Descrição gerada automaticamente

The scree plot (Figure 4) considered PC1, PC2 and P3 because PC1 and P2 added together reached a variance of only 64.1%. It is possible to observe that the accumulated variance of the three PCs is 82.8% (Table 5). Although this value is higher than the previous analysis, it is still lower than the desired one as it allows a large margin of error.

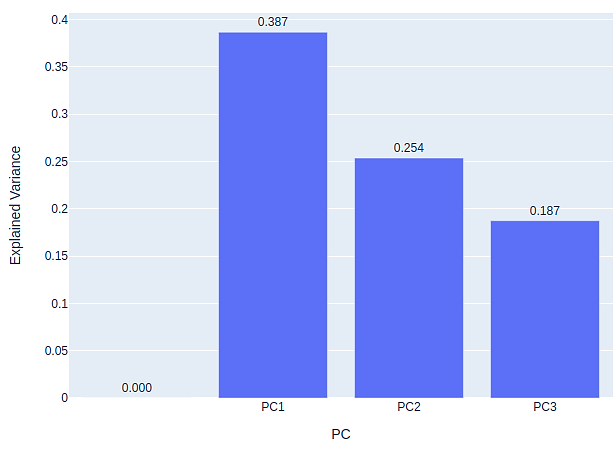


Figure 4 – Scree plot HCV sensor

Table 5 – Explained variance and cumulative variance HCV sensor

Tabela

Descrição gerada automaticamente

In the scores plot (Figure 5), among the specific electrodes, it can be said that all had a very similar response to HCV antibody concentrations, and in this analysis, the non-specific sensor for HCV antibody detection in the presence of HIV antibody, SF\_NS5A\_5bicp24, had an optimal separation of the specific sensors.

Gráfico, Gráfico de radar

Descrição gerada automaticamenteTexto

Descrição gerada automaticamente

Figure 5 – Scores plot HCV sensor

Regarding the loading plot and the magnitude and influence of each PC, we observed in Figure 6 that in PC1 the most important variables are the real impedance and imaginary impedance. In PC2, the predominance is of the HIV antibody and HCV antibody concentrations, and the PC3 is almost completely influenced by frequency (Table 6), the same observations we made in the analysis for the HIV sensor.

Gráfico

Descrição gerada automaticamente

Figure 6 – Loading plot HCV sensor

Table 6 – Influence of variables for each PC HCV sensor

Tabela

Descrição gerada automaticamente

* 1. *Comparison between HIV and HCV sensors*

The comparison between the two sensor's data set used the variables sensor type, frequency (Freq(Hz)), real impedance (Z'), imaginary impedance (Z''), anti-HIV antibody concentration (antiHIVmicrog/ml) and anti-HCV antibody concentration were considered (antiHCVmicrog/ml).

Only specific sensors for each type of antibody were considered in this analysis. The same concentrations as in the previous analyzes were adopted. For analysis purposes, data in PBS solution, phosphate buffered saline with zero antibody concentration, were not considered. Table 7 presents the values of the main components PC1, PC2 and PC3 and the target column selected for analysis.

Table 7 – Target column and Principal Components HCV versus HIV sensors

Tabela

Descrição gerada automaticamente

The scree plot described in Figure 7 considered PC1, PC2 and P3. If PC1 and P2 are added together, the reached a variance will be 52.1%, however adding PC3 to the cumulative variance, only the 68.1% mark was reached (Table 8), reaching a variance lower than that of previous analyses.

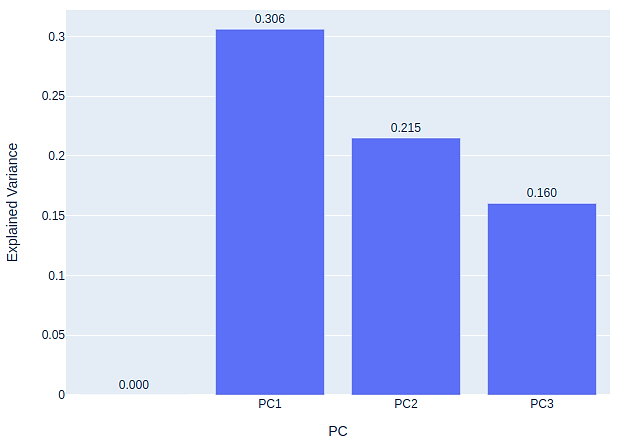


Figure 7 – Scree plot HCV versus HIV sensors

Table 8 – Explained variance and cumulative variance HCV versus HIV sensors

Tabela

Descrição gerada automaticamente

In the scores plot (Figure 8), we can see an overlapping of data in the function of PC2, a possible cause for the overlapping would be a similar response of both sensors to low antibody concentrations.

Gráfico

Descrição gerada automaticamente

Figure 8 – Scores plot HCV versus HIV sensors

In the analysis of the loading plot and the magnitude and influence of each PC represented by Figure 9, some small variations were observed in relation to previous analyses. For PC1 the most important variables are the real impedance and imaginary impedance. In PC2, the predominance is of the concentrations of HIV antibody and HCV antibody. In addition to the sensor type variable (Electrode) having considerable importance, an increase in the contribution of the other variables in this Principal Component was observed. The frequency almost completely influenced the PC3 in the previous analyzes, but in the comparative analysis of the two sensors, a large contribution of the sensor type and an increase of the contribution of the antibody concentrations were observed (Table 9).

Gráfico, Gráfico de dispersão, Gráfico de bolhas

Descrição gerada automaticamente

Figure 9 – Loading plot HCV versus HIV sensors

Table 9 – Influence of variables for each PC HCV versus HIV sensors

Texto

Descrição gerada automaticamente

With the motivation to understand the overlapping and which antibody concentrations were influencing the overlapping of the data, new models were created for analysis.

First, all concentrations of the order of 10-3 μg/mL were removed from the data set, but a small overlapping was still observed, so a new model was created, this time removing concentrations of the order of 10-2 μg/mL and this time, the separation between the sensors occurred perfectly in the function of the concentration of antibodies and their types, in PC2.

Figure 10 shows the new variances of each PC after the model changes. With the model changes, it was possible to observe an increase in the cumulative variance of 8.37%, reaching the mark of 73.8% for PC1, PC2 and PC3 (Table 10). This increase is mainly due to PC2, which increased by about 21% and has a greater predominance of variables related to antibody concentration.

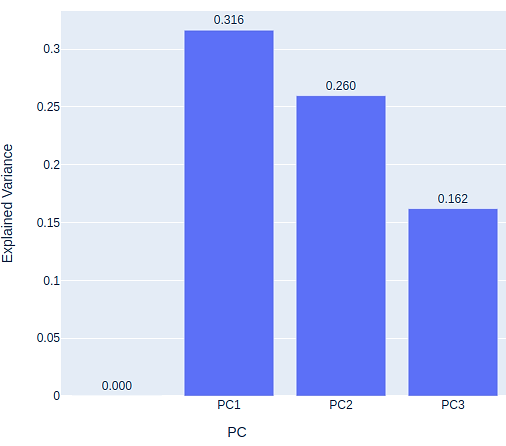


Figure 10 – Scree plot model changes HCV versus HIV sensors

Table 10 – Explained variance and cumulative variance model changes HCV versus HIV sensors

Tabela

Descrição gerada automaticamente

Analyzing the scores plot (Figure 11) for the new model, it is possible to verify the complete separation of the HIV and HCV sensors in the function of PC2.

Gráfico

Descrição gerada automaticamente

Figure 11 – Scores plot model changes HCV versus HIV sensors

The loading plot now has several variations (Figure 12). We can observe that the predominance in PC1 continues to be in function of the impedances, real and imaginary. However, we significantly increased PC1 regarding contributions of the HIV and HCV antibody concentrations. The PC2 also presented variations in the contribution of the impedances, real and imaginary, which increased greatly, while the contribution of the other variables presented a slight increase or decrease. The variations in PC3 were slight and the most important variables remain the frequency and sensor type (Table 11).

Gráfico, Gráfico de dispersão, Gráfico de bolhas

Descrição gerada automaticamente

Figure 12 – Loading plot model changes HCV versus HIV sensors

Table 11 – Influence of variables for each PC model changes HCV versus HIV sensors

Tabela

Descrição gerada automaticamente com confiança média

**4. Conclusions**

In the HCV sensor, there was a clear separation of the non-specific Sensor for HCV detection in the presence of anti-HIV (anti-p24) from the other sensors, it was possible to observe the classification according to PC2, which is characterized by the change in concentration.

As for the HIV sensor, we saw that the sensors follow the response and are more dependent on Lignin, which is the immobilization matrix adopted for them. It is possible to notice that at the lowest antibody concentrations, the specific sensor for detecting HIV in the presence of anti-HCV also followed the other responses, but it was identified and classified at higher concentrations as a function of PC2, separating it from the other samples.

When the responses of both sensors are compared, we can notice a variation and classification mostly due to PC1 and PC2 variations. Again, it is possible to note that we have overlapping data from both sensors at lower concentrations, and classification is only possible at higher concentrations.

It was possible to process and classify the data using the PCA statistical meth-od, but the results of the summed variance in PC3 did not exceed 90% in any scenario, indicating a considerable range of errors. Other statistical methods can be used in sets to improve the results' accuracy and the samples' representativeness. Other measurements can be made by adding new variables that can be treated by the PCA statistical method, enriching this study.

**References**

[1] K. I. Ozoemena and S. Carrara, “Biomedical electrochemical sensors for resource-limited countries,” *Curr. Opin. Electrochem.*, vol. 3, no. 1, pp. 51–56, 2017, doi: 10.1016/j.coelec.2017.06.002.

[2] B. M. Cerrutti, M. L. Moraes, S. H. Pulcinelli, and C. V. Santilli, “Lignin as immobilization matrix for HIV p17 peptide used in immunosensing,” *Biosens. Bioelectron.*, vol. 71, pp. 420–426, 2015, doi: 10.1016/j.bios.2015.04.054.

[3] L. R. Lima, A. M. B. Gonçalves, F. V. Paulovich, O. N. Oliveira, S. J. L. Ribeiro, and M. L. Moraes, “Electrical immunosensor made with antigenic peptide NS5A- 1 immobilized onto silk fibroin for diagnosing hepatitis C,” *J. Braz. Chem. Soc.*, vol. 29, no. 10, pp. 2054–2059, 2018, doi: 10.21577/0103-5053.20180080.

[4] C. Zhao and X. Liu, “A portable, paper-based multiplexing immunosensor for detection of HIV and HCV markers in serum,” *2015 Transducers - 2015 18th Int. Conf. Solid-State Sensors, Actuators Microsystems, TRANSDUCERS 2015*, pp. 1025–1028, 2015, doi: 10.1109/TRANSDUCERS.2015.7181100.

[5] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, “Classification models for heart disease prediction using feature selection and PCA,” *Informatics Med. Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[6] L. M. M. Ferro, S. G. Lemos, M. Ferreira, and F. Trivinho-Strixino, “Use of multivariate analysis on Fabry-Pérot interference spectra of nanoporous anodic alumina (NAA) for optical sensors purposes,” *Sensors Actuators, B Chem.*, vol. 248, pp. 718–723, 2017, doi: 10.1016/j.snb.2017.04.051.

[7] N. Vaswani and P. Narayanamurthy, “Static and Dynamic Robust PCA and Matrix Completion: A Review,” *Proc. IEEE*, vol. 106, no. 8, pp. 1359–1379, 2018, doi: 10.1109/JPROC.2018.2844126.

[8] M. A. H. Khan, B. Thomson, R. Debnath, A. Motayed, and M. V. Rao, “Nanowire-Based Sensor Array for Detection of Cross-Sensitive Gases Using PCA and Machine Learning Algorithms,” *IEEE Sens. J.*, vol. 20, no. 11, pp. 6020–6028, 2020, doi: 10.1109/JSEN.2020.2972542.

[9] K. Ahuja and A. K. Pani, “Software sensor development for product concentration monitoring in fed-batch fermentation process using dynamic principal component regression,” *ICSNS 2018 - Proc. IEEE Int. Conf. Soft-Computing Netw. Secur.*, no. 1, pp. 1–6, 2018, doi: 10.1109/ICSNS.2018.8573661.

[10] S. M. Ho, “Principal Components Analysis (Pca),” *Princ. Components Anal.*, no. December, p. 12, 2019.

[11] M. M. C. Ferreira, *Quimiometria: Conceitos, Métodos e Aplicações*. 2015.

[12] L. Smith, “A tutorial on PCSA,” *Dep. Comput. Sci. Univ. Otago.*, pp. 12–28, 2006.

[13] H. Shinzawa, K. Awa, W. Kanematsu, and Y. Ozaki, “Multivariate data analysis for Raman spectroscopic imaging,” *J. Raman Spectrosc.*, vol. 40, no. 12, pp. 1720–1725, 2009, doi: 10.1002/jrs.2525.

[14] C. R. Wall, J. V.; Jenkins, *Practical Statistics for Astronomers*, vol. 1999, no. December. 2006.

[15] N. Salkind, “Eigen decomposition,” *Encycl. Meas. Stat.*, pp. 1–10, 2013, doi: 10.4135/9781412952644.n150.

[16] N. S. Tomar, Geetam Singh; Barbosa, Jorge Luis V.; Aghwariya, Mahesh Kumar; Chaudhari, “International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019,” *Springer Nat. Singapore*, p. 1689, 2020.

[17] J. Brownlee, “Machine Learning Mastery With Python: Data Cleaning, Feature Selection, and Data Transforms in Python,” 2020.

[18] L. Gour, S. B. Maurya, G. K. Koutu, S. K. Singh, S. S. Shukla, and D. K. Mishra, “Characterization of rice (Orzya sativa L.) genotypes using principal component analysis including scree plot and rotated component matrix,” *Int. J. Chem. Stud.*, vol. 5, no. 4, pp. 975–983, 2017.

[19] J. D. Brown, “Questions and answers about language testing statistics : Choosing the Right Number of Components or Factors in PCA and EFA Choosing the Number of Components or Factors to Include in a PCA or EFA Examples Illustrating the Five Stopping Rules,” *JALT Test. Eval. SIG*, vol. 13, no. May, pp. 19–23, 2009.

[20] F. P. G. Márquez, *Handbook of Research on Big Data Clustering and Machine Learning*. 2019.

[21] J. A. Williams, J. R. Koup, R. Lalonde, and D. D. Christ, *Predictive Approaches in Drug Discovery and Development: Biomarkers and In Vitro/In Vivo Correlations*. 2012.

[22] L. P. Esbensen, Kim H.; Guyot, Dominique; Westad, Frank; Houmoller, “Multivariate Data Analysis In Practice : an Introduction to Multivariate Data Analysis and Experimental Design,” p. 598, 2002.

[23] S. Imran, S. Ahmadi, and K. Kerman, “Electrochemical biosensors for the detection of sars-cov-2 and other viruses,” *Micromachines*, vol. 12, no. 2, pp. 1–23, 2021, doi: 10.3390/mi12020174.