Scientia Generalis 2675-2999

v. 2, n.1, p. 143-151. *2021.*

Artigo Original

DETECÇÃO DO VÍRUS DA DENGUE POR ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO EM CONJUNTO COM ALGORITMO GENÉTICO E ANÁLISE DISCRIMINANTE LINEAR: um modelo quimiométrico

Gigliane Joice Santos da Silva [a], Marcos Claudino Batista dos Santos Filho [b], Paulo Venicius Messias dos Santos [c], Ricardo Fernandes dos Santos [d], Marfran Claudino Domingos dos Santos [ef*]

[a] *Universidade Federal do Rio Grande do Norte, Centro do Biociências, Natal, Rio Grande do Norte, Brasil.*

[b] *Universidade Federal do Rio Grande do Norte, Departamento de Física Teórica e Experimental, Natal, Rio Grande do Norte, Brasil.*

[c] *Instituto Federal de Ciência e Tecnologia do Rio Grande do Norte – Campus Pau dos Ferros, Pau dos Ferros, Rio Grande do Norte, Brasil.*

[d] *Instituto Federal de Ciência e Tecnologia do Rio Grande do Norte – Campus Ipanguaçu, Ipanguaçu, Rio Grande do Norte, Brasil.*

[e] *Instituto Federal de Ciência e Tecnologia do Sertão Pernambucano – Campus Floresta, Floresta, Pernambuco, Brasil.*

[f] *Universidade Federal de Rio Grande do Norte - Instituto de Química, Natal, Rio Grande do Norte, Brasil*

Resumo

O objetivo deste estudo foi desenvolver um modelo quimiométrico baseado na aplicação do algoritmo computacional GA-LDA na análise de dados de infravermelho médio para detectar o vírus da dengue a partir de um conjunto de espectros de amostras clínicas. Nessa abordagem, o GA seleciona as variáveis mais importantes no conjunto de espectros, enquanto o LDA trabalha na discriminação entre as classes, baseando-se nestas variáveis selecionadas. O estudo investigou as vantagens desta abordagem frente às técnicas padrão, uma vez que que existe a necessidade da existência de uma técnica que permita unir uma resposta rápida com custo baixo e boa especificidade. O conjunto de dados utilizados neste estudo foi disponibilizado cordialmente pelo Grupo de Pesquisa em Química Biológica e Quimiometria da UFRN. O modelo foi construído com o uso do software "MATLAB", onde foi feita a análise computacional e cálculos de medidas de qualidade (figuras de mérito). Constatou-se que a técnica conseguiu uma boa previsão, com 100% de sensibilidade e especificidade, da classe correta de todas as amostras utilizadas no conjunto teste. Além disso, demonstrou-se não precisar do uso de reagentes nem de kits para a análise, além de fornecer resultados mais rápidos frente às técnicas padrão utilizadas. A partir destes resultados foi possível verificar a importância da aplicação de uma ferramenta computacional no campo da virologia, uma vez que foi observado um grande potencial para esta aplicação. Contudo, são necessários estudos mais aprofundados.

Palavras-chave: Detecção da Dengue; Algoritmos computacionais; Reconhecimento de padrões; GA-LDA.

DENGUE VIRUS DETECTION BY MIDDLE INFRARED SPECTROSCOPY IN CONJUNCTION WITH GENETIC ALGORITHM AND LINEAR DISCRIMINANT ANALYSIS: A CHEMOMETRIC MODEL

[*] Autor correspondente - profmarfransantos@gmail.com

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

Abstract

The aim of this study was to develop a chemometric model based on the application of the GA-LDA computational algorithm in the analysis of mid-infrared data to detect dengue virus from a set of spectra of clinical samples. In this approach, the GA selects the most important variables in the set of spectra, while the LDA. works on the discrimination between classes, based on these selected variables. The study investigated the advantages of this approach compared to standard techniques, since there is a need for a technique that allows for a quick response with low cost and good specificity. The data set used in this study was cordially made available by the Research Group on Biological Chemistry and Chemometrics at UFRN. The model was built using the "MATLAB" software, where computational analysis and calculations of quality measures (figures of merit) were performed. It was found that the technique achieved a good prediction, with 100% sensitivity and specificity, of the correct class of all samples used in the test set. Furthermore, it has been shown not to need the use of reagents or kits for the analysis, in addition to providing faster results compared to the standard techniques used. From these results, it was possible to verify the importance of applying a computational tool in the field of virology, since a great potential for this application was observed. However, further studies are needed.

Keywords: Dengue detection; Computational algorithms; Pattern Recognition; GA-LDA.

DETECCIÓN DEL VIRUS DEL DENGUE MEDIANTE ESPECTROSCOPÍA INFRARROJA MEDIA EN CONJUNCIÓN CON ALGORITMO GENÉTICO Y ANÁLISIS DISCRIMINANTE LINEAL: un modelo quimiométrico

Resumen

El objetivo de este estudio fue desarrollar un modelo quimiométrico basado en la aplicación del algoritmo computacional GA-LDA en el análisis de datos de infrarrojo medio para detectar el virus del dengue a partir de un conjunto de espectros de muestras clínicas. En este enfoque, el GA selecciona las variables más importantes en el conjunto de espectros, mientras que el LDA trabaja en la discriminación entre clases, en función de estas variables seleccionadas. El estudio investigó las ventajas de este enfoque en comparación con las técnicas estándar, ya que existe la necesidad de una técnica que permita una respuesta rápida con bajo costo y buena especificidad. El conjunto de datos utilizado en este estudio fue puesto a disposición cordialmente por el Grupo de Investigación en Química Biológica y Quimiometría de la UFRN. El modelo se construyó utilizando el software "MATLAB", donde se realizaron análisis computacionales y cálculos de medidas de calidad (cifras de mérito). Se encontró que la técnica logró una buena predicción, con 100% de sensibilidad y especificidad, de la clase correcta de todas las muestras utilizadas en el conjunto de prueba. Además, se ha demostrado que no necesita el uso de reactivos o kits para el análisis, además de proporcionar resultados más rápidos en comparación con las técnicas estándar utilizadas. A partir de estos resultados se pudo constatar la importancia de aplicar una herramienta computacional en el campo de la virología, ya que se observó un gran potencial para esta aplicación. Sin embargo, se necesitan más estudios.

Palabras llave: Detección del dengue; Algoritmos computacionales; Reconocimiento de patrones; GA-LDA.

1. Introdução

Os vírus são os agentes infecciosos que atingem o maior número de pessoas no mundo. A chegada de alguns vírus em locais antes livres da presença destes, representa um potencial desafio para a saúde pública em muitos aspectos. Visto que toda a população é susceptível à infecção e não há vacinas disponíveis nem antivirais para o tratamento, a circulação destes vírus significa um problema considerável[1].

Uma constante ameaça em regiões tropicais são os arbovírus, devido às mudanças climáticas aceleradas, desmatamentos, migração da população e precariedade nas condições sanitárias que permitem a amplificação e transmissão dos vírus. O Brasil está situado em uma área, em sua grande parte, tropical, tendo, assim, locais propícios para a existência do vetor e, portanto, para a incidência de arboviroses. Dentre os arbovírus emergentes e reemergentes no Brasil citam-se:

Dengue, Zika, Febre Amarela, da família *Flaviviridae*; e Chikungunya da família *Togaviridae*, entre outros[2]. Dentre estes, o vírus da Dengue é, indubitavelmente, o de maior importância do ponto de vista endêmico.

O vírus da Dengue se apresenta em quatro sorotipos: DENV-1, DENV-2, DENV-3 e DENV-4, ressalta-se que são o mesmo vírus, porém distinguem-se por darem diferentes respostas a diferentes anticorpos. O vírus da Dengue é composto por proteínas de superfície: proteínas E e M (na figura 1; envelope viral, formado por uma bicamada lipídica); nucleocapsídeo (uma membrana formada por proteínas responsável por proteger o RNA); e o RNA que é a informação genética do vírus[1,2].

Com relação ao seu diagnóstico, classificam-se em dois grupos os métodos de diagnósticos: os diretos e os indiretos. Os métodos diretos são os mais específicos, porém, são mais custosos, precisam de mais tempo para fornecer os resultados e são menos acessíveis à população, são encontrados principalmente em hospitais ou clínicas de ponta e centros de estudo. Por outro lado, os métodos indiretos são menos específicos, porém, são mais baratos, fornecem resultados com maior rapidez e são encontrados facilmente em clínicas de diagnóstico[3, 4].

Acredita-se que em função do alto custo dos métodos diretos, as metodologias utilizadas nas rotinas de diagnóstico da maioria das unidades de saúde pública são do grupo dos métodos indiretos, ou seja, técnicas inespecíficas. Dessa forma, faz-se necessário uma técnica que consiga combinar resposta rápida, baixo custo e especificidade, sendo portanto uma justificativa plausível para pesquisas sobre métodos alternativos.

Assim sendo, neste trabalho tratou-se sobre uma nova abordagem para a detecção deste vírus através de métodos computacionais que têm sido aplicados na detecção de Dengue em amostras biológicas a partir de espectros de infravermelho médio destas amostras. Especialmente, apresentaram-se as vantagens de dois algoritmos que têm sido implementados em conjunto em dados de infravermelho médio: o Algoritmo Genético (GA) e Algoritmo da Análise Discriminante Linear (LDA). Por fim, foi realizada uma aplicação do GA-LDA na discriminação entre espectros de sangue saudável e infectados com Dengue.

2. Material e Métodos

2.1. Dados Espectroscópicos

Os espectros de infravermelho das amostras saudáveis e com dengue foram disponibilizados pelo grupo de pesquisa em Química Biológica e Quimiometria da UFRN, que forneceu gentilmente os dados espectroscópicos. O conjunto de dados foi formado por 16 espectros de amostras de sangue de pessoas saudáveis e 16 espectros de amostra de sangue de pessoas infectadas pelo vírus da dengue. As amostras são originais de pessoas do Rio Grande do Norte que passaram por diagnóstico para dengue através do RT-PCR.

2.2. Software

Os dados foram importados para o MATLAB R2014b (Math- works, Natick, EUA)[5], que é um software interativo de alta performance voltado para o cálculo numérico, podendo ser utilizado em estudos envolvendo cálculo com matrizes e processamento de sinais, onde ocorreu um pré-tratamento e o modelo de classificação foi aplicado.

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

2.3. Análise Computacional

Os dados das amostras foram pré-tratados com ajustamento de linha de base e suavização: mudanças físicas no ambiente tais como temperatura, anomalias no equipamento, pressão, etc, podem ser traduzidos como desvios na linha de base dos espectros ou aspecto ruidoso, por isso, é preciso aplicar pré-processamentos que eliminem estas informações físicas. Então, o Algoritmo Genético foi aplicado em conjunto com Análise Discriminante Linear (GA-LDA)[5].

O algoritmo clássico de amostragem uniforme de Kennard-Stone[5] foi implementado para separar as amostras em treinamento (6 amostras de cada classe), validação (6 amostras de cada classe) e teste (4 amostras de cada classe).

2.4. Medidas de Qualidade

O estudo foi validado a partir dos cálculos da sensibilidade e especificidade, que foram calculadas conforme as equações (1) e (2):

$$Sens = \frac{VP}{VP + FN} \times 100 \quad (1)$$

$$Esp = \frac{VN}{VN + FP} \times 100 \quad (2)$$

Onde VP é verdadeiro positivo, FP é falso positivo, VN é verdadeiro negativo e FN é falso negativo.

3. Resultados e Discussões

Na figura 1 encontram-se os espectros brutos, cortados e pré-processados das amostras de sangue saudáveis e com dengue. Como pode ser notado, existe uma grande similaridade entre os espectros, o que limita sua diferenciação visual. Por esse motivo, é necessário usar o GA-LDA, onde o GA consegue encontrar as variáveis que mais possibilitem o LDA discriminar as amostras em suas respectivas classes.

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

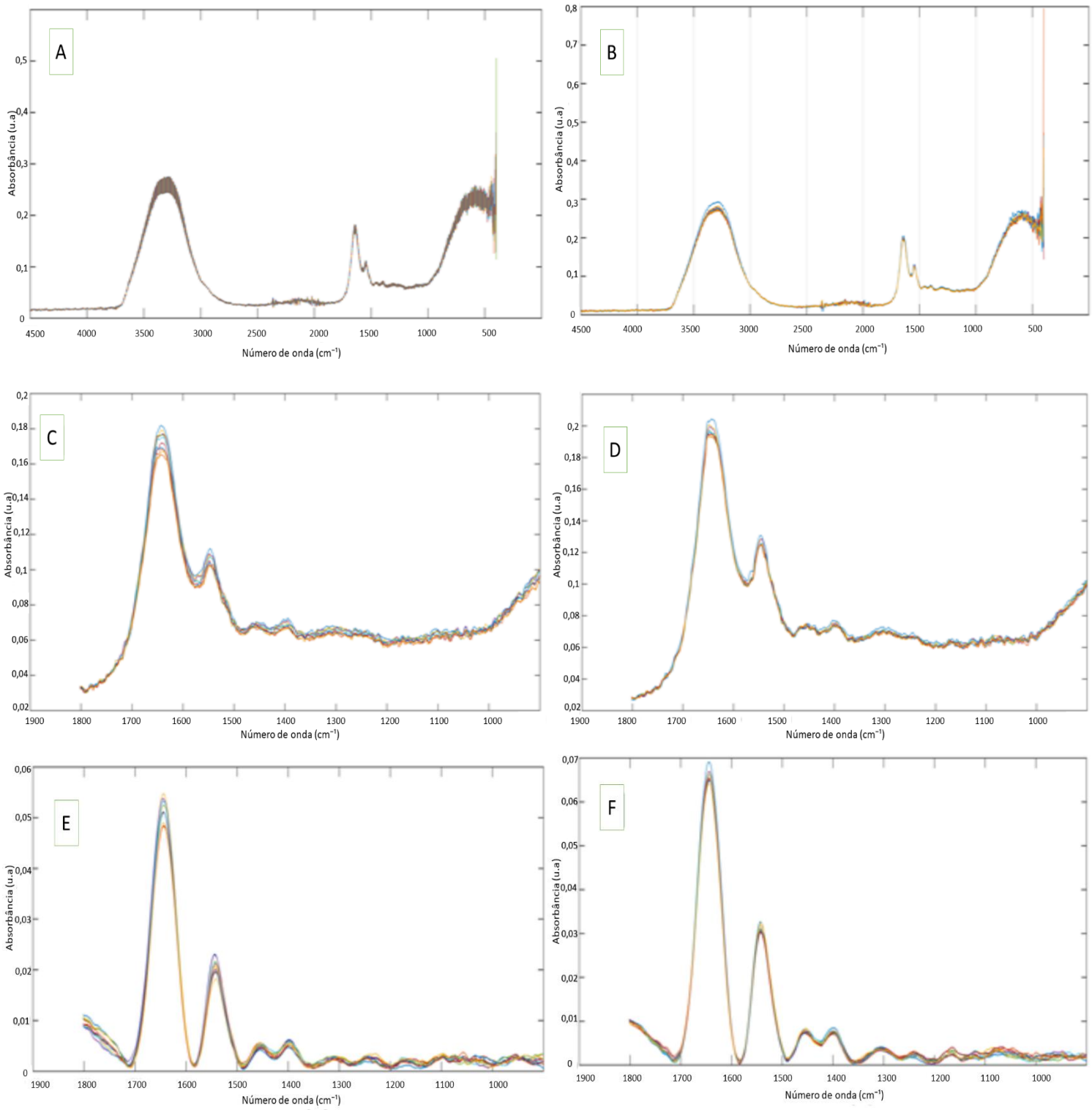

Fig. 1. Espectros brutos das amostras de sangue (a) saudável e (b) com dengue; Espectros cortados na faixa de impressão digital de amostras biológicas (1800-900 cm-1) das amostras de (c) sangue saudável e (d) com dengue; Espectros cortados e pré-processados das amostras de e) sangue saudável e (f) com dengue. Fonte: dados da pesquisa. (2019)

Na figura 2, pode-se ver a média dos espectros cortados e pré-processados para as duas classes, onde se observa

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

que a média do espectro da classe com dengue tem maior intensidade de absorbância que a média do espectro da classe saudável. Isto, provavelmente, deve ocorrer pela presença do vírus, que altera as espécies químicas presentes no meio, assim como a intensidade de radiação absorvida, acarretando uma maior absorbância.

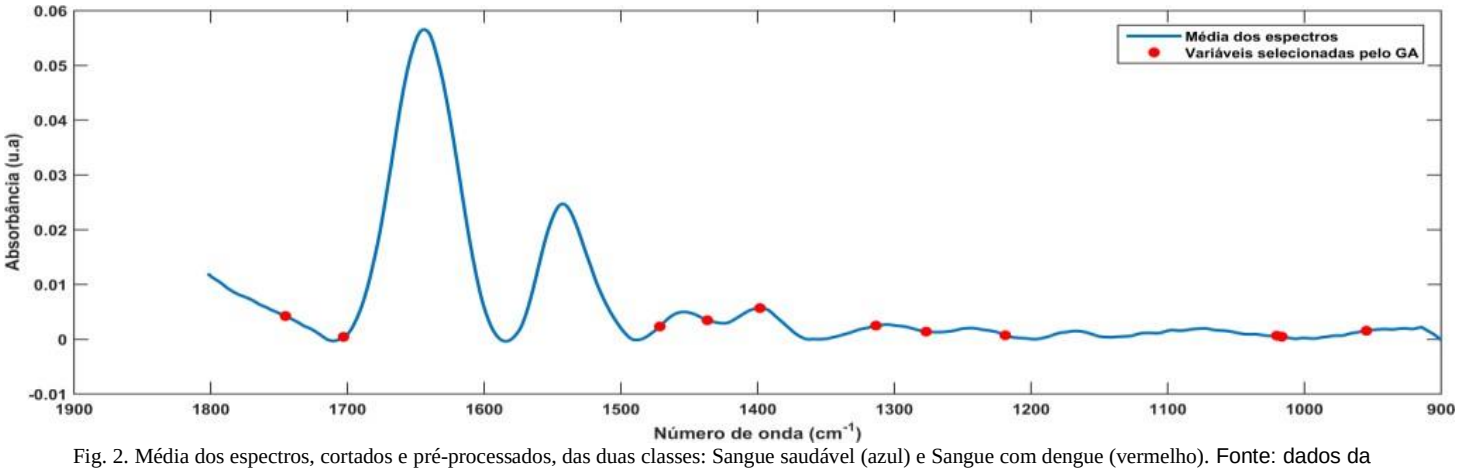

Fig. 2. Média dos espectros, cortados e pré-processados, das duas classes: Sangue saudável (azul) e Sangue com dengue (vermelho). Fonte: dados da pesquisa. (2019)

Ao aplicar o GA no conjunto de dados, observou-se a seleção de onze variáveis (números de onda). As variáveis selecionadas por GA podem ser vistas no quadro 1 e na figura 3. Estas variáveis podem ser compreendidas como "marcadores biológicos". As variáveis representam as informações que, segundo o GA, mais distinguem uma classe de outra.

Quadro 1 – Variáveis selecionadas pelo Algoritmo Genético.

Variáveis Selecionadas (cm^{-1})	1745 1703 1471 1437 1398 1313 1276 1219 1020 1016 954

Fonte: dados da pesquisa. (2019)

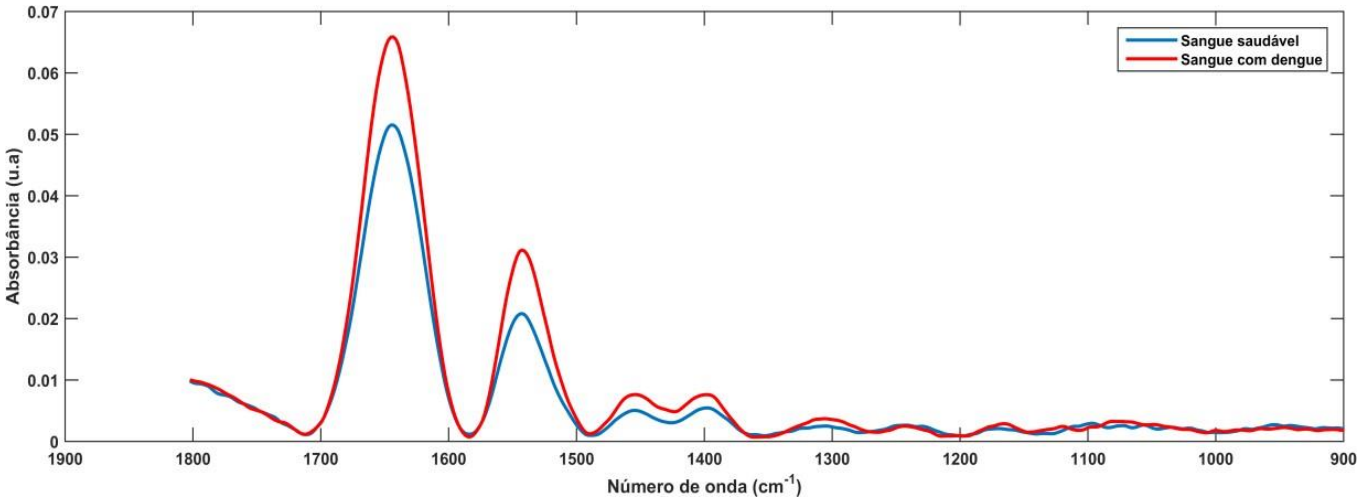

Fig. 3. Variáveis selecionadas pelo GA para serem usadas na discriminação. Fonte: dados da pesquisa. (2019)

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

Ao analisar as variáveis selecionadas pelo GA, observou-se variáveis relacionadas a vibrações associadas com estiramento em lipídios (~1745 cm^{-1}), Proteínas: (~1276 cm^{-1}, ~954 cm^{-1}), e RNA/DNA: (~1219 cm^{-1}).

Sabendo que o vírus da dengue é uma partícula composta por proteína, lipídio e RNA, pode-se inferir que o algoritmo identificou informações espectrais referentes ao vírus, que estavam presentes nas amostras infectadas, mas que não estavam presentes nas amostras não infectadas, e, a partir destas informações, o LDA fez a classificação, que pode ser analisada através do gráfico da função discriminante em função das amostras, mostrado na figura 4.

Como é apresentado, há uma considerável segregação das duas classes, com os seus respectivos pontos em regiões específicas do espaço, o que confere uma inegável classificação obtida pela combinação dos algoritmos GA-LDA.

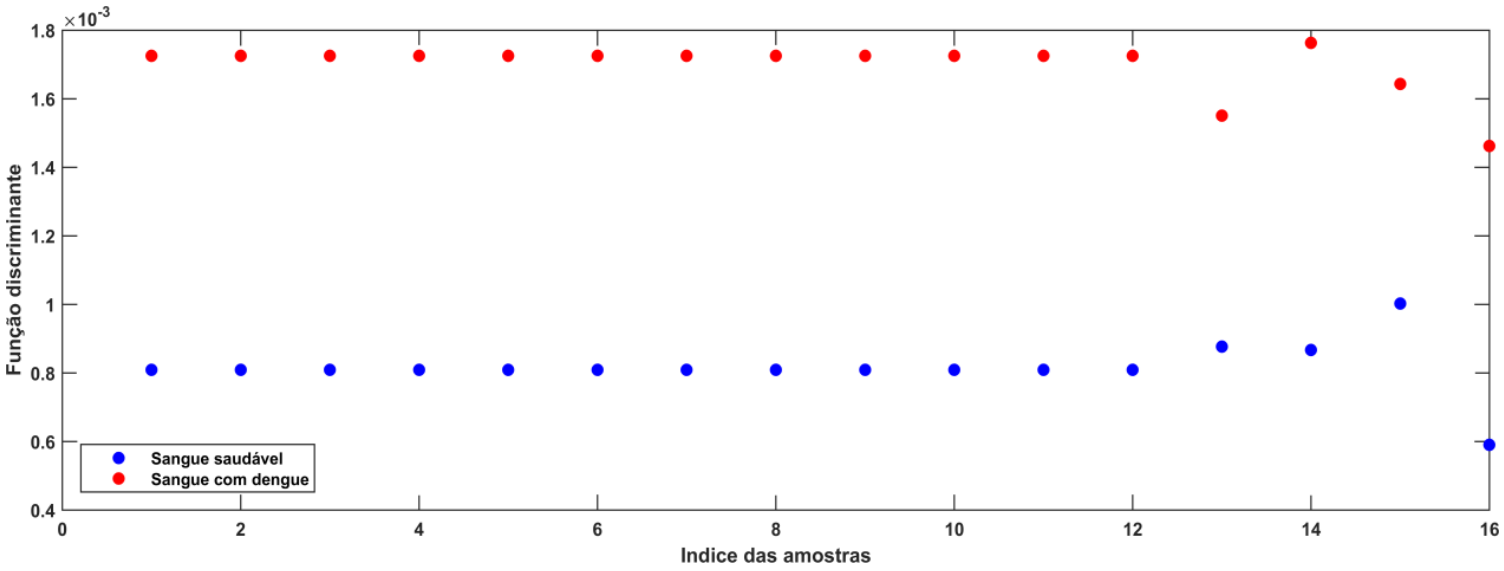

Fig. 4. Plote da função discriminante em função dos índices das amostras do grupo de teste (DF plot). Fonte: dados da pesquisa. (2019)

A partir dos valores de verdadeiro positivo e negativo e de falso positivo e negativo obtido na etapa de teste, foram calculadas a sensibilidade e especificidade, onde foram encontrados valores de 100% para ambas. Isto quer dizer que a técnica conseguiu prever com 100% todas as amostras utilizadas no conjunto teste.

A título de comparação, o teste rápido, uma das técnicas mais utilizadas no diagnóstico de dengue em clínicas e hospitais, possui sensibilidade e especificidade média na faixa de 70%, bem inferior ao encontrado neste estudo.

Outras vantagens que devem ser destacadas é o fato de que, para a técnica de Infravermelho médio, não se faz necessário o uso de reagentes nem kits para a análise, diferentemente de PCR, por exemplo, que se faz necessário o uso de primers para o diagnóstico de vírus. Ou seja, a análise é muito simples e ainda demanda uma quantidade de amostra muito menor, o que faz com que a abordagem da espectroscopia com ferramentas computacionais tenha um grande potencial de aplicação no campo da virologia. Além de tudo isso, a técnica fornece resultados mais céleres (cerca de 15 segundos) que as técnicas utilizadas como padrão.

Scientia Generalis, v. 2, n. 1, p. 143-151. 2021.

4. Proposta de uma Nova Ferramenta Computacional para Diagnosticar a Dengue

De acordo com os resultados obtidos neste trabalho e em outros trabalhos já publicados na literatura[6,7], propõe-se uma nova ferramenta computacional para o diagnóstico da Dengue. Esta ferramenta seria composta por Espectrofotômetro. Neste caso, sugere-se o uso de um espectrofotômetro de infravermelho próximo, pois seus componentes eletrônicos permitem a obtenção de instrumentos portáteis.

Nesse cenário, o espectrofotômetro faz a aquisição dos espectros das amostras e os enviariam diretamente para um computador acoplado, onde seria feito automaticamente todo o trabalho computacional como pré-processamento, e classificação, descritos neste trabalho. Finalmente, o software fornecerá um resultado positivo ou negativo para a Dengue. Estima-se que tudo isso poderia ser feito em aproximadamente um minuto.

Portanto, apresenta-se a possibilidade de uma técnica mais rápida, mais barata (não necessita da compra de reagentes ou kits de diagnósticos), e mais sensível e especifica. A figura 5 resume bem esta proposta de uma nova ferramenta para diagnóstico viral.

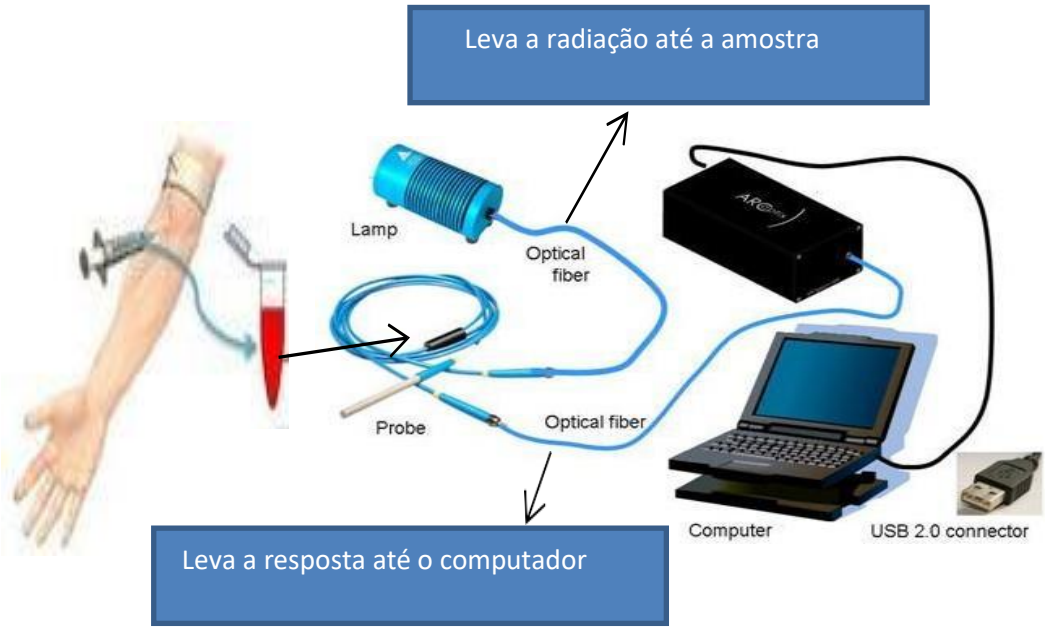

Fig. 5. Esquema da proposta de uma nova ferramenta computacional para o diagnóstico da Dengue. Fonte: Chemometrics UFRN[8]

5. Conclusão

Neste artigo pôde-se observar a relevância da aplicação de uma ferramenta computacional no campo da virologia. Um algoritmo de seleção de variáveis (GA) foi utilizado em conjunto com um algoritmo de classificação multivariada (LDA) para discriminar entre amostras de sangue sem a presença de vírus dengue, e com a presença do vírus.

Os resultados foram animadores e demonstraram o grande potencial desta abordagem, em que valores de 100%

de sensibilidade e especificidade foram obtidos. Estes valores são superiores a aqueles obtidos por algumas técnicas utilizadas como padrão. A ferramenta utilizada também necessita de menos amostra (cerca de 5 uL), e o resultado é mais rápido.

Aqui, demonstrou-se apenas uma aplicação breve, com um número de amostras menor, de uma ferramenta computacional que pode, em um futuro próximo, ser utilizada a favor da sociedade, demonstrando também, um dos inúmeros benefícios que a informática ainda pode trazer.

Referências

1. LIMA-CAMARA, T. N.. Arboviroses emergentes e novos desafios para a saúde pública no Brasil. **Revista de Saúde Pública**, [sl], v. 50, n. 36, p. 10-20, 2016

2. LOPES, N., NOZAWA, C., LINHARES, R. E. C.. Características gerais e epidemiologia dos arbovírus emergentes no Brasil. **Rev Pan-Amaz Saúde**, [sl], v. 5, n. 3, p. 55-64, 2014

3. PAHO. **Tool for the diagnosis and care of patients with suspected arboviral diseases**. 2017. Disponível em: http://iris.paho.org/xmlui/bitstream/handle/123456789/33895/9789275119365_eng.pd fsequence=1&isAllowed=y. Acesso em: 15/02/2021.

4. PEELING, R. W. *et al*. Evaluation of diagnostic tests: dengue. **Nature Reviews – Microbiology**, [sl], v. 8, n. 12, p. S30-S37, 2010.

5. KENNARD, R. W.; STONE, L. A. Projeto de experimentos auxiliado por computador. **Technometrics**, v. 11, n. 1, pág. 137-148, 1969.

6. SANTOS, M. C. D. *et al*. Spectroscopy with computational analysis in virological studies: A decade (2006-2016). **Trends in Analytical Chemistry**. [sl], v. 97, p. 244:256, 2017

7. SANTOS, M. C. D. *et al.* ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dengue vs. chikungunya vs. zika clinical samples. **Analytical Methods**. [sl], v. 10, p. 1280:1285, 2018.

8. CHEMOMETRICSUFRN.**Química Biológica e Quimiométrica** - Natal – Brasil Química Biológica e Quimiometria, 2020. Disponivel em: http://www.chemometricsufrn.org/. Acesso em: 23/03/2021.