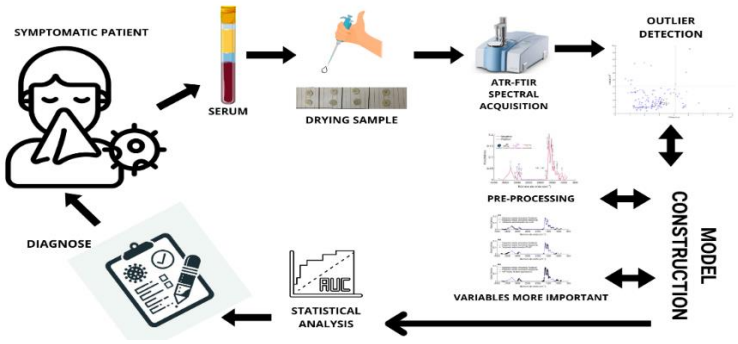
ISSN 2359-4799
Volume 10 / Número 1 / Ano 2024 – p. 01-12
DOI: 10.36524/ric.v10i1.2829



CLASSIFICAÇÃO MULTIVARIADA PARA TRIAGEM CLÍNICA DE COVID-19 POR MEIO DA BIOESPECTROSCOPIA

MULTIVARIATE CLASSIFICATION FOR CLINICAL SCREENING OF COVID-19 USING BIOSPECTROCOPY

Anne Louise Silva Torres[1,2,3]*, Carine Coneglian de Farias[1], Valerio Garrone Barauna[2], Livia C.M. Rodrigues[2], Wanderson Romão[1], Paulo Roberto Filgueiras[3] e Márcia Helena Cassago Nascimento[1,3]

[1] Coordenadoria de Biomedicina, Instituto Federal do Espírito Santo Campus Vila Velha, 29106-010 Vila Velha - ES, Brasil.

[2] Departamento de Ciências Fisiológicas, Universidade Federal do Espírito Santo Centro de Ciências da Saúde, 29047-105 Vitória - ES, Brasil.

[3] Laboratório de Quimiometria do Centro de Competência em Química do Petróleo − NCQP, Universidade Federal do Espírito Santo Centro de Ciências Exatas - 29075-910 Vitória - ES, Brasil.

* anne.torres@estudante.ifes.edu.br

Artigo submetido em 18/06/2024, aceito em 29/07/2024 e publicado em 07/10/2024.
ORCID- Anne Louise Silva Torres: https://orcid.org/0009-0001-8848-936X
ORCID – Carine Coneglian de Farias: https://orcid.org/0000-0003-2967-3420
ORCID - Valério Garrone Barauna: https://orcid.org/0000-0003-2832-0922
ORCID - Lívia do Carmo Melo Rodrigues: https://orcid.org/0000-0002-6004-7981
ORCID – Wanderson Romão: https://orcid.org/0000-0002-2254-6683
ORCID – Paulo Roberto Filgueiras: https://orcid.org/0000-0003-2617-1601
ORCID – Marcia Helena Cassago Nascimento: https://orcid.org/000-0001-5252-586X

Resumo: A COVID-19, causada pelo vírus SARS-CoV-2, é uma doença sistêmica detectada principalmente por métodos sorológicos e moleculares, laboratorialmente dependentes. No entanto, a espectroscopia na região do infravermelho por transformada de Fourier com reflectância total atenuada (ATR-FTIR) associada à quimiometria tem sido estudada para triagem de diversas doenças, inclusive a COVID-19, por ser uma técnica rápida que permite aquisição da informação ao nível molecular. Assim, este trabalho avaliou diferentes abordagens de classificação multivariada na distinção entre amostras de soro de indivíduos infectados pela COVID-19 e pessoas sintomáticas com diagnóstico negativo. Utilizou-se 167 amostras de soro de pacientes sintomáticos, 76 negativos e 91 positivos (Comitê de Ética UFES 51803621.1.0000.5060). Os espectros ATR-FTIR foram coletados com espectrômetro Bruker Alpha II (Bruker) no modo absorbância. Os dados foram pré-processados, divididos em conjunto de treinamento (n=117) e teste externo (n=50) e avaliados pelos métodos de seleção de variáveis e classificação: algoritmo genético com análise discriminante linear (GA-LDA), análise discriminante com mínimos quadrados parciais (PLS-DA) e floresta aleatória com peso de Fisher (PF-RF). O modelo de melhor performance, PF-RF, apresentou 85% de sensibilidade, 73,9% de especificidade e 80% de exatidão. Entre as variáveis, destacaram-se as regiões ~3500 cm^{-1} a ~3000 cm^{-1}, ~3000 cm^{-1} a ~2800 cm^{-1}, ~1700 cm^{-1} a ~1600 cm^{-1}, ~1595 cm^{-1} a ~1512 cm^{-1}, ~1196 cm^{-1} a ~1090 cm^{-1}, atribuídas as macromoléculas de lipídeos, ácidos graxos, proteínas, carboidratos e ácidos nucleicos, respectivamente. Com isso, reforça-se a aplicabilidade da utilização do ATR-FTIR de biofluido associada à classificação multivariada para triagem clínica de doenças.

Palavras-chave: bioespectroscopia; soro; reconhecimento de padrões; quimiometria; seleção de variáveis.

Abstract: COVID-19, caused by the SARS-CoV-2 virus, is classified as a systemic disease and is primarily detected by serological and molecular methods that are laboratory-dependent. However, Fourier-transform infrared spectroscopy with attenuated total reflectance (ATR-FTIR) associated with chemometric methods has been studied for screening of various diseases, including COVID-19, as it is a rapid technique that allows molecular-level information acquisition. This study evaluated different multivariate classification approaches in distinguishing between serum samples from individuals infected with COVID-19 and symptomatic individuals with a negative diagnosis. We used 167 serum samples from symptomatic patients, including 76 negatives and 91 positives (UFES Ethics Committee 51803621.1.0000.5060). ATR-FTIR spectra were collected using a Bruker Alpha II spectrometer (Bruker) in absorbance mode. The data were pre-processed, divided into a training set (n=117) and an external test set (n=50), and evaluated using variable selection and classification methods: genetic algorithm with linear discriminant analysis (GA-LDA), partial least squares discriminant analysis (PLS-DA) and random forest with Fisher's weight (FW-RF). The better performance model, *i.e.* FW-RF demonstrated 85% sensitivity, 73.9% specificity, and 80% accuracy. The variables of interest were predominantly in the regions ~3500 cm^{-1} to ~3000 cm^{-1}, ~3000 cm^{-1} to ~2800 cm^{-1}, ~1700 cm^{-1} to ~1600 cm^{-1}, ~1595 cm^{-1} to ~1512 cm^{-1}, and ~1196 cm^{-1} to ~1090 cm^{-1}, assigned to macromolecules of lipids, fatty acids, proteins, carbohydrates, and nucleic acids, respectively. This reinforces the applicability of ATR-FTIR of biofluids associated with multivariate classification for clinical disease screening.

Keywords: biospectroscopy; serum; pattern recognition; chemometrics; variable selection.

1 INTRODUÇÃO

O vírus SARS-CoV-2, causador da pandemia de COVID-19, é um tipo de coronavírus, de RNA envelopado, fita simples com forma pleomórfica ou esférica e possui projeções de glicoproteína (KHAN; REHMAN, 2020). A COVID-19 é classificada como uma síndrome respiratória, entretanto, estudos revelam ser uma doença sistêmica devido o comprometimento do sistema cardiovascular, respiratório, gastrointestinal, neurológico, imunológico e hematopoiético (ROTHAN; BYRAREDDY, 2020). Os métodos mais empregados para diagnóstico da COVID-19 utilizam métodos sorológicos ou métodos moleculares, sendo esse o padrão-ouro (OLIVEIRA et al., 2022). No entanto, esses métodos são dependentes de reagentes químicos e requerem recursos laboratoriais, além de precisarem de um tempo significativo para liberação de resultados (BRUYNE et al., 2018).

Em contrapartida aos métodos convencionais de diagnóstico para COVID-19, a espectroscopia na região do infravermelho por transformada de Fourier com reflectância total atenuada (ATR-FTIR) tem sido apresentada como uma alternativa de bom custo-benefício, é uma técnica não destrutível, rápida e não necessita de reagentes (BRUYNE et al., 2018). Essa técnica baseia-se na medição da interação da radiação, na região do infravermelho médio, com a matéria em diferentes comprimentos de onda. Isso é possível pelas vibrações moleculares quando ocorre mudança no momento de dipolo das ligações químicas envolvidas na interação, o que possibilita a obtenção da informação química ao nível molecular de biomoléculas presentes nos biofluidos como saliva, plasma, soro, urina (PALUSZKIEWICZ et al.,2020; NASCIMENTO ET AL., 2022; CROCCO et al., 2023; GULEKEN et al., 2022; GIAMOUGIANNIS et al., 2021)

A espectroscopia ATR-FTIR tem sido utilizada associada a métodos de reconhecimento de padrões para triagem clínica de diversas doenças (PALUSZKIEWICZ et al.,2020; NASCIMENTO ET AL., 2022; CROCCO et al., 2023; GULEKEN et al., 2022; GIAMOUGIANNIS et al., 2021; HOFFNER et al., 2014; DEPCIUCH et al., 2019; YANG et al., 2021; BARAUNA et al., 2021; ZHANG et al., 2021; FARIA et al., 2023; BRUN et al., 2024). Uma importante característica dessa técnica é que os espectros ATR-FTIR podem gerar em torno de 1798 variáveis com alto grau de correlação e informação redundante (FOLLI et al., 2023). Ademais, as amostras biológicas apresentam alta variabilidade a partir dos indivíduos e das patologias investigadas, tornando a amostra biológica uma das mais complexas matrizes. Dessa forma, diversas abordagens têm sido propostas para extração de informações relevantes a partir dessa matriz, tais como redução da dimensão e seleção das variáveis para discriminação de grupos. Entre os métodos amplamente utilizados, podemos citar análise de componentes principais (PCA) (BRO; SMILDE, 2014), análise discriminante linear associada a algoritmo genético (GA-LDA) (SHAFFER; SMALL, 1996), análise discriminante associada a mínimos quadrados parciais (PLS-DA) (BALLABIO; CONSONNI, 2013) e floresta aleatória (RF) (BREIMAN, 2001).

Em estudos anteriores do nosso grupo de pesquisa, foram propostos métodos para detecção de COVID-19 baseado em análises de *swab* nasofaríngeo (BARAUNA et al., 2021), saliva (NASCIMENTO et al., 2022) e meios de transporte viral (NOGUEIRA et al., 2021) por meio do ATR-FTIR associado a análises multivariada. No primeiro estudo desenvolvido pelo grupo (BARAUNA et al., 2021), obteve-se 95% de sensibilidade e 89% de especificidade para classificação de indivíduos com COVID-19. Além disso, Zhang et al. (2021) obtiveram 87% de sensibilidade e 98% de especificidade para

identificação de amostras de soro de pacientes com COVID-19 por meio da ATR-FTIR associada a métodos de análise discriminante de mínimos quadrados parciais.

O método de seleção de variáveis mais utilizado para bioespectroscopia é o algoritmo genético associado a análise discriminante linear (GA-LDA) (SANTOS; MORAIS; LIMA, 2020). Esse método combina o algoritmo genético, que é uma técnica de seleção de variáveis para redução da dimensão dos dados espectrais em um número menor de variáveis de maior significância, com a análise discriminante linear que realiza a classificação dos espectros. Isso ocorre por meio de simulação de um processo evolutivo de variáveis, baseado na seleção natural darwiniana, partindo-se de uma população inicial e concluindo com seleção das variáveis de maior aptidão na análise discriminante. Outro método utilizado para seleção de variáveis é o Peso de Fisher (LOVATTI et al., 2020) que é utilizado para maximizar a separação entre classes diferentes e minimizar a variabilidade dentro de cada classe. Para isso, são calculados as médias e a matriz de covariância para cada classe e depois a matriz de dispersão dentro das classes e entre as classes, possibilitando encontrar um vetor que maximiza a razão da dispersão entre as classes. Para discriminação dos grupos o método de PLS-DA (BALLABIO; CONSONNI, 2013) é amplamente utilizado, esse é um método de análise de reconhecimento de padrões supervisionada, ou seja, que passa por um processo de treinamento de modelo. PLS-DA utiliza variáveis latentes para capturar as informações importantes da matriz de dados (**X**) com o vetor de classes (**y**), dessa forma, é calculado uma probabilidade para cada classe e a classificação das amostras é realizada pela classe com maior probabilidade. Por fim, em alguns casos em que se necessita de análise de maior complexidade, ou problemas com pequeno conjunto amostral, o método comumente empregado para classificação é a Floresta aleatória (RF) (BREIMAN, 2001), um modelo conhecido como *ensemble*, que significa um conjunto de modelos para encontrar uma solução combinada. O RF é um método de aprendizado de máquina que utiliza múltiplas árvores de decisão. As árvores formam partições recursivas resultando em nós hierárquicos. Cada nó corresponde a um valor de corte que é equivalente a uma variável dividida em folha (nó final) e uma ramificação para outro nó. A árvore cresce a partir de um nó raiz, que é uma divisão de variável com maior contribuição para distinguir as amostras entra as classes. Assim, partições binárias recursivas são realizadas até que seja alcançado um nó final para distinção das classes (LOVATTI et al., 2019).

Dessa forma, o objetivo deste trabalho foi avaliar diferentes abordagens de classificação multivariada na distinção entre amostras de soro de indivíduos sintomáticos infectados pela COVID-19 e com diagnóstico negativo.

2 MATERIAIS E MÉTODOS

2.1 AMOSTRAS

Foram utilizadas 167 amostras de soro com resultado de teste imunológico, coletadas pelo Laboratório Tommasi-Vitória ES e cedidas sem identificação nominal (76 negativas e 91 positivas). Todas as amostras eram de pacientes com suspeita de infecção por coronavírus (sintomáticos). As amostras foram manuseadas segundo os princípios éticos e aprovado pelo Comitê de Ética em Pesquisa da UFES (nº de aprovação: 51803621.1.0000.5060). Os materiais recebidos do laboratório, foram identificados e organizados de acordo com seus respectivos diagnósticos e mantidos

em freezer (-20ºC) até a aquisição dos espectros.

2.2 AQUISIÇÃO ESPECTRAL ATR-FTIR

O equipamento utilizado na aquisição dos espectros foi o espectrômetro Bruker Optics spectrometer ALPHA II, configurado para leituras no modo de absorbância, na faixa de 4000 a 400 cm^{-1}, programado com resolução de 4 cm^{-1} e com varredura de 32 *scans* para amostra e *background*. A cada medição, realizou-se o procedimento de *background* previsto pelo fabricante para diminuição dos interferentes, principalmente gás carbônico e vapores de água. Pipetou-se 20 µL de cada amostra em uma placa de alumínio que permaneceu para secagem ambiente (18,7ºC e 43,5%) durante um tempo mínimo de 2 horas. A placa de alumínio foi transferida para o cristal ATR e pressionada com o acessório de amostragem, para garantir o contato de um filme fino da amostra sólida com o cristal ATR e realizou-se a medição do espectro pelo software OPUS 8.5. Entre cada análise, o equipamento foi limpo com água milli-q e etanol 70% v/v. Os espectros ATR-FTIR foram adquiridos em triplicata.

2.3 ANÁLISE MULTIVARIADA

Os espectros médios das triplicatas foram inicialmente pré-processados para correção de linha de base por meio do algoritmo airPLS (ZHANG; CHEN; LIANG, 2010) com finalidade de remover a variação que não é inerente às amostras e suavização pelo filtro de Savitzky-Golay (janela de 5 pontos) (SAVITZKY; GOLAY, 1964). Em seguida, o conjunto de dados foi dividido em treinamento (70%, n=117) e teste (30%, n=50), pelo método Kennard Stone (KENNARD; STONE, 1969) garantindo a proporção existente entre as diferentes classes. Na etapa de otimização e construção dos modelos foram testados diversos métodos de pré-processamentos espectrais, tais como: padronização normal de sinal (SNV, do inglês *Standard Normal Variate*), normalização vetorial, 1ª e 2ª derivada, correção multiplicativa de espalhamento (MSC, do inglês *multiplicative scatter correction*), além de suas combinações. Os pré-processamentos foram utilizados para corrigir oscilações, minimizar ruídos instrumentais e diminuir sobreposições de bandas. Maiores informações a respeito do funcionamento desses métodos podem ser encontradas em Peris-Díaz e Krezel (2021). Aos dados pré-processados foram aplicados métodos de análise supervisionada para discriminação das classes como: análise linear discriminante com algoritmo genético (GA-LDA), análise linear discriminante pelos mínimos quadrados parciais (PLS-DA) e Floresta aleatória com peso de Fisher (PF-RF) utilizando 1000 árvores. O conjunto de treinamento foi utilizado para seleção de variáveis e construção dos modelos. Além disso, os modelos foram otimizados por meio de validação cruzada pelo método *venetian blinds cross-validation* K-fold (K=10). Todos os pré-processamentos e análise multivariada foram realizadas no software Matlab (2024-A).

2.4 PARÂMETROS DE AVALIAÇÃO DOS MODELOS

O conjunto de teste externo foi aplicado para avaliação do modelo pelos parâmetros de sensibilidade, especificidade, taxa de falso positivo e de falso negativo, exatidão e curva ROC. A sensibilidade é equivalente à taxa de verdadeiros positivos, ou seja, as amostras que são positivas e o modelo as classificou corretamente como positivas (equação 1). Já a especificidade é a taxa de verdadeiros negativos (equação 2). Considerando como classe alvo a classe de indivíduos com diagnóstico positivo para SARS-CoV-2, a especificidade é a taxa de pacientes com diagnóstico negativo que o modelo classificou corretamente como não positivo. A exatidão é a acurácia do modelo (equação 3). Por fim, para análise de qual modelo foi mais preditivo, elaborou-se uma curva ROC. A Curva ROC é uma forma gráfica de avaliar a qualidade do modelo

baseado na taxa de falso positivo e sensibilidade analisada pela área abaixo da curva (AUC) do gráfico. Dessa forma, quanto mais próxima de 1 a AUC melhor será a capacidade preditiva do modelo. Em contrapartida, quanto mais próximo a curva estiver da linha diagonal do gráfico a AUC fica em torno de 0,5, demonstrando uma aleatoriedade do modelo para classificação dos grupos.

$$Sensibilidade\ (\%) = \frac{VP}{VP + FN} \cdot 100 \quad (1)$$

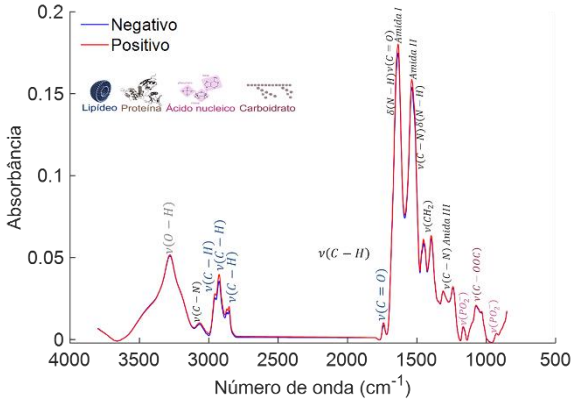
$$Especificidade\ (\%) = \frac{VN}{FP + VN} \cdot 100 \quad (2)$$

$$Exatidão\ (\%) = \frac{(VP + VN)}{VP + FN + FP + VN} \cdot 100 \quad (3)$$

4 RESULTADOS E DISCUSSÃO

O perfil espectral médio dos grupos positivo e negativo está apresentado na Figura 1. O perfil espectral apresenta bandas de absorção características de biofluidos (GULEKEN et al., 2022; BANDEIRA et al., 2022). Para este estudo, foi considerado o espectro ATR-FTIR total com o intuito de obter toda a informação do conjunto de dados, principalmente as regiões que representam ligações de lipídios, que tem sido associada a alterações causadas pela infecção COVID-19 (SONG et al., 2020; SPICK et al, 2021). No entanto, foi desconsiderada a região espectral entre 2800 cm^{-1} e 1800 cm^{-1} que não possui bandas de absorção característica para o biofluido utilizado (CALLERY; ROWBOTTOM, 2021) e a região das extremidades do espectro que devido a limitação de leitura do equipamento apresenta somente ruído. Dessa forma, o espectro final analisado possuía 950 variáveis.

Figura 1- Espectro ATR-FTIR médio pré-processado pela padronização normal de sinal (SNV) de conjunto de amostras negativas (n=76) e positivas (n=91) para classificação multivariada de COVID-19.

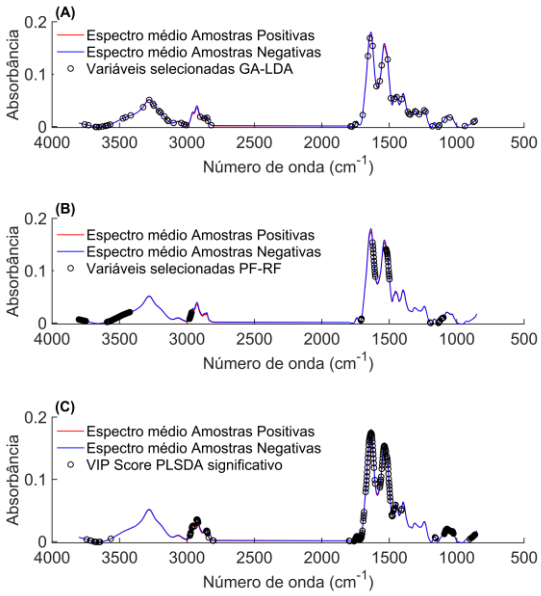


Fonte: elaborado pelos autores.

Neste estudo foram utilizados dois diferentes métodos de seleção de variáveis, Peso de Fisher (PF) (LOVATTI et al., 2020) e Algoritmo Genético (GA) (SHAFFER; SMALL, 1996), com o intuito de extrair as informações mais relevantes para construção de modelos de classificação com resultados satisfatórios O algoritmo de PF reduziu a dimensão das variáveis em 92% (76 variáveis), já o GA reduziu em 93% a dimensão dos dados (66 variáveis) para construção do modelo, o que foi coerente com as características de seleção pela aptidão a partir de uma função custo, do método GA, e de seleção pela maximização da variância inerente ao método de peso de Fisher. Enquanto no caso do PLS-DA, as variáveis são modeladas a partir da redução de dimensão e avaliadas quanto a sua influência no treinamento do modelo. A identificação das variáveis de maior relevância no PLS-DA foi realizada pela pontuação da importância variável de projeção (VIP). O VIP é um índice que mostra quanto uma variável contribuiu para construção do modelo discriminativo, e as variáveis são consideradas importantes se apresentarem pontuações de VIP>1 (CHONG; JUN, 2005).

Dentre as variáveis destacadas na Figura 2, que compreendem as regiões selecionadas em comum pelos métodos de seleção de variáveis (GA e PF) e pela análise VIP, observam-se as regiões de 3500 cm^{-1} – 3000 cm^{-1} relacionada ao estiramento da ligação N-H, atribuída a proteínas e também a estiramento de O-H atribuída as biomoléculas de lipídios e carboidratos; região de 3000 cm^{-1} – 2800 cm^{-1}, atribuída a ácidos graxos; regiões de 1700 cm^{-1} – 1512 cm^{-1} de ligação N-H atribuídas na literatura como banda de amida I (1700 cm^{-1} – 1600 cm^{-1}) e amida II (1595 cm^{-1} – 1512 cm^{-1}) correlacionada a proteínas; região de 1196 cm^{-1} – 1090 cm^{-1}, correlacionada a estiramento simétrico de grupos fosfatos, proteínas, DNA e RNA. Além disso, o GA selecionou ainda variáveis na região de 1780 cm^{-1} – 1750 cm^{-1} correlacionada aos lipídios, 1408 cm^{-1} – 1231 cm^{-1} atribuída a fosfolipídeos, aminoácidos e proteínas, descrita na literatura como região de amida III, mais precisamente na região de 1350 cm^{-1} – 1240 cm^{-1}. Ademais, a análise VIP realizada pelo PLS-DA destacou também as variáveis na região de deformação N-H, 1761 cm^{-1} – 1712 cm^{-1}, atribuída a proteínas e aminoácidos (NASEER; ALI.; QAZI, 2020).

Figura 2- Variáveis mais relevantes selecionadas para construção dos modelos de classificação. A) Algoritmo GA. B) Algoritmo PF. C) VIP >1 PLS-DA.



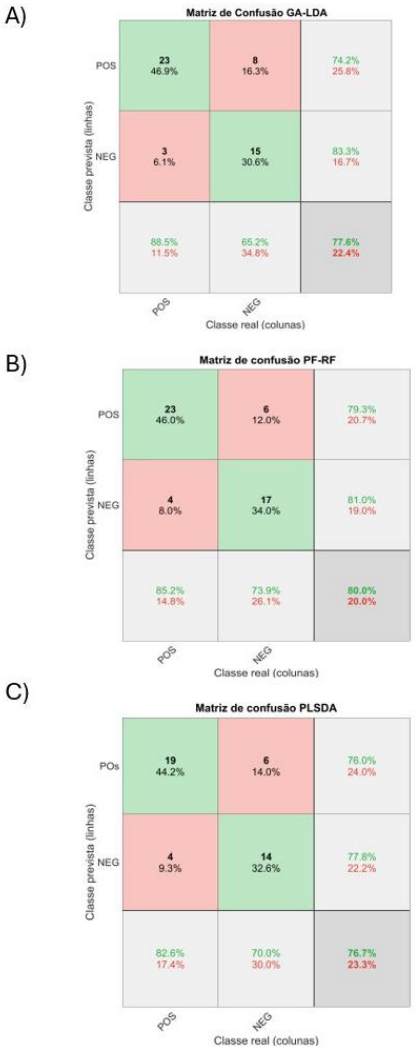
Fonte: elaborado pelos autores.

Ao analisar as regiões destacadas pelos 3 métodos (GA, PF e VIP-*score*) para construção do modelo de classificação para COVID-19, é possível destacar que estão correlacionadas principalmente as biomoléculas de lipídios, proteínas, carboidratos e ácidos graxos, sendo essas biomoléculas utilizadas para diferenciar o soro de pacientes com COVID-19 de pessoas com diagnóstico negativo. Há poucos estudos sobre evidências de uma relação entre os níveis de lipídeos e infecções por COVID-19 em biofluidos ou outras fontes (SONG et al., 2020; SPICK et al, 2021). Entretanto, a região de proteínas, principalmente amida I e amida II, tem sido retratada na literatura com grande importância para diagnóstico de COVID-19 GULEKEN et al., 2022; (BARAUNA et al., 2021). Guleken et al. (2022) argumentam sobre a importância da região de carboidratos para diferenciação dos grupos controle e doente, destacadas em seu estudo na região de 1390 cm^{-1} a 1421 cm^{-1}, devido à presença dessa biomolécula em estruturas que compõem o sistema imunológico

humano que é ativado contra o vírus, como a produção de anticorpos. Além disso, a informação adquirida pela espectroscopia ATR-FTIR das amostras de soro de maior relação às alterações fisiológicas em função da severidade da infecção pela COVID-19 tem sido associada à região espectral de 1718 cm^{-1} a 1788 cm^{-1}, que foi selecionada pelos modelos de seleção de variáveis GA e PF e pela análise VIP do PLS-DA, ademais, esta região foi reportada por Bandeira et al. (2022) como um marcador espectral do grau de glicosilação de IgG referente ao grau de gravidade da infecção por COVID-19.

O modelo de classificação PF-RF com SNV (BARNES; DHANOA; LISTER, 1989) como pré-processamento obteve 85,2% de sensibilidade e 73,9% de especificidade para o conjunto de teste externo para triagem clínica de COVID-19. O modelo GA-LDA também utilizou SNV como pré-processamento e apresentou 88,4% de sensibilidade e 65,2% de especificidade para discriminação dos grupos positivo e negativo para COVID-19 no conjunto de teste externo. Por fim, o modelo PLS-DA com SNV e 2ª derivada (SAVITZKY; GOLAY, 1964) demonstrou 82,6% de sensibilidade e 70% de especificidade no conjunto de teste externo. A matriz de confusão obtida para cada modelo permitiu a observação da classificação dos grupos feita pelos modelos, linhas da matriz, e a classificação real dos grupos obtida pelo imunoensaio, colunas da matriz (Figura 3). O GA-LDA apresentou uma taxa de falso-negativo de 6,1% (Figura 3a), para o PF-RF a taxa de falso-negativo foi de 8% (Figura 3b), o modelo PLS-DA obteve 9,3% (Figura 3c) tendo como classe alvo os positivos.

Figura 3- Matriz de Confusão dos modelos de classificação aplicados. A) GA-LDA. B) PF-RF. C) PLS-DA.



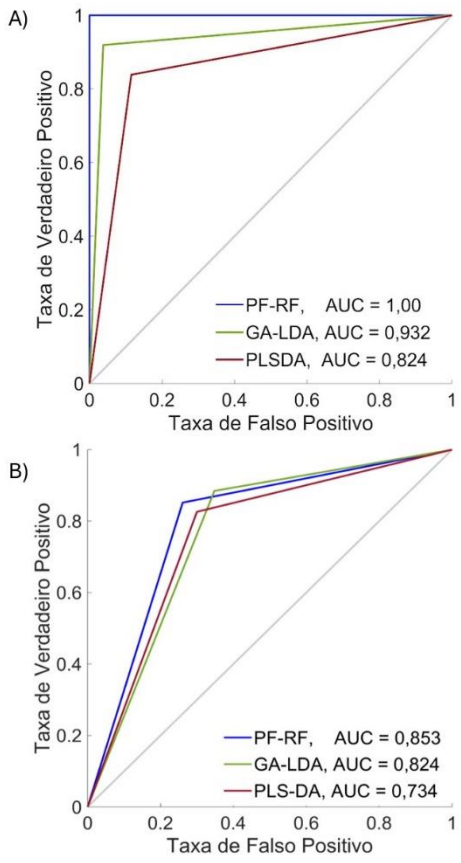
Fonte: elaborado pelos autores.

No contexto da COVID-19, o isolamento do indivíduo infectado torna-se crucial para a não disseminação da doença já que sua forma de transmissão ocorre em larga escala por meio do contato com gotículas de saliva/aerossóis de pessoas contaminadas com o SARS-CoV-2 (SALIAN et al., 2021). Dessa forma, um exame mais robusto e eficaz para a triagem clínica de COVID-19 deve focar em uma alta taxa de verdadeiros-positivos e uma baixa taxa de falsos-negativos, permitindo que os pacientes infectados sejam devidamente diagnosticados e possam realizar um rápido isolamento social

impedindo a maior disseminação da doença. No entanto, os parâmetros estatísticos devem ser analisados em conjunto para obter como resultado um modelo de classificação estatisticamente significativo e confiável. Assim, deve-se analisar a sensibilidade, especificidade e a taxa de falsos-negativos para que o modelo desenvolvido tenha uma melhor capacidade preditiva para ser utilizado no ambiente clínico.

Para comparar a qualidade dos modelos abordados neste estudo de forma resumida e estatística foi elaborada uma curva ROC.

Figura 4- Curva ROC dos modelos de classificação. A) Curva ROC grupo de treino. B) Curva ROC grupo de teste.



Fonte: elaborado pelos autores.

O modelo de PF-RF obteve uma AUC de 0,85, o PLS-DA uma AUC de 0,73 e o GA-LDA AUC de 0,82 (Figura 4b). Assim, sugere-se uma maior qualidade dos modelos GA-LDA e o PF-RF, pelos maiores valores de AUC, ambos modelos que utilizaram seleção de variáreis para a classificação dos grupos. Uma possível interpretação para esse fenômeno é o tipo de algoritmo utilizado. O PLS-DA realiza uma análise por meio do agrupamento de varáveis transformadas em variáveis latentes e constrói uma função linear para a classificação. O GA-LDA reduz as variáveis para a construção de um modelo linear por meio da análise individual de cada variável, partindo de um conjunto aleatório de variáveis. Já o PF-RF analisa as variáveis de forma individual, e inicialmente aleatória, sendo avaliada como preditora pelo índice de Gini durante construção de cada árvore, além de ser um método não paramétrico, ou seja, não tem uma função paramétrica definida para realizar a modelagem, que é direcionada pelas características dos dados (LOVATTI et al., 2019). Portanto, analisando as métricas do grupo de teste externo, o modelo PF-RF apresentou maior capacidade preditiva devido sua AUC de 0,85, sensibilidade de 85,2% e especificidade de 73,9%.

5 CONCLUSÃO & PERSPECTIVAS

O método PF-RF obteve uma sensibilidade de 85,2% e 73,9% de especificidade, o GA-LDA 88,4% de sensibilidade e 65,2% de especificidade de e o PLS-DA sensibilidade e especificidade de 82,6% e 70%, respectivamente para distinção das amostras de soro de pacientes sintomáticos infectados por COVID-19 e pacientes sintomáticos com diagnóstico negativo por imunoensaio. Dessa forma, é possível inferir que o método que obteve maior capacidade preditiva foi o PF-RF que também obteve uma baixa taxa de falso negativo (8%) que é uma métrica primordial para um bom teste diagnóstico para a COVID-19 devido sua forma de disseminação.

Além disso, é importante destacar que os métodos que obtiveram maior capacidade preditiva utilizaram algoritmo de seleção de variáveis para construção de modelo de classificação, PF-RF e GA-LDA, demostrando a relevância de utilizar esses algoritmos em dados espectroscópicos que possuem muitas variáveis que se correlacionam entre si, principalmente na análise de matrizes complexas como a biológica.

As regiões espectrais destacadas e utilizadas para discriminação dos grupos foi na região de ~3500 cm^{-1} a ~3000 cm^{-1}, ~3000 cm^{-1} a ~2800 cm^{-1}, ~1700 cm^{-1} a ~1600 cm^{-1}, ~1595 cm^{-1} a ~1512 cm^{-1}, ~1196 cm^{-1} a ~1090 cm^{-1}, atribuídas as macromoléculas de lipídeos, ácidos graxos, proteínas, carboidratos e ácidos nucléicos, respectivamente.

Dessa forma, a exemplo deste modelo construído para COVID-19, a bioespectroscopia é um método alternativo com potencialidade para contribuir em triagens clínicas no contexto de altas demandas e novas pandemias. Porém, este é um estudo inicial que necessita de maior aprofundamento com outros métodos de análise multivariada, modelos de classificação de consenso, modelos de ensemble que é uma metodologia de conjuntos de modelos de classificação, possíveis comparações estatísticas e identificação de fatores limitantes para a aplicabilidade do método proposto. Por fim, reforça-se a aplicabilidade da utilização do ATR-FTIR nas ciências biomédicas.

AGRADECIMENTOS

Os autores agradecem ao Laboratório Tommasi por ter cedido as amostras para realização deste estudo. Este estudo foi financiado pela Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Universal e PROFIX – n.691/2022, 1036/2022 e 2023-2NC6S), pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – n.4097/2022) e pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). VGB é Bolsista de produtividade CNPq.

3 REFERÊNCIAS

BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. **Analytical Methods**, v. 5, n. 16, p. 3790, 2013.

BANDEIRA, L.C. et al. Micro-Fourier-transform infrared reflectance spectroscopy as tools for probing IgG glycosylation in COVID-19 patients. **Scientific Reports**, v. 12, n. 1, 11 mar. 2022.

BARAUNA, V. G. et al. Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. **Analytical Chemistry**, v. 93, n. 5, p. 2950–2958, 22 jan. 2021.

BARNES, R. J.; DHANOA, M. S.; LISTER, S. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. **Applied Spectroscopy**, v. 43, n. 5, p. 772–777, jul. 1989.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRO, R.; SMILDE, A. K. Principal component analysis. **Anal. Methods**, v. 6, n. 9, p. 2812–2831, 2014.

BRUYNE S., et al. Applications of mid-infraredspectroscopy in the clinical laboratory setting. **Crit Rev Clin Lab Sci**, v. 55, n. 1, p. 1-20, 2018.

BRUN, B. F. et al. Fast screening using attenuated total reflectance- fourier transform infrared (ATR-FTIR) spectroscopy of patients based on D-dimer

threshold value. **Talanta**, v. 269, p. 125482, 1 mar. 2024.

CALLERY, E. L.; ROWBOTTOM, A. W. Vibrational spectroscopy and multivariate analysis techniques in the clinical immunology laboratory: a review of current applications and requirements for diagnostic use. **Applied Spectroscopy Reviews**, p. 1–30, 4 ago. 2021.

CHONG, I.-G.; JUN, C.-H. Performance of some variable selection methods when multicollinearity is present. **Chemometrics and Intelligent Laboratory Systems**, v. 78, n. 1-2, p. 103–112, jul. 2005.

CROCCO, M. A., et al. ATR-FTIR spectroscopy of plasma supported by multivariate analysis discriminates multiple sclerosis disease. **Scientific Reports**. v. 13, n. 1, 13 fev. 2023.

DEPCIUCH, J. et al. FTIR Spectroscopy of Cerebrospinal Fluid Reveals Variations in the Lipid: Protein Ratio at Different Stages of Alzheimer's Disease. v. 68, n. 1, p. 281–293, 1 jan. 2019.

FARIA, R. A. et al. Potential Role of Fourier Transform Infrared Spectroscopy as a Screening Approach for Breast Cancer. **Applied spectroscopy**, v. 77, n. 4, p. 405–417, 9 mar. 2023.

FOLLI, G. S. et al. Correlation analysis of modern analytical data – a chemometric dissection of spectral and chromatographic variables. **Analytical methods**, v. 15, n. 33, p. 4119–4133, 1 jan. 2023.

GIAMOUGIANNIS, P. et al. Detection of ovarian cancer (± neo-adjuvant chemotherapy effects) via ATR-FTIR spectroscopy: comparative analysis of blood and urine biofluids in a large patient cohort. **Analytical and Bioanalytical Chemistry**, v. 413, n. 20, p. 5095–5107, 1 jul. 2021.

GULEKEN, Z. et al. Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy and machine learning classifications. **Talanta**, v. 237, p. 122916, jan. 2022.

HOFFNER, G., et al. Synchrotron-based infrared spectroscopy brings to light the structure of protein aggregates in neurodegenerative diseases, **Rev. Anal. Chem**. 33 (2014) 231 e 243.

KENNARD, R. W.; STONE, L. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1, p. 137–148, 1 fev. 1969.

KHAN, R. S.; REHMAN, I. U. Spectroscopy as a Tool for Detection and Monitoring of Coronavirus (COVID-19). **Expert Review of Molecular Diagnostics**. Taylor and Francis Ltd July 2, 2020, pp 647–649.

LOVATTI, B. P. O. et al. Use of Random Forest in the identification of important variables. **Microchemical Journal**, v. 145, p. 1129–1134, mar. 2019.

LOVATTI, B. P. O. et al. Different strategies for the use of random forest in NMR spectra. v. 34, n. 12, 2 mar. 2020.

MANUEL DAVID PERIS-DÍAZ; ARTUR KRĘŻEL. A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry. **TrAC Trends in Analytical Chemistry** v. 135, p. 116157–116157, 1 fev. 2021.

NASCIMENTO, M. H. C. et al. Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. **Analytical Chemistry**, v. 94, n. 5, p. 2425–2433, 25 jan. 2022.

NASEER, K.; ALI, S.; QAZI, J. ATR-FTIR spectroscopy as the future of diagnostics: a systematic review of the approach using bio-fluids. **Applied Spectroscopy Reviews**, v. 56, n. 2, p. 85–97, 18 mar. 2020.

NOGUEIRA, M. S. et al. Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. **Scientific Reports**, v. 11, n. 1, 11 out. 2021.

OLIVEIRA, M. et al. TESTES DIAGNÓSTICOS PARA O SARS-COV-2: UMA REFLEXÃO CRÍTICA. **Química Nova**, 2022.

PALUSZKIEWICZ, C., et al. Saliva as first-line diagnostic tool: a spectral challenge for identification of cancer biomarkers, **J. Mol. Liq.** 307 (2020).

ROTHAN, H. A.; BYRAREDDY, S. N. The Epidemiology and Pathogenesis of Coronavirus Disease (COVID-19) Outbreak. **Journal of Autoimmunity**, v. 109, n. 102433, p. 102433, fev. 2020.

SALIAN, V. S. et al. COVID-19 Transmission, Current Treatment, and Future Therapeutic Strategies. **Molecular Pharmaceutics**, v. 18, n. 3, p. 754–771, 1 mar. 2021.

SANTOS, M. C. D.; MORAIS, C. L. M.; LIMA, K. M. G. ATR-FTIR spectroscopy for virus identification: A powerful alternative. **Biomedical Spectroscopy and Imaging**, p. 1–16, 9 jun. 2020.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627–1639, jul. 1964.

SHAFFER, R. E.; SMALL, G. W. Genetic algorithms for the optimization of piecewise linear discriminants. **Chemometrics and Intelligent Laboratory Systems**, v. 35, n. 1, p. 87–104, 1 nov. 1996.

SONG, J. W.et al. Omics-Driven Systems Interrogation of Metabolic Dysregulation in COVID-19 Pathogenesis. **Cell Metab** 2020, 32 (2), 188-202.e5.

SPICK, M.et al. Changes to the Sebum Lipidome upon COVID-19 Infection Observed via Rapid Sampling from the Skin. **EClinicalMedicine** 2021, 33, 100786.

YANG, X. et al. Diagnosis of Lung Cancer by ATR-FTIR Spectroscopy and Chemometrics. **Frontiers in Oncology**, v. 11, 30 set. 2021.

ZHANG, L. et al. Fast Screening and Primary Diagnosis of COVID-19 by ATR–FT-IR. **Analytical Chemistry**, v. 93, n. 4, p. 2191–2199, 11 jan. 2021.

ZHANG, Z.-M.; CHEN, S.; LIANG, Y.-Z. Baseline correction using adaptive iteratively reweighted penalized least squares. **The Analyst**, v. 135, n. 5, p. 1138, 2010.