Machine Learning methods for micro-FTIR imaging classification of human skin tumors

Matheus Del Valle
Center for Lasers and Applications
Instituto de Pesquisas Energéticas e Nucleares IPEN/CNEN
São Paulo, Brazil
matheus.valle@usp.br

Kleber Stancari
Center for Lasers and Applications
Instituto de Pesquisas Energéticas e Nucleares IPEN/CNEN
São Paulo, Brazil
kleberstancari@usp.br

Pedro Arthur Augusto de Castro
Center for Lasers and Applications
Instituto de Pesquisas Energéticas e Nucleares IPEN/CNEN
São Paulo, Brazil
pedroarthur@usp.br

Moisés Oliveira dos Santos
Center for Lasers and Applications
Instituto de Pesquisas Energéticas e Nucleares IPEN/CNEN
São Paulo, Brazil
and Technology School (EST)
Amazonas State University (UEA)
Amazonas, Brazil
mosantos@uea.edu.br

Denise Maria Zezell*
Center for Lasers and Applications
Instituto de Pesquisas Energéticas e Nucleares IPEN/CNEN
São Paulo, Brazil
http://orcid.org/0000-0001-7404-9606

***Abstract*— This review presents some methods applied to micro-FTIR imaging for classification of human skin tumors. It is a collection of the pre-processing pipeline and machine learning classification models. The aim of this review is to update and summarize the current methods which are applied in our skin tumor research.**

Keywords*— *micro-FTIR imaging, skin tumor, machine learning, spectral bands

I. INTRODUCTION

The Fourier-transform infrared (FTIR) spectroscopy has emerged as one of the important tools to study biological materials and enable cell biology analysis. It is a so-called label free, a technique relatively simple, reproducible, non-destructive to the tissue and with substantially accurate results

In recent years, infrared spectroscopy analysis is often confronted with large amounts of data and the fundamental information may not be readily evident [1]. The infrared imaging collected during analysis of the sample presents observations regarding the peak and band area, as the wavelength of these data Each one can be analyzed by a different dimension. The classical statistics, through its models (parametric and non parametric), always had a strong preference for low-dimensional parametric models [2], not being able to handle the growing increase in the volume of data generated and its high dimensional [3]

The use of techniques capable to extract information that is hidden in the infrared spectra, are increasingly vital in analytical chemistry. Analysis that goes beyond the one dimensional space, revealing characteristics or properties in collected data from the samples, have been the main appeal of the researchers In this scenario, the multivariate data analysis has many advantages to be explored and some are already published in the scientific literature [4] [5].

This review highlights and briefly discusses the multivariate classification of spectroscopy analysis data from human skin tumor The spectral data analysis pipeline, according to Morais, C. et al [6], is compose of following stages: data acquisition, outlier detection, preprocessing, data selection, model construction and validation. To restrict the scope of this review we will briefly discuss the preprocessing steps and the modeling that our group are using in the skin tumor analysis.

II. PREPROCESSING

Preprocessing is applied to improve data quality, decreasing undesired signal contributions (not related to the target sample). Even though there is no unique preprocessing pipeline for all kinds of spectra and analysis, the steps must follow a logical order with adequate parameters, otherwise preprocessing may mask the signal of interest or add bias to the data. General preprocessing steps are [6-8]:

1. Quality test: evaluate the data quality using raw spectra. Usually, Amide I and II region (1700 to 1500 cm^{-1}) is used to check biological tissue signal, while 1900 to 1800 cm^{1} is known as the dead region, without tissue signals. A more robust technique may be used, such as Hotelling's T^2 vs Q residuals chart.
2. Region of interest truncation: the region to be selected is related to the slide type, the tissue being analyzed, and the equipment used. However, the 1800 to 900 cm^{-1} usually used, as it is the biofingerprint region.
3. Removal of substrate contributions: paraffin may be removed by truncation or by digital filters, as adding digital de-waxing in the extended multiplicative signal correction (EMSC); glass slides also demand special truncations.
4. Smoothing: necessary to remove random noise from the spectra. The most used method is the Savitzky-Golay (SG) algorithm.
5. Light scattering: to correct this issue, MSC and standard normal variate (SNV) are usually applied.

2021 SBFoton International Optics and Photonics Conference (SBFoton IOPC) | 978-1-6654-1948-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SBFOTONIOPC50774.2021.9461969

978-0-7381-2418-6/21/$31.00 ©2021 IEEE

6. Baseline correction: the spectra present background absorption interferences The EMSC technique is the most used technique for baseline correction. Other techniques such as automatic weighted least squares (AWLS) may also be employed.
7 Differentiation: first and second derivative can be applied to correct the baseline and also to evidence overlapped bands It is usually coupled with the smoothing step by the SG algorithm.
8. Normalization: used to correct different sample thickness and concentration. Amide I and vector normalization are usually employed, while the EMSC and SNV methods already applies their own normalizations
9. Outlier detection: usually the Hotelling's T^2 vs Q residuals chart is used for outlier detection, although several clustering techniques may be used
10. Data selection and extraction: as spectra contain a large number of features, these techniques may be applied before data modeling. The principal component analysis (PCA) is one of the most used techniques to obtain dimensionality reduction.

III. Machine Learning Classification Algorithms

Supervised machine learning classification models can be trained to predict samples labels. To make this possible, the training step needs to have both the spectra and the ground truth label, so that the model can learn from it. There are a huge number of models, where the main algorithms for spectra analysis are discussed in the following.

A Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) calculates discriminant projections vectors to achieve a data dimensionality reduction [9] These projections form the maximum intraclass separation and minimum interclass distance. Given two labeled data x_1 and x_2, the linear projection w that maximizes the interclass distance is obtained by the following:

1. Calculate the classes center M
2 Calculate the intraclass and interclass scatter matrices.

$$S_{intra} = S_1 + S_2$$

$$S_{inter} = (m_1 - m_2)(m_1 - m_2)^T$$

where S_1 and S_2 are given by

$$S_i = \sum \quad (x - m_i)(x - m_i)^T$$

3. Compute the LDA projection by maximizing J

$$J = \frac{w^T S_{inter} w}{w^T S_{intra} w}$$

4. To solve the J optimization problem, i.e., the minimum w for $w^T S_{intra} x$ and $w^T S_{inter} w = c$, where c is a non-zero value, the Lagrangian can be constructed as:

$$L(w, \lambda_{lda}) = w^T S_{intra} w - \lambda_{lda}(w^T S_{inter} w - c)$$

$$\frac{\partial L(w, \lambda_{lda})}{\partial w} = S_{intra} w \qquad \lambda_{lda} S_{inter} w$$

5. The optimal w can be obtained as

$$S_{intra} w = \lambda_{lda} S_{inter} w$$

LDA assumes that the data is gaussian shaped (normal distribution) Spectral data usually does not present normal distribution, although LDA can handle modeling this kind of data, and according to the central limit theorem, increasing the dataset size can help to overcome this issue. The number of samples should be significantly larger than the number of features. In addition, LDA is negatively affected when the attribute variance is not equally distributed, which is usually the case of complex biological media [6].

B Partial Least Squares Discriminant Analysis

The Partial Least Squares (PLS) PLS seeks to maximize the covariance, finding the direction of the space of the predictors that explains the greatest variance of class space The PLS in its classical form is based on nonlinear iterative partial least squares (NIPALS) algorithm Given a sample data X, a its label matrix Y, and a number of latent variables, NIPALS steps are:

1. Calculate the sample weights and normalize it

$$w = X^T y$$

$$w = w / ||w||$$

2. Obtain the data scores

$$t = Xw$$

3. Calculate the weight of the labels and normalize it

$$c_y = Y^T t \,/\, (t^T t)$$

$$c_y = c_y / \left|\left|c_y\right|\right|$$

4. Get the vector u

$$u = Y c_y$$

5. Iterate steps 1 to 4 until t converges (current t equal to last t)

6. Compute the loading vector of $X$and Y, respectively

$$p = X^T t \,/\, t^T t$$

$$q = Y^T t \,/\, t^T t$$

7. Calculate the coefficient b

$$b = u^T t \,/\, (t^T t)$$

8. Update data and label matrices as

$$X = X - tq^T$$

$$Y = Y - tq^T$$

9 The vector t, p, u, b and $q$can be saved and the next component can be obtained restarting the first step

The PLS model look for the multidimensional direction in the data space that explains the maximum variance direction in the label space:

$$X = TP^T + E$$

$$Y + UQ^T + F$$

$$U = TD + H$$

where T and U are the score matrices, $P, Y$and Q the loadings and $E, F$and $H$are the residuals.

The discriminant analysis term (DA) for PLS refers to the use of a threshold after the decomposition to enable a classification. As it is a binary classifier, the threshold is usually set to 0.5.

PLS-DA is negatively affected by unbalanced classes and the number of latent variables requires a gridsearch optimization.

C. K-nearest neighbors

The KNN algorithm calculates the distance between a test data and the training data (labeled) [10]. Euclidean Distance is usually used, but others may also be used, as they are special cases of a more general family of distance functions, L_k:

$$d_k(p,q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k} = \left(\sum_{i=1}^{d} |p_i - q_i|^k\right)^{1/k}$$

where $p, q$are a set of points. When k=1,Manhattan distance is calculated; k=2 the Euclidean Distance; k=3 the Maximum Component.

The nearest K data define the classification of the tested data by a voting system. Voting can be uniform, where all the data has the same weight, or weights can be assigned to the data, such as the inverse of the distance, increasing the influence of closer data. K values usually range from 3 to 50, and the optimal value should be grid searched.

For a classification with a large number of data, search optimization algorithms, such as voronoi diagrams and k-d tree, are used, partitioning the dimensional space. The effectiveness of these optimizations and the high training time are highly affected by the number of features, and dimensionality reduction techniques, such as PCA, may be applied before modeling.

KNN can be modeled into almost all type of data and distribution, although unbalanced classes may add a larger class bias and lead to an overfitting issue When a new sample has to be classified, KNN must re-run all the model training, calculating the distance metrics again, thus being considered a lazy model [6].

D Support Vector Machines

The Support Vector Machines (SVMs) look for a hyperplane that presents the best separation between two classes, based on the points closest to it, called support vectors [10]. If a point is erroneously separated, that is, it is not on the separation side of its correct class, then its distance to the hyperplane is given as an error. The model error is calculated through the sum of all errors, where a cost constant is added as a penalty for incorrect classification. The separating plane of a linear SVM can be written as:

$$w \cdot x - b = 0$$

where $w$is a vector of coefficients and $x$the input variables. Thus, the constraints for class 1 and class 1 for each $x_i$point are, respectively:

$$w \;\; x - b \geq 1$$

$$w \;\; x - b \leq -1$$

The problem optimization for classes y_i of $x_i$is:

$max||w||$, where $y_i(w \;\; x_i - b) \geq 1$ for all $1 \leq i \leq n$

For application in non-linear models, SVM uses the concept of kernel, where a function is applied to the predictors, in order to increase dimensionality and make separation possible. Most used kernels are the polynomial and the radial basis function (RBF or gaussian), defined respectively by:

$$K(x_1, x_2) = (a + {x_1}^T \;\; x_2)^b$$

$$K(x_1, x_2) = exp(-\gamma(d_{12})^2)$$

where b is the polynomial order and $a$is a constant term; $d_{12}$is the euclidean distance between $x_1$and x_2, and $\gamma$is the inverse of the radius of influence of samples selected by the model as support vectors.

The kernel type and parameters optimization is a laborious step, yet RBF usually presents the best adaptation to several data distributions. As it is a binary classifier, multiclass problems can be solved using one versus rest (OvR) or one versus one (OvO) approach.

SVM performs well in higher dimension, but it tends to struggle with the increasing of data complexity and size. Overlapped classes also compromises the model performance [6].

E. Random Forest

The Random Forest (RF) algorithm is an ensemble learning method using decision trees [10]. A decision tree is generated by dividing the data into nodes, which are divided into subsequent nodes through binary choices of a predictor, for example, if the intensity in a wave number is greater or less than a certain value, until it arrives at a classification, given by final nodes, called leaves This structure is also called classification and regression trees (CART).

To evaluate each predicate and how it will contribute to partitioning the set $S$of the samples it is used information entropy or Gini impurity. Given f_i as the fraction of S, the information entropy is defined as

$$H(S) = -\sum_{i=1}^{m} f_i \log_2 f_i$$

The potential split is evaluated by how much it decreases the system entropy. Considering a predicate splitting Sin two subsets, the information gain function is given by

$$IG(S) = H(S) - \sum_{j=1}^{2} \frac{|S_i|}{|S|} H(S_i \ )$$

If Gini impurity is used, then it is based on another quantity ($f_i(1-f_i)$):

$$IG(f) = \sum_{i=1}^{m} f_i(1-f_i)$$

The number of knots can be limited through the concept of pruning, in order to reduce overfitting. When N trees are built, where N is defined by the user, usually with initial tests of 20 to 100, an RF is obtained. The final classification is given by a voting system for each of the trees, which can follow a uniform or weighted voting. This strategy of combining different classifiers into one is called ensemble learning. To avoid high correlation between trees, the bagging or bootstrap approach is used to build the best possible tree in small random subsets of predictors.

RF is robust to outliers and present lower overfitting than most of the machine learning algorithms, as it performs feature selection and generate uncorrelated decision trees. The subsampling method also makes it a good model for dealing with data with large feature quantity. Sparse data usually decreases the model performance

F. Xtreme Gradient Boost

The Xtreme Gradient Boost (XGBoost) is a gradient tree boosting implementation [11] Boosting is a process to weight the classifiers based on how hard the classification is. The gradient descent algorithm is used to minimize the classification errors. In this way, each tree learns from the previous, instead of random non-related trees like in the RF algorithm. The learn the set of functions used in the model, the following objective function is minimized

$$L(\phi) = \sum_{i} \ l(\hat{y}_i, y_i) + \sum_{k} \ \Omega(f_k)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$

The first term in L is related to the loss function, where l calculates the residuals of the predicted $\hat{y}_i$ and the true y_i values. The second part, Ω, is a regularization term, where the number of leaves T is pruned by the penalty $\gamma$and the leaf weights w is regularized by the λterm. The objective function has functions as parameters and cannot be optimized by traditional methods, hence it is trained in an additive manner, where the $f_t$that most improves the model is added.

$$L(t) = \sum_{i=1}^{n} \ l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

XGBoost shows an amazingly fast model training in comparison to most machine learning algorithms. As each tree learns from the previous, it is more prone to overfitting than RF ensemble learning, however, it usually learns better from the features and presents a higher performance To overcome the overfitting issue, the regularization term helps to generalize the model Preprocessing steps to remove data noise may also play an important role for the modeling.

IV. SAMPLES

The success of analysis depends on experimental conditions used (humidity and CO_2) and preprocessing Established the minimum acquisition conditions, preprocessing contributes significantly to the reduction of analytical errors and the success in the application of ML models The Figure 1 shows the mean spectra of skin cancer samples (BCC, SCC, Melanoma and healthy skin), after applied preprocessing pipeline The following steps were applied: truncation to the biofingerprint region (1000 to 1800 cm^{-1}), Savitzky–Golay filter (window lenght = 7), Second Derivative and EMSC.

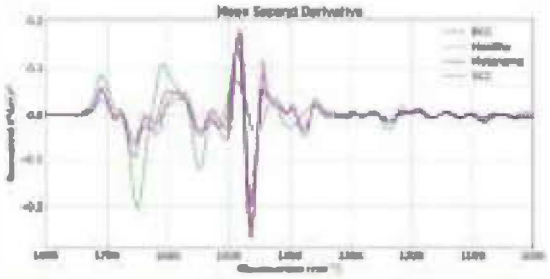

Figure 1- A mean second derivative infrared spectra of basal cell carcinoma (BCC), healthy skin, melanoma, and squamous cell carcinoma (SCC) at biofingerprint region (1000 to 1800 cm^{-1})

V. CONCLUDING REMARKS

The micro-FTIR imaging integrated with multivariate data analysis, has great potential for improving analysis, and facilitate the implementation of the infrared spectroscopy technique as an auxiliary tool in the clinical diagnosis of skin lesions

ACKNOWLEDGMENT

The authors would like to thank the financial support granted by São Paulo Research Foundation FAPESP (CEPID 05/51689 2, 17/50332-0 and 13/26113-6), National Council for Scientific and Technological Development CNPQ (INCT 465763/2014-6, PQ 309902/2017 7) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROCAD 88881 068505/2014-01)

REFERENCES

1 Gautam, R , Vanga, S , Ariese, F , and Umapathy, S.: 'Review of multidimensional data processing approaches for Raman and infrared spectroscopy', EPJ Techniques and Instrumentation, 2015, 2, (1), pp. 1-38

2 Efron, B., and Hastie, T.: 'Computer age statistical inference' (Cambridge University Press, 2016 2016)

3 Bzdok, D., and Yeo, B.T.: 'Inference in the age of big data: Future perspectives on neuroscience', Neuroimage, 2017, 155, pp. 549-564

4 Pereira, T M , Zezell, D.M , Bird, B , Miljković, M., and Diem, M.: 'The characterization of normal thyroid tissue by micro FTIR spectroscopy', Analyst, 2013, 138, (23), pp. 7094-7100

5 Shakya, B R., Shrestha, P , Teppo, H R , and Rieppo, L.: 'The use of Fourier Transform Infrared (FTIR) spectroscopy in skin cancer research: a systematic review', Applied Spectroscopy Reviews, 2020, pp. 1-33

6 Morais, C L., Lima, K.M , Singh, M , and Martin, F.L.: 'Tutorial: multivariate classification for vibrational spectroscopy in biological samples', Nature Protocols, 2020, 15, (7), pp. 2143-2162

7 Baker, M J , Trevisan, J , Bassan, P , Bhargava, R , Butler, H.J., Dorling, K.M., Fielden, P.R., Fogarty, S.W., Fullwood, N J , and Heys, K.A : 'Using Fourier transform IR spectroscopy to analyze biological materials', Nature protocols, 2014, 9, (8), pp 1771

8 Morais, C.L., Paraskevaidi, M., Cui, L., Fullwood, N J , Isabelle, M , Lima, K M , Martin Hirsch, P L , Sreedhar, H., Trevisan, J., and Walsh, M.J.: 'Standardization of complex biologically derived spectrochemical datasets', Nature protocols, 2019, 14, (5), pp 1546 1577

9 Tang, L., Peng, S., Bi, Y., Shan, P., and Hu, X.: 'A new method combining LDA and PLS for dimension reduction', PloS one, 2014, 9, (5), pp. e96944

10 Skiena, S S : 'The data science design manual' (Springer, 2017. 2017)

11 Chen, T : 'guestrin, C Xgboost: A scalable tree boosting system', in Editor (Ed.)^(Eds.): 'Book guestrin, C. Xgboost: A scalable tree boosting system' (2016, edn), pp 785-794