



**UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS DA SAÚDE E
BIOLÓGICAS**

BRUNO FONSECA OLIVEIRA COELHO

**USO DA ESPECTROSCOPIA VIS-NIR NA DETECÇÃO DO
SARS-COV-2 COMO TÉCNICA AUXILIAR DE DIAGNÓSTICO DA
COVID-19**

**PETROLINA-PE
2023**

BRUNO FONSECA OLIVEIRA COELHO

**USO DA ESPECTROSCOPIA VIS-NIR NA DETECÇÃO DO
SARS-COV-2 COMO TÉCNICA AUXILIAR DE DIAGNÓSTICO DA
COVID-19**

Trabalho apresentado à Universidade Federal do Vale do São Francisco - UNIVASF, Campus Petrolina, como requisito da obtenção do grau de mestre em Ciências da Saúde e Biológicas.

Orientador: Prof. Dr. Rodrigo Pereira Ramos
Coorientador: Prof. Dr. Rodrigo Feliciano do Carmo

PETROLINA-PE

2023

Coelho, Bruno Fonseca Oliveira
C672u Uso da espectroscopia Vis-NIR na detecção do SARS-CoV-2
 como técnica auxiliar de diagnóstico da Covid-19 / Bruno Fonseca
 Oliveira Coelho. – Petrolina-PE, 2023.
 xix, 65 f.: il.; 29 cm.

Dissertação (Mestrado em Ciências da Saúde e Biológicas) -
Universidade Federal do Vale do São Francisco, Campus Petrolina-
PE, 2023

Orientador: Prof. Dr. Rodrigo Pereira Ramos.

Inclui referências.

1. Covid-19 - Diagnóstico. 2. Espectroscopia Vis-NIR. 3.
Detecção automática. 4. Aprendizado de máquina, I. Título. II.
Ramos, Rodrigo Pereira. III. Universidade Federal do Vale do São
Francisco.

CDD 539.60287

UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
PÓS-GRADUAÇÃO CIÊNCIAS DA SAÚDE E BIOLÓGICAS

FOLHA DE APROVAÇÃO

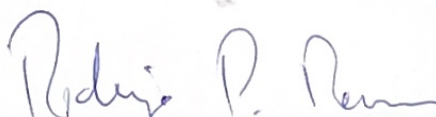
BRUNO FONSECA OLIVEIRA COELHO

**USO DA ESPECTROSCOPIA VIS-NIR NA DETECÇÃO DO SARS-COV-2 COMO
TÉCNICA AUXILIAR DE DIAGNÓSTICO DA COVID-19**

Dissertação apresentada como requisito para obtenção do título de Mestre em Ciências com ênfase na linha de pesquisa: Fundamentação Conceitual e Metodologias Inovadoras Integradoras em Ambiente, Tecnologia e Saúde, pela Universidade Federal do Vale do São Francisco.

Aprovada em: 26 de janeiro de 2023

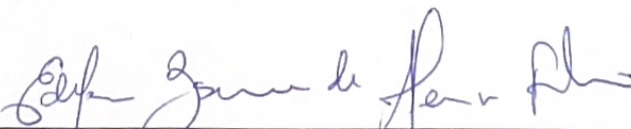
Banca Examinadora



Rodrigo Pereira Ramos, Doutor
Universidade Federal do Vale do São Francisco – Univasf



Rosalvo Ferreira de Oliveira Neto, Doutor
Universidade Federal do Vale do São Francisco – Univasf



Edilson Beserra de Alencar Filho, Doutor
Universidade Federal do Vale do São Francisco – Univasf

DEDICATÓRIA

Dedico este trabalho a toda minha família.

AGRADECIMENTOS

Gostaria de agradecer aos meus amigos Alanderson, João Vitor, Yalle e Juliana (que chegou no meio do caminho) por estarem sempre comigo desde os tempos de colégio e que eu sei que continuarão por perto.

Agradeço aos amigos que fiz durante a graduação, Eugênio, Kevin e Welton. Agradeço ao meu amigo Rodrigo Ramos por me inspirar a ser um profissional e ser humano cada vez melhor. Agradeço principalmente a Bia, por compartilhar muitos momentos bons e difíceis, e por me apoiar durante esses 5 anos de graduação.

Agradeço a toda a minha família: avós, tios, padrinhos, primos. Em especial, sou grato ao meu primo/irmão Segundo, e as minhas irmãs Bruna, Karla e Kátia, por todo amor e carinho.

Acima de tudo, sou grato aos meus pais: Adonis e Marluce, por sempre incentivarem a minha educação desde criança. Lembro-me que liam livros pra mim em voz alta, sempre que eu tinha preguiça de estudar. Muito obrigado, vocês me ensinaram a não desistir!

Agradeço à Universidade Federal do Vale do São Francisco, ao Programa de Pós-Graduação Ciências da Saúde e Biológicas e a todos os professores, técnicos e funcionários que contribuíram para a minha formação.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Não são as coisas que perturbam as pessoas, mas os pareceres a respeito das coisas"

Epicteto

RESUMO

A pandemia causada pela Covid-19 ainda está presente ao redor de todo o mundo. Apesar dos avanços no combate a doença como o desenvolvimento da vacina, a identificação de indivíduos contaminados ainda é essencial para otimizar o controle da transmissão do vírus entre humanos. A principal técnica para realizar a detecção do vírus, conduzindo a um diagnóstico, é o método RT-qPCR que, embora tenha um alto custo financeiro, possui uma alta acurácia na detecção do novo coronavírus. Diante disso, um método que seja capaz de realizar o diagnóstico de forma rápida, precisa e barata se faz necessário. Assim, este trabalho teve como objetivo analisar a viabilidade de uma nova técnica para identificação do SARS-CoV-2 através do uso de espectroscopia óptica na faixa do visível e do infravermelho próximo (Vis-NIR) em conjunto com algoritmos de aprendizado de máquina. Os sinais espectrais foram obtidos de amostras de *swab* nasofaríngeo previamente analisadas através do método RT-qPCR. Os exemplares foram fornecidos pelo Laboratório de Diagnóstico Molecular de Covid-19 da Univasf. Ao todo, 677 amostras foram analisadas, sendo 84 testadas positivo e 593 testadas negativo para Covid-19. Na sequência, técnicas de processamento digital de sinais, como filtros do tipo Savitzky-Golay, transformadas tempo-frequenciais e métodos estatísticos, foram utilizadas para eliminar dos dados originais elementos indesejados e extrair características relevantes. Algoritmos de aprendizado de máquina supervisionado como o SVM, *Random Forest* e o classificador do tipo *Naive Bayes* foram usados para realizar a identificação automática das amostras. Para avaliar o desempenho dos modelos, foi utilizada a técnica de validação cruzada *5-fold*. Com a metodologia proposta, foi possível alcançar uma acurácia de 93%, uma sensibilidade de 96% e uma especificidade de 89%, além de uma área sob a curva ROC de 0,95, na identificação de amostras de *swab* nasofaríngeo de indivíduos previamente diagnosticados. A partir desses resultados, se pode concluir que a espectroscopia Vis-NIR é uma técnica promissora para o diagnóstico do SARS-CoV-2.

Palavras-chave: Covid-19, espectro Vis-NIR, aprendizado de máquina, detecção automática

ABSTRACT

The pandemic caused by Covid-19 is still present around the world. Despite advances in combating the disease, such as vaccine development, identifying infected individuals is still essential to optimize the control of human-to-human transmission of the virus. The main technique for detecting the virus, leading to a diagnosis, is the RT-qPCR method, which, although it has a high financial cost, has a high accuracy in detecting the new coronavirus. In view of this, a method that is capable of performing the diagnosis quickly, accurately and inexpensively is necessary. Thus, this work aimed to analyze the feasibility of a new technique for identifying SARS-CoV-2 through the use of optical spectroscopy in the visible and near infrared range (Vis-NIR) in set with machine learning algorithms. Spectral signals were obtained from nasopharyngeal swab samples previously analyzed using the RT-qPCR method. The specimens were provided by the Molecular Diagnosis Laboratory of Covid-19 at Univasf. In all, 677 samples were analyzed, with 84 tested positive and 593 tested negative for Covid-19. Next, digital signal processing techniques, such as Savitzky-Golay filters, time-frequency transforms and statistical methods, were used to eliminate unwanted elements from the original data and extract relevant features. Supervised machine learning algorithms such as SVM, Random Forest and the Naive Bayes classifier were used to perform automatic sample identification. To evaluate the performance of the models, the 5-fold cross-validation technique was used. With the proposed methodology, it was possible to reach an accuracy of 93%, a sensitivity of 96% and a specificity of 89%, in addition to an area under the ROC curve of 0.95, in the identification of nasopharyngeal swab samples from previously diagnosed individuals. From these results, it can be concluded that Vis-NIR spectroscopy is a promising technique for the diagnosis of SARS-CoV-2.

Keywords: SARS-CoV-2, spectroscopy, machine learning, diagnosis

LISTA DE FIGURAS

Figura 1	Esquemático de um espectrômetro de dispersão. Em aplicações práticas, um anteparo reflexivo branco pode ser utilizado como amostra de referência.	25
Figura 2	Reflectância de amostras de <i>swab</i> testadas para SARS-CoV-2. Os sinais foram obtidos com o uso do espectrômetro <i>FieldSpec 3[®]</i> . Nela, é possível observar o comportamento da reflectância para cada comprimento de onda na faixa de 350 nm até 1400 nm.	26
Figura 3	Exemplo de sinal coletado com o espectrômetro <i>Tellspec Enterprise Sensor</i> . Na Figura, é o mostrado o sinal antes e depois do processo de remoção da linha de base pelo método dos mínimos quadrados modificado.	29
Figura 4	Entradas de classificadores que utilizam aprendizado supervisionado.	32
Figura 5	Esquemático da metodologia adotada para realizar a diferenciação entre amostras de <i>swab</i> positivas e negativas para SARS-CoV-2. . .	36
Figura 6	Configuração utilizada para aquisição dos dados com o <i>FieldSpec 3</i> . A lâmpada (1) e o sensor do espectrômetro (2) são posicionados manualmente a uma distância padronizada da amostra (3). Todo o sistema foi montado dentro de uma caixa preta, de forma que fontes luminosas externas não interferiram no experimento.	37
Figura 7	Configuração utilizada durante a aquisição dos dados com o <i>Tellspec Enterprise Sensor</i> . Como as lâmpadas já são integradas ao sistema, é necessário apenas aproximar o espectrômetro (1) da amostra. Um anteparo reflexivo de cor branca (2) é posicionado atrás da amostra com o objetivo de amenizar o efeito de transmitância através do líquido.	38

Figura 8	Processo de aquisição de dados para o Banco de Dados 3. a) O líquido foi manipulado com uma micropipeta de volume ajustável. 5 μL de líquido foram colocados nas janelas de cristal para formar um sanduíche. b) Foi utilizada uma impressora 3D para fabricar um <i>case</i> para o espectrômetro TES. Este <i>case</i> permitiu que as placas ficassem entre o sensor do espectrômetro e um anteparo branco reflexivo. c) Logo após a aquisição das medidas, as placas foram limpas com álcool etílico (99%). Todas as medidas de biossegurança foram cuidadosamente tomadas durante essa etapa.	39
Figura 9	Sinais coletados utilizando o FieldSpec 3. Na imagem, é possível observar uma região com pouca informação para comprimentos de onda acima de 1500 nm. Também é possível observar que o sinal apresenta uma descontinuidade em 1000 nm.	43
Figura 10	Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.	44
Figura 11	Curvas ROC obtidas para os 3 classificadores utilizados. É possível observar que todos os classificadores alcançaram um baixo desempenho, com o melhor resultado sendo obtido através da segunda derivada e do classificador <i>Naive Bayes</i> , com $AUC = 0,62$	45
Figura 12	Gráficos dos coeficientes da transformada <i>wavelet</i> para as duas metodologias de pré-processamento. Foram utilizados 9 níveis de decomposição e uma função <i>wavelet-mãe</i> do tipo Daubechies 1.	46
Figura 13	Curvas ROC obtidas para os 3 classificadores utilizados. O desempenho geral foi ruim em todas as situações, com nenhum modelo alcançando uma AUC superior a 0,5.	47
Figura 14	Curvas ROC obtidas pelos 3 classificadores em cada caso. É possível observar que todos os classificadores atribuíram previsões aleatórias.	48
Figura 15	Distribuição das amostras com base nos <i>scores</i> de suas componentes principais. No gráfico, são representadas as 3 componentes de maior variância explicada.	49

Figura 16	Curvas ROC obtidas pelos 3 classificadores em cada caso. Os classificadores obtiveram um ótimo desempenho, principalmente quando utilizada a segunda derivada dos sinais.	50
Figura 17	Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.	51
Figura 18	Curvas ROC obtidas pelos 3 classificadores em cada caso. Os classificadores obtiveram um ótimo desempenho, principalmente quando utilizada a segunda derivada dos sinais.	52
Figura 19	Projeção das 3 componentes principais com maior variância explicada. Nesse caso, os grupos formados pelo classificador <i>k-means</i> não representa nenhum conjunto de dados específico.	52
Figura 20	Sinais coletados utilizando o Tellspec Enterprise Sensor. Diferentemente do caso em que foi utilizado o FieldSpec 3, os sinais agora não apresentam uma descontinuidade em 1000 nm.	53
Figura 21	Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.	54
Figura 22	Grupos formados pelo classificador K-means. Uma primeira inspeção visual identificou a possível formação de dois grupos, por isso os parâmetros do classificador foram ajustados para a identificação de apenas dois. Apesar do agrupamento, os grupos formados não correspondem à amostras com metadados em comum.	55
Figura 23	Curvas ROC obtidas pelos 3 classificadores em cada caso. O classificador do tipo <i>Naive Bayes</i> alcançou os maiores valores de área sob a curva em ambos os cenários, com $AUC = 0,81$ e $AUC = 0,74$	56
Figura 24	Sinais coletados utilizando o Tellspec Enterprise Sensor em conjunto com as janelas de cristal. Mais uma vez os sinais apresentam um aspecto ruidoso devido à baixa resolução do espectrômetro.	57
Figura 25	Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.	58

Figura 26 Curva ROC para os 3 classificadores quando utilizados os comprimentos de onda selecionados com o teste F . É possível notar que as medidas extraídas da segunda derivada não apresentaram um bom desempenho. Já quando extraídas dos sinais cuja linha de base foi removida por interpolação, os classificadores tiveram uma boa performance, sendo o modelo do tipo *Naive Bayes* aquele com a maior área sobre a curva $AUC = 0,75$. Entretanto, a acurácia alcançada pelo SVM (72%) foi superior a dos algoritmos *Random Forest* (69%) e *Naive Bayes* (69%).

LISTA DE TABELAS

Tabela 1	Matriz de Confusão.	33
Tabela 2	Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Apesar do desempenho dos classificadores ter sido ruim, com acurácia em torno de 50%, destaca-se a especificidade média de 73% obtida com o classificador SVM para o caso em que os sinais tiveram sua linha de base removida por interpolação.	44
Tabela 3	Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Nesse caso foram utilizadas as energias médias para cada nível de decomposição da transformada <i>wavelet</i>	45
Tabela 4	Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Os <i>scores</i> das componentes principais foram usados como parâmetros de entrada.	46
Tabela 5	Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os 10 comprimentos de onda com maior valor de F . Houve uma melhora considerável nas acurácias obtidas, com destaque para o modelo em que a segunda derivada dos sinais foi usada em conjunto com os classificadores SVM e <i>Naive Bayes</i> , obtendo uma acurácia de 91% e 93% respectivamente. Também é possível notar que a segunda derivada obteve um melhor desempenho geral.	49
Tabela 6	Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os 10 comprimentos de onda com maior valor de F . Houve uma melhora considerável nas acurácias obtidas, com destaque para o modelo em que a segunda derivada dos sinais foi usada em conjunto com os classificadores SVM e <i>Naive Bayes</i> , obtendo uma acurácia de 83%. Também é possível notar que a segunda derivada obteve um melhor desempenho geral.	52

Tabela 7 Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os comprimentos de onda com maior valor de F . Ambas as metodologias de pré-processamento alcançaram acurácias semelhantes. Entretanto, os sinais que tiveram a linha de base removida por interpolação obtiveram uma sensibilidade maior. . 56

LISTA DE ABREVIATURAS E SIGLAS

ACC	Acurácia
ANOVA	Análise de variância
AUC	Área sob a curva ROC
CoV	Coronavirus
ESP	Especificidade
FN	Falso negativo
FP	Falso positivo
HCoV	Coronavirus que infectam humanos
IBMP	Instituto de Biologia Molecular do Paraná
KNN	<i>K nearest neighbors</i>
MTV Laborclin	Meio de transporte viral Laborclin
NaCl	Cloreto de sódio
PCA	Análise de componentes principais
PLS	<i>Partial least squares</i>
RNA	Rede neural artificial
ROC	<i>Receiver operating characteristic</i>
SARS-CoV	<i>Severe acute respiratory syndrome coronavirus</i>
SEN	Sensibilidade
SIMCA	<i>Soft modeling of class analogy</i>
SVM	<i>Support vector machine</i>
TES	Tellspec Enterprise Sensor

VN	Verdadeiro negativo
VP	Verdadeiro positivo
ATR-FTIR	<i>Attenuated total reflectance Fourier-transform infrared</i>
Covid-19	<i>Coronavirus disease 2019</i>
MERS-CoV	<i>Middle East respiratory syndrome coronavirus</i>
RT-qPCR	<i>Real-time reverse transcriptase polymerase reaction chain</i>
SARS-CoV-2	<i>Severe acute respiratory syndrome coronavirus 2</i>
Vis-NIR	<i>Visible near infrared</i>

SUMÁRIO

1	INTRODUÇÃO	20
2	OBJETIVOS	21
2.1	OBJETIVO GERAL	21
2.2	OBJETIVOS ESPECÍFICOS	21
3	REFERENCIAL TEÓRICO	21
3.1	SARS-CoV-2	21
3.2	ESPECTROSCOPIA Vis-NIR	23
3.3	TÉCNICAS DE ANÁLISE E PROCESSAMENTO DE SINAIS	26
3.3.1	Filtro Savitzky-Golay	26
3.3.2	Remoção da linha de base por interpolação	28
3.3.3	Análise de variância (ANOVA)	29
3.3.4	Transformada Wavelet	30
3.3.5	Análise de componentes principais	31
3.4	APRENDIZADO DE MÁQUINA	31
3.4.1	Avaliação dos modelos	33
3.5	TRABALHOS CORRELATOS	34
4	METODOLOGIA	35
4.1	OBTENÇÃO DAS AMOSTRAS DE SWAB NASOFARÍNGEO	36
4.2	AQUISIÇÃO DOS DADOS ESPECTRAIS	37
4.2.1	Banco de Dados 1	37
4.2.2	Banco de Dados 2	37
4.2.3	Banco de Dados 3	38
4.3	PRÉ-PROCESSAMENTO DOS SINAIS	40
4.4	EXTRAÇÃO DE CARACTERÍSTICAS	41
4.5	CLASSIFICAÇÃO E AVALIAÇÃO	41
5	RESULTADOS	42
5.1	Banco de Dados 1	42
5.1.1	Faixa de 350-1000 nm	42

5.1.2	Faixa de 1000-1500 nm	50
5.2	Banco de Dados 2	53
5.3	Banco de Dados 3	57
6	DISCUSSÃO	59
7	CONCLUSÃO E TRABALHOS FUTUROS	61
	REFERÊNCIAS	64

1 INTRODUÇÃO

O novo coronavírus, denominado mais precisamente vírus da síndrome respiratória aguda grave tipo 2 (SARS-CoV-2, do termo em inglês), causador da Covid-19 (do inglês, *coronavirus disease 2019*), está sendo responsável pela maior catástrofe sanitária do século XXI até o momento. Devido à sua alta transmissibilidade e alta taxa de mortalidade em indivíduos com comorbidades, a pandemia de SARS-CoV-2 já causou prejuízos bilionários, além de ter ceifado milhões de vidas (1).

Apesar dos avanços terapêuticos e do desenvolvimento da vacina, o diagnóstico rápido e eficiente ainda é essencial para otimizar o controle da transmissão e minimizar os efeitos mais graves da doença (2). O método considerado padrão ouro para o diagnóstico preciso da Covid-19 é o exame RT-qPCR (do inglês, *real-time reverse transcriptase polymerase reaction chain*). Apesar de ter uma ótima acurácia para o SARS-CoV-2, essa técnica possui a desvantagem de ter um custo elevado, além de necessitar do uso de equipamentos e material especializados (3). Como técnica alternativa ao RT-qPCR, o uso de métodos baseados em tomografia computadorizada do tórax também demonstram uma alta acurácia na identificação do SARS-CoV-2. Porém, assim como no caso anterior, é uma técnica que possui alto custo financeiro, além de não permitir um diagnóstico precoce (4). Uma técnica de baixo custo que é bastante empregada são os testes rápidos de imunocromatografia baseados em reações de imunoglobulinas IgM/IgG, cujo maior problema é a alta taxa de falsos-negativos e não terem finalidade diagnóstica (5).

Dessa forma, é possível observar que existe uma demanda por métodos de detecção do SARS-CoV-2 que sejam rápidos, eficientes e de baixo custo. Uma possível técnica é a espectroscopia óptica na faixa do visível e infravermelho próximo (Vis-NIR, do inglês *visibile near infrared*) que, em conjunto com algoritmos de inteligência artificial, já foi utilizada para detecção de alguns vírus (6–9). Assim, o uso da espectroscopia Vis-NIR com ferramentas de aprendizado de máquina se apresenta como uma técnica promissora para identificação do SARS-CoV-2.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver um método de detecção do SARS-CoV-2 em amostras de *swab* na-sofaríngeo utilizando técnicas de espectroscopia Vis-NIR e algoritmos de inteligência artificial.

2.2 OBJETIVOS ESPECÍFICOS

- Utilizar um espectrômetro comercial para obter os dados do espectro Vis-NIR de amostras de *swab* de casos suspeitos de Covid-19;
- Criar e organizar um banco de dados espectrais destinado ao estudo do SARS-CoV-2;
- Identificar o pré-processamento dos sinais espectrais que promova a eliminação de ruídos, artefatos e informações irrelevantes;
- Obter descritores dos sinais espectrais para detecção do SARS-Cov-2 em amostras de *swab*;
- Discriminar amostras de *swab* de casos suspeitos de Covid-19 com uso de algoritmos de aprendizado de máquina.

3 REFERENCIAL TEÓRICO

3.1 SARS-CoV-2

Os *Nidovirales* são uma ordem de vírus envelopados que possuem RNA não segmentado de sentido positivo (10). Essa ordem é dividida em quatro famílias: Coronaviridae, Arteriviridae, Mesoniviridae e Roniviridae. A família Coronaviridae possui ainda duas subfamílias, sendo uma delas a subfamília Coronavirinae, que compreende o grupo de vírus denominados de coronavírus (CoV) e é subdividida filogeneticamente em quatro gêneros: alfa, beta, teta e gama.

Dentre as espécies de coronavírus conhecidas, algumas são capazes de infectar humanos (HCoV). Duas dessas espécies pertencem ao gênero alfa, nomeadas: HCoV-229E e HCoV-NL63, e outras cinco ao gênero beta, nomeadas: HCoV-HKU1, HCoV-OC43, MERS-CoV (do inglês, *Middle East respiratory syndrome coronavirus*),

SARS-CoV (do inglês, *severe acute respiratory syndrome coronavirus*) e SARS-CoV-2 (11).

A maioria desses HCoV são considerados endêmicos, causando de 15% a 30% das infecções respiratórias registradas anualmente (10). Entretanto, três desses vírus já causaram surtos de infecção a nível mundial, sendo eles o SARS-CoV, o MERS-CoV e o SARS-CoV-2.

O surto de SARS-CoV teve início na China no fim de 2002 e chegou a contaminar 8.000 pessoas ao redor do mundo, provocando cerca de 800 mortes até meados de 2003 (12). Os sintomas apresentados nos indivíduos infectados se assemelhavam aos de uma gripe comum (febre, dor de cabeça, calafrios), porém mais graves. Especula-se que o vírus tenha sido transmitido aos humanos através de civetas que eram comercializadas em mercados de animais vivos como alimentos exóticos. Após esse surto, nenhum novo caso de infecção por SARS-CoV foi reportado em humanos desde 2004, sendo então considerada uma doença erradicada.

Identificado pela primeira vez na Arábia Saudita em abril de 2012, o MERS-CoV causou uma epidemia que se espalhou por 27 países, infectando 2547 pessoas e provocando 886 mortes até março de 2021. A transmissão do vírus acontece principalmente através do contato direto ou indireto com dromedários contaminados (13). Os sintomas iniciais são similares ao da infecção causada pelo SARS-CoV, porém alguns pacientes são assintomáticos. Nos casos mais graves, o quadro clínico evolui rapidamente, apresentando sintomas como síndrome do desconforto respiratório agudo, choque séptico, falência múltipla dos órgãos e morte (14).

Já o SARS-CoV-2 é o vírus causador da Covid-19, doença responsável pela maior pandemia do século XXI, sendo diagnosticada em 651.198.402 indivíduos e provocando 6.656.601 mortes até 26 de dezembro de 2022 (15). Esse novo coronavírus foi identificado pela primeira vez em dezembro de 2019, na China, estando associado ao aparecimento de uma pneumonia misteriosa em pessoas que frequentavam um mercado de animais silvestres vivos na cidade de Wuhan (16). A análise do genoma do SARS-CoV-2 mostrou uma semelhança de 96% com o coronavírus encontrado em morcegos (CoV RaTG13) na província de Yunna, também na China, o que leva a crer que os humanos adquiriram o vírus através desses animais (17).

A transmissão do SARS-CoV-2 se dá principalmente pelo contato direto com in-

divíduos contaminados e é considerado um vírus com alta transmissibilidade, tendo um número básico de reprodução entre 2 e 2,5 (18). Os sintomas provocados pela Covid-19 não são específicos, podendo haver desde infecções assintomáticas ou apenas sintomas leves, como febre, tosse, mialgia e fadiga, até infecções em que os pacientes apresentam uma pneumonia severa que pode levar à morte (19).

A principal ferramenta para auxiliar no diagnóstico do SARS-CoV-2 é a análise de amostras de *swab* de nasofaringe por meio do método RT-qPCR. Esse método é considerado o padrão ouro para o diagnóstico da Covid-19, uma vez que possui alta sensibilidade e não tem reação cruzada com outros coronavírus (3). Por necessitar de mão de obra e equipamentos especializados, o RT-qPCR é considerado uma técnica de alto valor financeiro. Outra técnica que se destaca por seu baixo custo é a análise sorológica por meio de testes rápidos baseados em reações de imunoglobulinas IgM/IgG, cujo principal problema é a alta taxa de falsos-negativos, em especial no início da infecção (5). Atualmente, esses testes rápidos para pesquisa de antígeno são os mais utilizados para detectar a infecção.

O tratamento da Covid-19 é feito primeiramente através da administração de antivirais e tratamentos baseados no uso de anticorpos. Quando a infecção já se encontra em um estágio avançado, são usados fármacos anti-inflamatórios como corticosteroides e imunomoduladores (20). Apesar de haver um protocolo de tratamento e fármacos comerciais específicos para o tratamento da doença, as medidas profiláticas de isolamento e a imunização da população com vacinas ainda são as melhores estratégias para o enfrentamento da pandemia.

3.2 ESPECTROSCOPIA Vis-NIR

A espectroscopia é o estudo da absorção e emissão da luz, ou qualquer outra forma de radiação, pela matéria (21). Ao longo da história, a espectroscopia tem sido crucial no estudo da física, em especial nos campos da mecânica quântica, relatividade e eletrodinâmica quântica. Dentre as contribuições mais importantes, destacam-se o uso da espectroscopia na descoberta do efeito fotoelétrico (22) e no estudo da radiação emitida por corpos negros (23).

Uma das formas de realizar esse estudo consiste em analisar a interação entre um objeto e ondas eletromagnéticas com comprimentos de onda entre 350 nm e

2500 nm, que compreendem a faixa do espectro visível e próximo ao infravermelho. A análise dessa interação se dá por meio dos parâmetros de reflectância, absorvância e transmitância. A reflectância é uma grandeza adimensional e indica a razão entre o fluxo de radiação refletido pelo corpo e o fluxo incidente. A absorvância e a transmitância também são grandezas adimensionais e indicam, respectivamente, a fração de radiação incidente que foi absorvida pelo corpo e a fração de radiação incidente que atravessou o corpo.

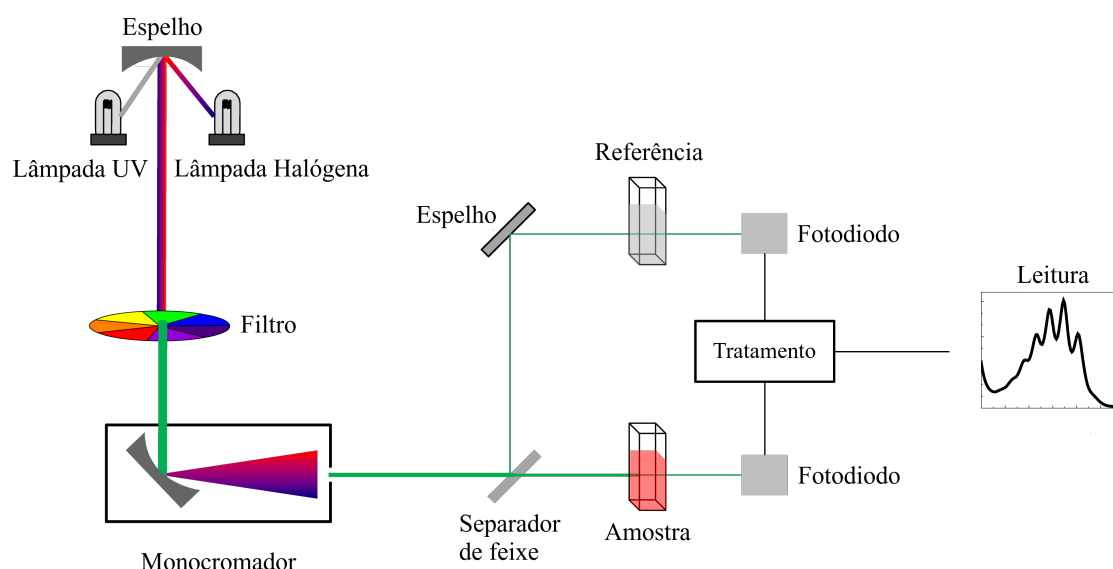
A medição desses parâmetros é feita através de equipamentos conhecidos como espectrômetros de infravermelho ou espectrofotômetros. Os dois tipos de espectrômetros de infravermelho mais utilizados atualmente são os espectrofotômetros dispersivos e os espectrofotômetros de transformada de Fourier. Ambos os aparelhos produzem resultados semelhantes, porém este último é capaz de realizar a medição mais rapidamente (24). Os equipamentos utilizados nesse trabalho são do tipo de dispersão, e seu princípio de funcionamento é mostrado em mais detalhes a seguir.

A Figura 1 ilustra de maneira simplificada o funcionamento de um espectrofotômetro de dispersão. As lâmpadas em conjunto com um espelho são responsáveis por gerar um feixe de radiação eletromagnética com os comprimentos de onda de interesse. Esse feixe é então direcionado para um primeiro filtro que funciona como uma rede de difração, separando as bandas frequenciais do feixe original. Esse filtro gira lentamente, de forma que todas as bandas são analisadas em algum momento.

Em seguida, o monocromador transforma o feixe nele incidente em uma luz monocromática, ou seja, uma faixa de luz visível composta por apenas um único comprimento de onda. Vale ressaltar que o monocromador não é um instrumento ideal. Dessa forma, a luz monocromática gerada por ele também não é ideal. Sendo assim, o monocromador deve ser considerado como um filtro de banda estreita. A luz monocromática é então dividida em dois feixes de igual intensidade, sendo que um deles é direcionado para a amostra de teste e o outro para uma amostra de referência. Os fotodiodos são então excitados pelos feixes luz que atravessaram os corpos de prova gerando um sinal elétrico. As diferenças entre os sinais elétricos gerados pela amostra de teste e pela amostra de referência são então computados para gerar o espectro (24).

Para se obter um espectro de qualidade, as amostras devem ser preparadas e

Figura 1 – Esquemático de um espectrômetro de dispersão. Em aplicações práticas, um anteparo reflexivo branco pode ser utilizado como amostra de referência.



Fonte: Wikipedia (2023).

colocadas em um recipiente ou cela ideais para esse propósito. Vidros e plásticos absorvem muito em quase todas as regiões do espectro Vis-NIR, e por isso devem ser evitados como recipientes (24).

Para compostos sólidos, um método de preparação bem consolidado consiste em moer a amostra sólida e a misturar com brometo de potássio (KBr). Em seguida, a mistura é comprimida sob alta pressão, formando uma pastilha de KBr que pode ser levada ao espectrômetro. Esse método é eficiente devido ao fato do KBr ser transparente até um comprimento de onda de 25.000 nm, ou seja, não absorve nenhum comprimento de onda na faixa do espectro Vis-NIR (24).

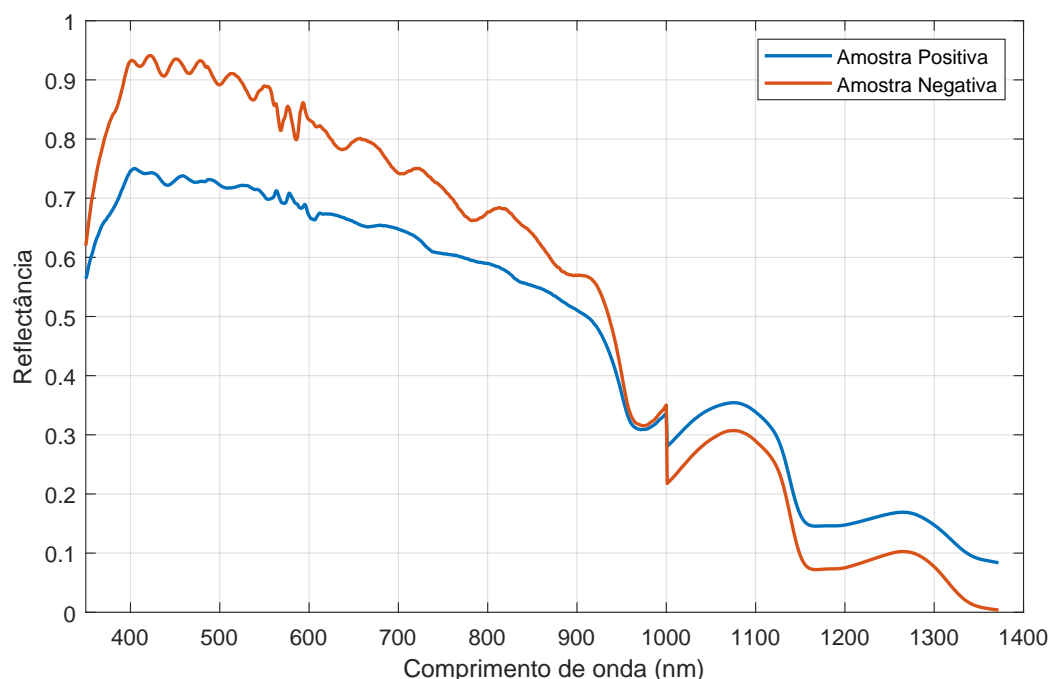
Já para líquidos, uma gota do composto orgânico deve ser colocada entre duas placas polidas de cloreto de sódio (NaCl) ou brometo de potássio. Essas placas são chamadas de placas de sal ou janelas de cristal. Quando apertadas uma contra a outra, as janelas de cristal formam um “sanduíche” com o líquido a ser analisado. Vale ressaltar que essas placas são frágeis e solúveis em água (24).

Por fim, alguns instrumentos, em especial os espectrofotômetros de transformada de Fourier, possuem um módulo denominado acessório de reflectância total atenuada. Com esse módulo, a preparação das amostras é dispensada tanto para

sólidos quanto para líquidos.

A Figura 2 mostra um exemplo de gráfico de reflectância de duas amostras de *swab* de nasofaringe que testaram positivo e negativo para SARS-CoV-2.

Figura 2 – Reflectância de amostras de *swab* testadas para SARS-CoV-2. Os sinais foram obtidos com o uso do espectrômetro *FieldSpec 3®*. Nela, é possível observar o comportamento da reflectância para cada comprimento de onda na faixa de 350 nm até 1400 nm.



Fonte: O autor.

Atualmente, a espectroscopia Vis-NIR tem sido muito utilizada na análise de alimentos e na agricultura (25, 26). Além disso, devido à capacidade de determinar alguns parâmetros de ligações químicas, essa técnica também já foi utilizada de forma eficiente na detecção automática dos vírus da dengue, zika e *influenza* (6–8).

3.3 TÉCNICAS DE ANÁLISE E PROCESSAMENTO DE SINAIS

A seguir, serão apresentadas técnicas de análise e processamento de sinais nos domínios temporal e frequencial. Essas técnicas serão utilizadas para caracterizar os sinais de espectroscopia obtidos. Vale ressaltar que, embora os sinais de espectroscopia não sejam sinais temporais, as técnicas de análise nos domínios temporal e frequencial podem ser empregadas de forma análoga.

3.3.1 Filtro Savitzky-Golay

Os filtros do tipo Savitzky-Golay foram propostos com o objetivo de resolver a interpolação de um polinômio de grau n através do cálculo de uma convolução (27). Além da eficiência no que diz respeito ao custo computacional, esses filtros possuem a capacidade de eliminar ruídos sem deformar os picos presentes no sinal original, tornando-os particularmente úteis para estudos de espectroscopia.

Dada uma janela de tamanho $M = 2m + 1$, em que m é um número inteiro que define a largura da janela de interpolação, tem-se um polinômio interpolador de grau n dado pela equação:

$$f(i) = \sum_{k=0}^n b_{n,k} i^k, \quad (3.1)$$

em que i são os inteiros no intervalo $-m < i < m$. Percebe-se que a janela escolhida deve ser simétrica, de tamanho ímpar e de tal forma que $i = 0$ represente seu ponto central. Dessa forma, uma vez encontrado o polinômio $f(i)$, o ponto central é definido por $f(i = 0)$. Na sequência, a janela é deslocada em uma posição, um novo polinômio é encontrado, e o novo ponto central é definido.

O problema dos mínimos quadrados consiste em encontrar os valores de $b_{n,k}$ que minimizam as diferenças quadradas entre o polinômio interpolador, $f(i)$, e as observações feitas na variável, y_i , dentro da janela especificada. Assim, deve-se resolver a equação:

$$\frac{\partial}{\partial b_{n,k}} \left[\sum_{i=-m}^m (f(i) - y_i)^2 \right] = 0. \quad (3.2)$$

Para determinar um coeficiente $b_{n,r}$ qualquer, utiliza-se a equação:

$$2 \sum_{i=-m}^m \left[\left(\sum_{k=0}^n b_{n,k} i^k \right) - y_i \right] i^r = 0, \quad (3.3)$$

que pode ser reescrita na forma:

$$\sum_{i=-m}^m \left[\left(\sum_{k=0}^n b_{n,k} i^k \right) - y_i \right] i^r = 0.$$

Invertendo a ordem do somatório e reorganizando os termos:

$$\sum_{k=0}^n b_{n,k} \sum_{i=-m}^m i^{k+r} = \sum_{i=-m}^m y_i i^r.$$

Por fim, fazendo $S_{k+r} = \sum_{i=-m}^m i^{k+r}$ e $F_r = \sum_{i=-m}^m y_i i^r$, tem-se a equação:

$$\sum_{k=0}^n b_{n,k} S_{k+r} = F_r. \quad (3.4)$$

Percebe-se que o membro esquerdo da Eq. (3.4) tem a forma de uma soma de convolução. Além disso, S_{r+k} depende apenas do tamanho da janela, da ordem do polinômio e do grau do termo que se deseja determinar o valor. O valor de F_r também é determinado apenas por variáveis já definidas pelo usuário e pelos valores que se deseja interpolar. Assim, é possível obter, da Eq. (3.4), $n + 1$ expressões que são suficientes para montar um sistema linear cujas variáveis sejam $b_{n,k}$, solucionando então o problema de minimização.

Filtros do tipo Savitzky-Golay também podem ser facilmente implementados para o cálculo de derivadas, uma vez que as derivadas de um polinômio dependem apenas dos seus coeficientes, que nesse caso são $b_{n,k}$. Ademais, derivadas são úteis na análise espectroscópica, já que se trata de uma forma de eliminar o problema da linha de base comum a sinais de espectroscopia.

3.3.2 Remoção da linha de base por interpolação

Como mostrado anteriormente, o uso de derivadas pode ser eficiente na remoção da linha de base. Isso se dá devido ao fato da linha de base corresponder a constantes ou a funções com pouca variação que são somadas ao sinal de interesse na forma de ruído de baixa frequência. Portanto, a derivada da linha de base em geral tende para valores próximos a zero. Em contrapartida, a derivada distorce o sinal de interesse original, o que pode ser, ou não, uma desvantagem na análise.

Uma forma eficiente de remoção da linha de base sem causar distorção é através da interpolação do espectro por polinômios. O método dos mínimos quadrados é o mais utilizado para realizar essa tarefa. Entretanto, diferentemente do filtro Savitzky-Golay, a interpolação não é feita de forma janelada, mas sim utilizando todo o espectro.

Em (28), os autores propuseram uma modificação no método dos mínimos quadrados para tornar a identificação e remoção da linha de base mais eficiente na espectroscopia de Raman. Apesar disso, a metodologia pode ser aplicada para outros

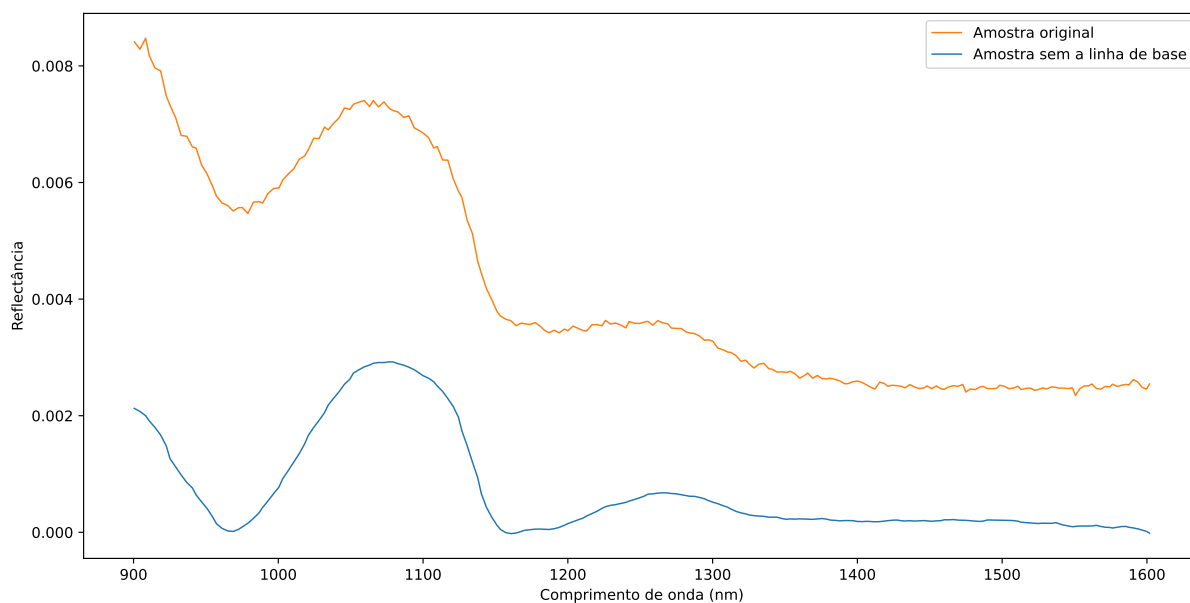


Figura 3 – Exemplo de sinal coletado com o espectrômetro TellSpec Enterprise Sensor. Na Figura, é o mostrado o sinal antes e depois do processo de remoção da linha de base pelo método dos mínimos quadrados modificado.

tipos de estudos espectroscópicos, como na espectroscopia Vis-NIR. O método consiste em interpolar um polinômio pelo método dos mínimos quadrados mas, caso o polinômio gerado em algum ponto alcance um valor superior ao do sinal original, o valor desse ponto automaticamente assume o valor do sinal original. O processo é repetido iterativamente até que o polinômio interpolado seja inferior ao sinal original em todos os pontos. A Figura 3 apresenta um exemplo cujo sinal teve a linha de base removida pelo método descrito por (28).

3.3.3 Análise de variância (ANOVA)

A ANOVA faz a comparação de duas ou mais médias populacionais ou de tratamento e determina, a partir de um valor F , o quão próximas essas médias estão umas das outras (29). Seja $X_{i,j}$ a variável aleatória que representa a j -ésima medida da i -ésima população e $x_{i,j}$ o valor observado de $X_{i,j}$ durante o experimento. É possível definir a média populacional \bar{X}_i por:

$$\bar{X}_i = \frac{\sum_{j=1}^J X_{i,j}}{J} \quad i = 1, 2, 3, \dots, I$$

sendo J o número de observações e I o número de populações. Define-se também a média global \bar{X} e as variâncias amostrais S_i^2 , como mostrado nas Equações (3.5) e

(3.6).

$$\bar{X} = \frac{\sum_{i=1}^I \sum_{j=1}^J X_{i,j}}{IJ} \quad (3.5)$$

$$S_i^2 = \frac{\sum_{j=1}^J (X_{i,j} - \bar{X}_i)^2}{J - 1} \quad i = 1, 2, 3, \dots, I \quad (3.6)$$

A estatística do teste ANOVA busca relacionar a medida das diferenças entre as médias populacionais \bar{X}_i e a média global \bar{X} com a medida de variação S_i^2 calculada dentro de cada população (29). Para isso, é necessário definir os chamados quadrado médio dos tratamentos (QMTr) e quadrado médio do erro (QME), que são apresentados nas equações (3.7) e (3.8).

$$\text{QMTr} = \frac{J}{I - 1} \sum_i (\bar{X}_i - \bar{X})^2 \quad (3.7)$$

$$\text{QME} = \frac{S_1^2 + S_2^2 + \dots + S_I^2}{I} \quad (3.8)$$

O valor F é então definido como $F = \text{QMTr}/\text{QME}$. Para o caso em que as médias calculadas estão próximas uma da outra, o valor de QMTr tende a ser pequeno, implicando também um menor valor de F . Dessa forma, quanto maior o valor de F , maior é a distância entre as médias populacionais. Essa estatística pode ser útil na seleção de variáveis que melhor discriminam dois grupos. No caso dos estudos de espectroscopia, é possível calcular o valor de F para cada comprimento de onda, relacionando a média populacional do grupo formado pelas amostras negativas com a média populacional do grupo formado pelas amostras positivas, e selecionando apenas as variáveis que alcançaram os maiores valores de F .

3.3.4 Transformada Wavelet

Wavelets são uma família de funções construídas a partir de translações e dilatações de uma única função $\psi(t)$ chamada de *wavelet-mãe* (30). Esta função é definida pela equação:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \quad (3.9)$$

Na Equação (3.9), os parâmetros a e b se referem, respectivamente, à dilatação e à translação da *wavelet* mãe. De forma análoga à transformada de Fourier, a transformada *wavelet* fornece os coeficientes necessários para descrever um sinal contínuo no tempo, $x(t)$, como uma combinação linear de funções $\psi_{a,b}(t)$. A transformada *wavelet* de tempo contínuo $W(a, b)$ é definida pela equação:

$$W(a, b) = \int_{-\infty}^{\infty} x(t)\psi_{a,b}(t)dt \quad (3.10)$$

3.3.5 Análise de componentes principais

Na análise de componentes principais, busca-se uma transformação linear para um conjunto de dados de modo que o novo espaço vetorial gerado, de menor dimensão e linearmente independente, contenha as informações necessárias para descrever a variação nos dados originais com boa precisão.

Para atingir esse objetivo, a transformação deve ser tal que o ruído e a colinearidade presente no conjunto de dados a ser transformado sejam eliminados. Dessa forma, a PCA determina uma base ortonormal formada pelas chamadas componentes principais, que são obtidas através de uma combinação linear das variáveis originais (31). Os valores das variáveis nesse novo espaço vetorial são chamamos de *scores*, que podem ser interpretados como projeções dos dados originais nas componentes principais.

3.4 APRENDIZADO DE MÁQUINA

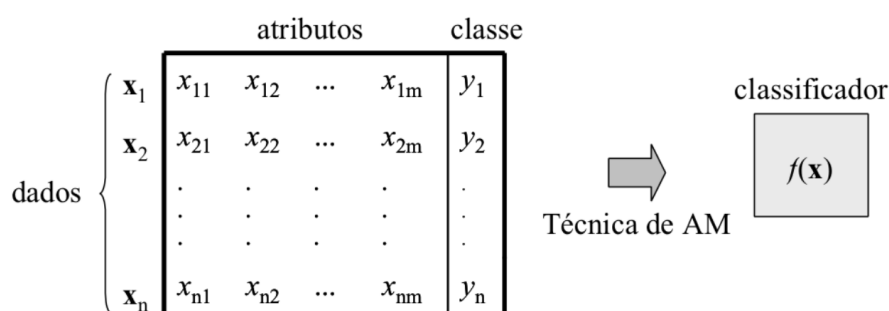
As técnicas de aprendizado de máquina empregam um princípio de inferência denominado indução, no qual são obtidas conclusões genéricas a partir de um conjunto particular de exemplos (32).

Em algoritmos de aprendizado de máquina que se valem do aprendizado dito supervisionado, os dados de treinamento fornecidos ao classificador são formados por entradas rotuladas na forma (x_i, y_i) , em que x_i representa um exemplo e y_i seu respectivo rótulo (classe à qual o exemplo pertence). Com base nesses dados de treinamento, os algoritmos identificam padrões que são utilizados posteriormente para

predizer corretamente os rótulos de novos dados não fornecidos anteriormente.

Os exemplos x_i , em geral, são representados por um vetor de características ou atributos. Os atributos podem ser dados qualitativos, como cor e forma, ou quantitativos, como velocidade e comprimento. Os rótulos y_i devem ser formados por valores inteiros que indiquem a que classe os exemplos pertencem. A Figura 4 mostra de forma visual como são montadas as entradas de um classificador que utiliza técnicas de aprendizado supervisionado.

Figura 4 – Entradas de classificadores que utilizam aprendizado supervisionado.



Fonte: (LORENA; CARVALHO, 2007).

A Figura 4 demonstra uma das maneiras de formar matrizes de entrada para treinamento desses algoritmos. Cada linha representa um exemplo distinto, enquanto as colunas representam os atributos desses exemplos. Na coluna final, são apresentados os rótulos indicando a qual classe os exemplos pertencem.

Dentro do aprendizado de máquina, diversos algoritmos são amplamente utilizados, como máquinas de vetor de suporte (SVM, do inglês, *support vector machines*), *Naive Bayes* e *Random Forest*, que são classificadores não probabilísticos e supervisionados.

Além do aprendizado supervisionado, tem-se também os modelos de aprendizado ditos não supervisionado, sendo esses conhecidos como *clusterizadores*. Nesse caso, os rótulos das amostras não são fornecidos ao algoritmo, cabendo ao modelo a determinação dos rótulos com base em algum padrão identificado nas amostras de treinamento. Um exemplo de classificador desse tipo é o chamado *k-means*, que agrupa um conjunto de dados em k *clusters* com base na distância euclidiana entre as amostras num dado espaço vetorial (33).

3.4.1 Avaliação dos modelos

A principal ferramenta de análise de desempenho dos classificadores são medidas extraídas da chamada matriz de confusão (34). A Tabela 1 apresenta uma matriz de confusão de dimensão 2×2 .

Tabela 1 – Matriz de Confusão.

		Verdade	
		Positivo	Negativo
Predição	Positivo	VP	FP
	Negativo	FN	VN

Fonte: O Autor.

A diagonal principal da matriz apresenta as predições corretas do classificador, em que VP indica os verdadeiros positivos e VN os verdadeiros negativos. A diagonal secundária mostra as predições incorretas do classificador, sendo FP os falsos positivos e FN os falsos negativos.

Da matriz de confusão, podem ser extraídas outras métricas importantes, como a sensibilidade (SEN), a especificidade (ESP) e a acurácia (ACU). A obtenção dessas medidas pode ser feita, respectivamente, com o uso das equações:

$$SEN = \frac{VP}{VP + FN} \quad (3.11)$$

$$ESP = \frac{VN}{FP + VN} \quad (3.12)$$

$$ACU = \frac{VP + VN}{VP + FP + FN + VN} \quad (3.13)$$

A sensibilidade indica a capacidade do classificador predizer a classe positiva, enquanto a especificidade a classe negativa. A acurácia indica a razão da quantidade de total de predições corretas pelo total de predições. Por levar em consideração tanto a classe positiva quanto a negativa, a acurácia é a principal medida de avaliação de classificadores.

Outra forma de avaliar o desempenho de classificadores é através da chamada

curva característica de operação do receptor (ROC, do inglês *Receiver Operating Characteristic*). Nessa curva, o eixo das ordenadas representa a sensibilidade e o eixo das abcissas o valor de 1-ESP. Durante a etapa de predição, os modelos atribuem às amostras um valor correspondente à probabilidade dessa pertencer à classe positiva. Se essa probabilidade for superior a um certo limiar, a amostra é considerada positiva, caso contrário a amostra é considerada negativa. Assim, para levantar a curva, o limiar de decisão dos modelos é gradualmente elevado, e a cada novo ponto os valores de sensibilidade e especificidade são determinados e inseridos no gráfico (35).

Dessa forma, através de uma inspeção visual, é possível analisar o 'custo' para o classificador obter um verdadeiro positivo em um dado limiar, uma vez que a quantidade de verdadeiros positivos em função da quantidade de falsos positivos é relacionada diretamente pela curva. Além disso, a área sob a curva ROC (AUC, do inglês *area under a ROC curve*) é considerada uma medida de desempenho geral do classificador, representando a capacidade do modelo atribuir para uma amostra positiva uma probabilidade superior à de uma amostra negativa (35).

3.5 TRABALHOS CORRELATOS

O uso da espectroscopia na identificação de diferentes tipos de vírus é algo comum na literatura científica. Em (7), os autores utilizaram a espectroscopia do infravermelho próximo para identificar mosquitos infectados pelo vírus da zika. Essa identificação é importante para controlar eventuais surtos, uma vez que permite combater a proliferação do vetor da doença. Nesse estudo, um espectrofotômetro foi utilizado para coletar os sinais diretamente do tórax e da cabeça dos mosquitos diagnosticados pelo método RT-qPCR. Na etapa de classificação, foi utilizado um modelo de regressão por mínimos quadrados parciais. Seguindo esse método, os autores conseguiram uma acurácia acima de 99% na identificação dos mosquitos infectados.

O trabalho desenvolvido pelos pesquisadores em (9) buscou identificar o vírus da hepatite B utilizando espectroscopia de Raman. Um banco de dados de 1.000 amostras de soro sanguíneo coletadas de pacientes testados para hepatite B foi utilizado. A análise de componentes principais (PCA) foi aplicada para extrair dos sinais características que serviram de entrada para um classificador SVM. Os hiperparâmetros do classificador foram ajustados através de um método de otimização por enxame de

partículas. O modelo proposto chegou a uma acurácia de 93%.

No projeto desenvolvido em (8), os pesquisadores utilizaram espectroscopia Vis-NIR na identificação de pacientes infectados pelo vírus *influenza*. Para isso, foram coletados os sinais espectrais de amostras de aspiração nasal de indivíduos diagnosticados por testes de imunocromatografia. Para etapa de extração de características, foi aplicado um modelo PCA. Já na etapa de classificação, os autores utilizaram um classificador do tipo SIMCA (do inglês, *soft modeling of class analogy*). O método proposto demonstrou um bom desempenho na identificação do vírus *influenza*, com uma acurácia acima de 96%. Também foi avaliada a capacidade do modelo em diferenciar pacientes infectados pelo vírus *influenza* de pacientes infectados pelo vírus sincicial respiratório. Nessa tarefa, o classificador não obteve uma boa performance, com acurácias em torno de 50%.

Por fim, ainda que raros, também é possível encontrar trabalhos que tratam do uso da espectroscopia para o diagnóstico do SARS-CoV-2. Em (36), os autores utilizaram um espectrofotômetro de transformada de Fourier para coletar sinais espectrais de amostras de RNA com e sem a presença do vírus SARS-CoV-2. Foram aplicados métodos de PCA e quadrados mínimos parciais para extrair características que serviram de entrada para um classificador SVM. Os pesquisadores conseguiram uma acurácia de 98% com esse método. Já no trabalho desenvolvido em (37), foi utilizada uma metodologia semelhante, mas nesse caso os dados espectrais eram obtidos diretamente das amostras de *swab*. Com esse modelo, se alcançou uma acurácia acima de 78%.

4 METODOLOGIA

O trabalho teve como objetivo desenvolver uma metodologia de aquisição de sinais espectrais Vis-NIR para detectar amostras de *swab* nasofaríngeo de pacientes infectados pelo SARS-CoV-2. A detecção se dá por meio da utilização de algoritmos computacionais de aprendizado de máquina, em que as amostras são apresentadas para classificação.

Esta seção descreve o método adotado na aquisição dos dados e na análise e classificação dos sinais espectrais. O fluxograma da Figura 5 resume as etapas relacionadas à execução do projeto que serão detalhadas nas subseções seguintes.

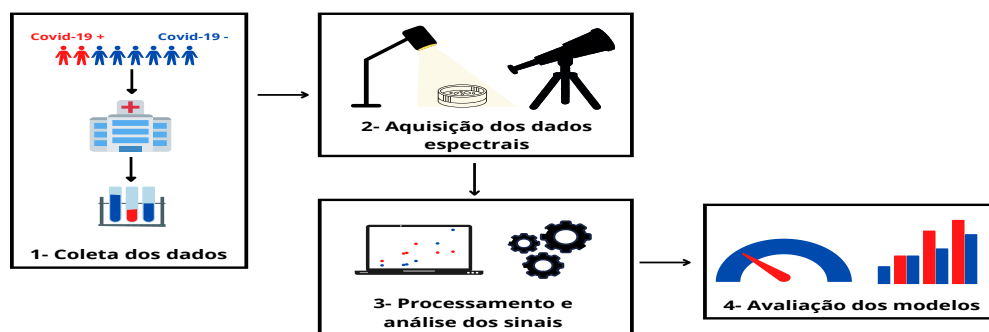


Figura 5 – Esquemático da metodologia adotada para realizar a diferenciação entre amostras de *swab* positivas e negativas para SARS-CoV-2.

4.1 OBTENÇÃO DAS AMOSTRAS DE SWAB NASOFARÍNGEO

As amostras de *swab* nasofaríngeo usadas nesse estudo foram fornecidas pelo Laboratório de Diagnóstico Molecular de Covid-19 da Univasf, provenientes de pacientes com suspeita de COVID-19 residentes em municípios que compõe a VIII Gerência Regional de Saúde de Pernambuco: Petrolina, Cabrobó, Orocó, Lagoa Grande, Santa Maria da Boa Vista, Dormentes e Afrânio. Foram utilizadas amostras de pacientes encaminhadas ao laboratório da Univasf no período de Agosto de 2021 até Março de 2022. Essas amostras foram mergulhadas em tubos de ensaio contendo uma solução de transporte e identificadas por um código numérico sem nenhuma informação pessoal sobre os pacientes. O laboratório também forneceu o diagnóstico de cada amostra com base nos resultados do teste RT-qPCR obtidos utilizando o BIOMOL OneStep/Covid-19 kit (IBMP).

Os critérios de inclusão dos espécimens para esse estudo foram: amostras coletadas de pacientes entre o terceiro e o décimo dia de sintomas; amostras armazenadas utilizando como solução de transporte um meio de transporte viral fabricado pela empresa Laborclin (MTV Laborclin) ou uma solução salina de NaCl (0,1%); amostras com tempo de armazenamento inferior a 48h; amostras com informação referente ao ciclo de quantificação (Cq).

Vale ressaltar que uma identificação adicional foi feita para indicar em qual solução de transporte cada amostra foi acondicionada. Além disso, também foram identificadas as amostras cujo volume de solução se mostrou excessivo.

4.2 AQUISIÇÃO DOS DADOS ESPECTRAIS

Nesse estudo, três diferentes bancos de dados espectrais foram montados, cada um deles sob diferentes condições de aquisição, que permitiram avaliar o desempenho do sistema quando analisadas diferentes faixas espectrais, equipamentos de coleta e soluções de transporte.

4.2.1 Banco de Dados 1

O espectrômetro FieldSpec 3, fabricado pela empresa Analytical Spectral Devices, foi utilizado para realizar a aquisição dos dados espectrais. Esse modelo possui uma faixa de operação entre 350 nm e 2500 nm com uma resolução de 3-10 nm. Uma vez que esse equipamento não possui uma fonte de iluminação embutida, uma lâmpada externa de halogênio-tungstênio de 50 W foi necessária para iluminar as amostras.

Os espectros foram obtidos diretamente da mistura entre o *swab* nasofaríngeo e a solução de transporte (MTV Laborclin ou solução salina de NaCl) presentes nos tubos de ensaio, dispensando o uso das janelas de cristal. Um total de 523 amostras espectrais foram coletadas utilizando o FieldSpec 3 entre os dias 30 de agosto de 2021 e 10 de dezembro de 2021, sendo que 60 dessas testaram positivo para Covid-19 e 463 testaram negativo de acordo com o método RT-qPCR. A Figura 6 apresenta a configuração utilizada para aquisição dos dados com o FieldSpec 3.

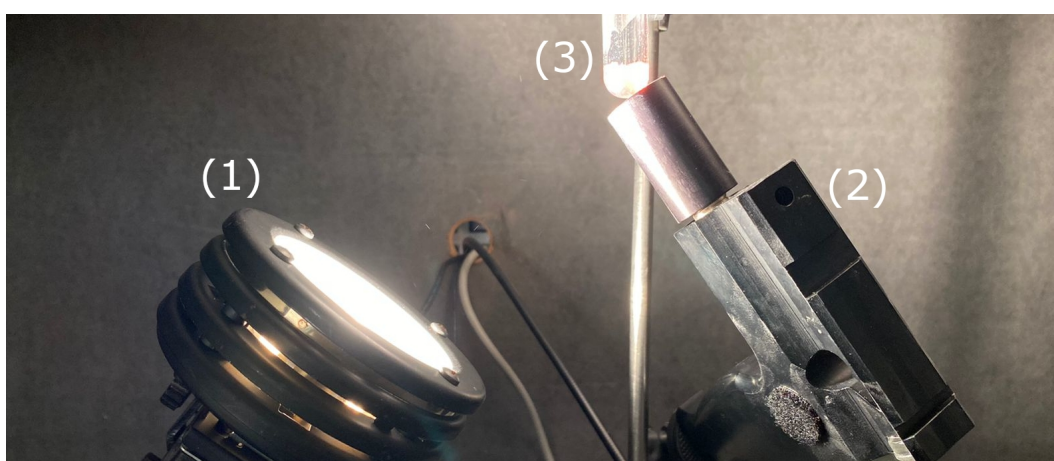


Figura 6 – Configuração utilizada para aquisição dos dados com o FieldSpec 3. A lâmpada (1) e o sensor do espectrômetro (2) são posicionados manualmente a uma distância padronizada da amostra (3). Todo o sistema foi montado dentro de uma caixa preta, de forma que fontes luminosas externas não interferiram no experimento.

4.2.2 Banco de Dados 2

O espectrômetro Telspec Enterprise Sensor (TES), fabricado pela empresa Telspec, foi utilizado para realizar a aquisição dos espectros. Este espectrômetro é construído com base no módulo DLP NIRScan da Texas Instruments. Vale notar que sua faixa de operação se limita ao infravermelho próximo, entre os comprimentos de onda 900-1700 nm, com uma resolução de 10 nm. Esse aparelho possui duas lâmpadas embutidas de halogênio-tungstênio de 0,7 W responsáveis por iluminar as amostras, dessa forma, não é necessária uma lâmpada externa para realizar a aquisição dos dados.

As medidas mais uma vez foram tomadas diretamente do líquido presente nos tubos de ensaio. Além disso, 152 amostras de diferentes indivíduos foram usadas, sendo que 23 deles testaram positivo e 129 testaram negativo para Covid-19 considerando a técnica RT-qPCR. O número de amostras coletadas para esse banco de dados foi inferior ao caso anterior. Isso se deu devido a menor disponibilidade de amostras que atendessem os critérios de inclusão durante o período de coleta, que aconteceu entre os dias 22 de fevereiro de 2022 e 16 de março de 2022. A Figura 7 apresenta a configuração usada para adquirir os dados espectrais para este caso.



Figura 7 – Configuração utilizada durante a aquisição dos dados com o Telspec Enterprise Sensor. Como as lâmpadas já são integradas ao sistema, é necessário apenas aproximar o espectrômetro (1) da amostra. Um anteparo reflexivo de cor branca (2) é posicionado atrás da amostra com o objetivo de amenizar o efeito de transmitância através do líquido.

4.2.3 Banco de Dados 3

Mais uma vez, o espectrômetro TES foi usado para coletar os dados espectrais. Diferentemente do Banco de Dados 2, em que as medidas eram obtidas diretamente dos tubos de ensaio, nesse caso foram utilizadas janelas de cristal de cloreto de sódio

ou brometo de potássio. Essa configuração permitiu avaliar a influência do vidro dos tubos de ensaio no desempenho do sistema, bem como no comportamento dos sinais espectrais.

Dois espécimes, que utilizaram o MTV Laborclin como solução de transporte, foram utilizados para criar este conjunto de dados. Um deles testou positivo para Covid-19, enquanto o outro testou negativo. Essas amostras foram então re-amostradas 50 vezes cada, utilizando $5 \mu\text{L}$ de fluido (*swab* + MTV Laborclin) que foi colocado nas janelas de cristal para formar uma espécie de sanduíche. Essas placas foram então fixadas a um anteparo reflexivo branco, e os espectros foram coletados usando o TES. A limitação no número de amostras desse banco de dados se justifica devido ao período de coleta (setembro de 2022), em que a quantidade de espécimens que atendessem aos critérios de inclusão era escassa. Além disso, para esse caso, a aquisição dos dados espectrais exigiu um rigoroso protocolo de biossegurança, o que impediu a análise de uma grande quantidade de amostras. A Figura 8 ilustra melhor esse processo.

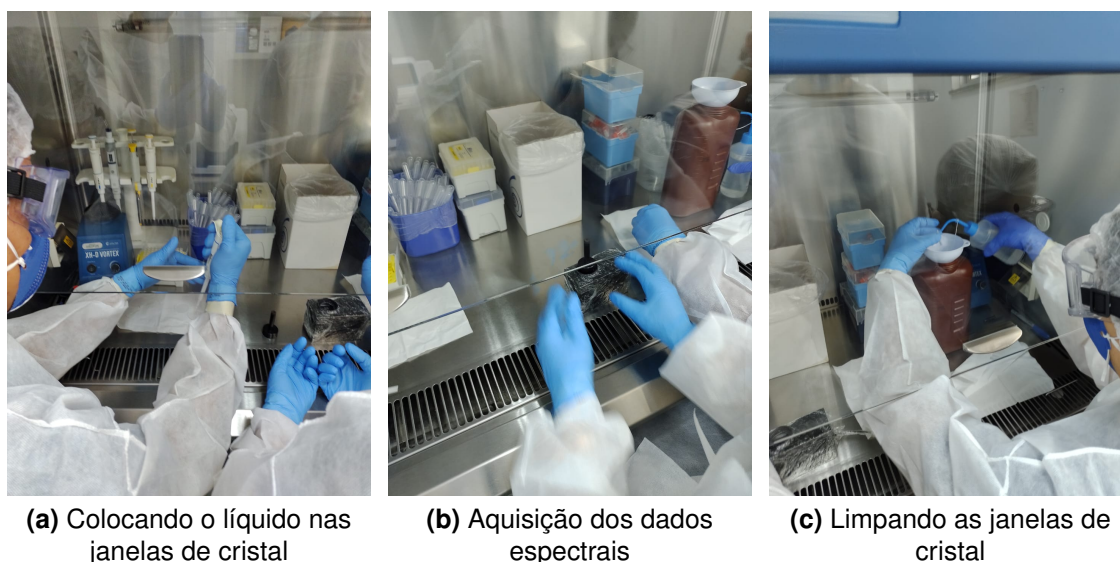


Figura 8 – Processo de aquisição de dados para o Banco de Dados 3. a) O líquido foi manipulado com uma micropipeta de volume ajustável. $5 \mu\text{L}$ de líquido foram colocados nas janelas de cristal para formar um sanduíche. b) Foi utilizada uma impressora 3D para fabricar um *case* para o espectrômetro TES. Este *case* permitiu que as placas ficassem entre o sensor do espectrômetro e um anteparo branco reflexivo. c) Logo após a aquisição das medidas, as placas foram limpas com álcool etílico (99%). Todas as medidas de biossegurança foram cuidadosamente tomadas durante essa etapa.

O Quadro 1 apresenta um resumo das metodologias adotadas para realizar a aquisição e montagem de cada um dos banco de dados mencionados. Também é

indicado o número de amostras positivas e negativas em cada situação.

Quadro 1 – Resumo das metodologias de aquisição e número de amostras para cada um dos bancos de dados montado.

	Espectrômetro	Recipiente	Nº de Amostras
Banco de Dados 1	FieldSpec 3	Tubos de ensaio	60 positivas e 463 negativas
Banco de Dados 2	TES	Tubos de ensaio	23 positivas e 129 negativas
Banco de Dados 3	TES	Janelas de cristal	2 re-amostradas 50x cada

4.3 PRÉ-PROCESSAMENTO DOS SINAIS

As análises foram realizadas utilizando a linguagem Python 3.8.8 no ambiente de desenvolvimento Spyder IDE. Para esse estudo, foram considerados apenas os dados referentes aos sinais de reflectância. Primeiramente, as regiões do espectro consideradas muito ruidosas ou sem informação foram removidas. Dessa forma, para o Banco de Dados 1, que utilizou o FieldSpec 3, foi utilizada apenas a faixa de 350-1500 nm, enquanto para os bancos de dados 2 e 3 a faixa de 900-1600 nm. A faixa espectral para o Banco de Dados 1 foi ainda dividida em duas faixas menores, sendo a primeira de 350-1000 nm e a segunda de 1000-1500 nm. Essa divisão foi necessária devido a uma descontinuidade observada nesses sinais no comprimento de onda de 1000 nm. Essa descontinuidade foi causada devido a alguma falha ou configuração do FieldSpec 3, uma vez que esse fenômeno não se repetiu para os outros bancos de dados, que utilizaram outro equipamento.

Na sequência, duas metodologias de remoção da linha de base foram aplicadas. A primeira consiste na extração da segunda derivada utilizando um filtro do tipo Savitzky-Golay com uma janela de 11 pontos e um polinômio de grau 2. Além de realizar o cálculo da segunda derivada, essa filtragem também elimina ruídos presentes no sinal original.

O segundo método consiste em aplicar esse mesmo filtro apenas para reduzir os ruídos do sinal original, sem realizar o cálculo da derivada, sendo a linha de base posteriormente removida por interpolação, seguindo a metodologia descrita por (28). Por fim, os sinais foram normalizados subtraindo de cada sinal sua média e dividindo pelo seu desvio padrão.

4.4 EXTRAÇÃO DE CARACTERÍSTICAS

A etapa de extração de características foi feita através de três diferentes técnicas: análise de componentes principais; seleção de atributos com base em testes estatísticos; energia média dos coeficientes da transformada *wavelet*.

Para o PCA, foi também analisado o comportamento dos *scores* num gráfico de dispersão e a variância explicada das componentes. Dessa forma, os *scores* das componentes que em conjunto eram responsáveis por explicar 90% da variância dos dados foram utilizados como atributos na etapa de classificação.

Outra métrica de extração de características utilizada foi a análise de variância (ANOVA). Nesse caso, os 10 comprimentos de onda que obtiveram os maiores valores de F foram utilizados como atributos. Por fim, as energias médias dos coeficientes da transformada *wavelet* com função *wavelet*-mãe do tipo Daubechies e 9 níveis de decomposição também foram utilizadas como características de entrada na etapa de classificação.

As características foram extraídas diretamente dos sinais após as etapas de pré-processamento. Dessa forma, as características supracitadas foram extraídas da segunda derivada dos sinais, bem como dos sinais que apenas tiveram sua linha de base removida por interpolação.

4.5 CLASSIFICAÇÃO E AVALIAÇÃO

Para a etapa de classificação, os classificadores SVM, *Random Forest* e *Naive Bayes* foram utilizados. Como o número de amostras em cada um dos bancos de dados é pequeno, não foi possível realizar o ajuste dos hiperparâmetros de forma eficiente, sendo os valores padrão os que demonstraram uma maior eficiência. Assim, para o classificador SVM, foi utilizado um *kernel* com função de base radial e os seguintes hiperparâmetros: $C = 1$, $gamma = 1/(n_{feat} \times X_{var})$, em que n_{feat} é o número de parâmetros da matriz de entrada e X_{var} a variância da matriz de entrada.

Para o classificador *Random Forest*, foi utilizado um modelo com 100 árvores. Já para o algoritmo *Naive Bayes*, foi utilizado um *kernel* gaussiano com uma suavização de variância de 10^{-9} .

Devido ao fato de que nos bancos de dados 1 e 2 o número de amostras negativas (N) é muito superior ao número de amostras positivas (P), foram criados 20

diferentes subgrupos formados pela seleção aleatória de P amostras testadas negativamente para Covid-19. Os classificadores foram então treinados e validados iterativamente por 20 vezes, sendo que a cada iteração um diferente subgrupo de amostras negativas era utilizado enquanto o grupo de amostras positivas se mantinha o mesmo. Essa etapa iterativa foi dispensada para o caso do Banco de Dados 3, uma vez que o número de amostras negativas e positivas é igual.

A metodologia de validação *5-fold cross-validation* foi utilizada para validar os modelos a cada iteração. A avaliação de performance dos modelos foi feita com base nas métricas de desempenho obtidas após as 20 iterações. Medidas extraídas da matriz de confusão como acurácia, sensibilidade e especificidades foram consideradas. Além disso, a curva característica de operação do receptor, bem como a área sob a curva ROC também foram avaliadas.

5 RESULTADOS

Os resultados obtidos serão apresentados em subseções separadas para cada um dos bancos de dados descritos anteriormente.

5.1 Banco de Dados 1

A Figura 9 apresenta exemplos de sinais coletados com o FieldSpec 3 antes da aplicação de qualquer tipo de pré-processamento. Esses sinais representam o comportamento dos níveis de reflectância em função do comprimento de onda da luz incidente na amostra.

De forma a eliminar os efeitos da região acima de 1500 nm e da descontinuidade em 1000 nm, o espectro foi dividido em duas faixas, de 350-1000 nm e de 1000-1500 nm, que foram então tratadas e analisadas separadamente. A seguir, serão mostrados os resultados para a faixa de 350-1000 nm e na sequência para a faixa de 1000-1500 nm.

5.1.1 Faixa de 350-1000 nm

A Figura 10 mostra exemplos de sinais após serem aplicadas as etapas de pré-processamento discutidas anteriormente.

Ao observar a Figura 10a, é possível notar que o sinal apresenta um compor-

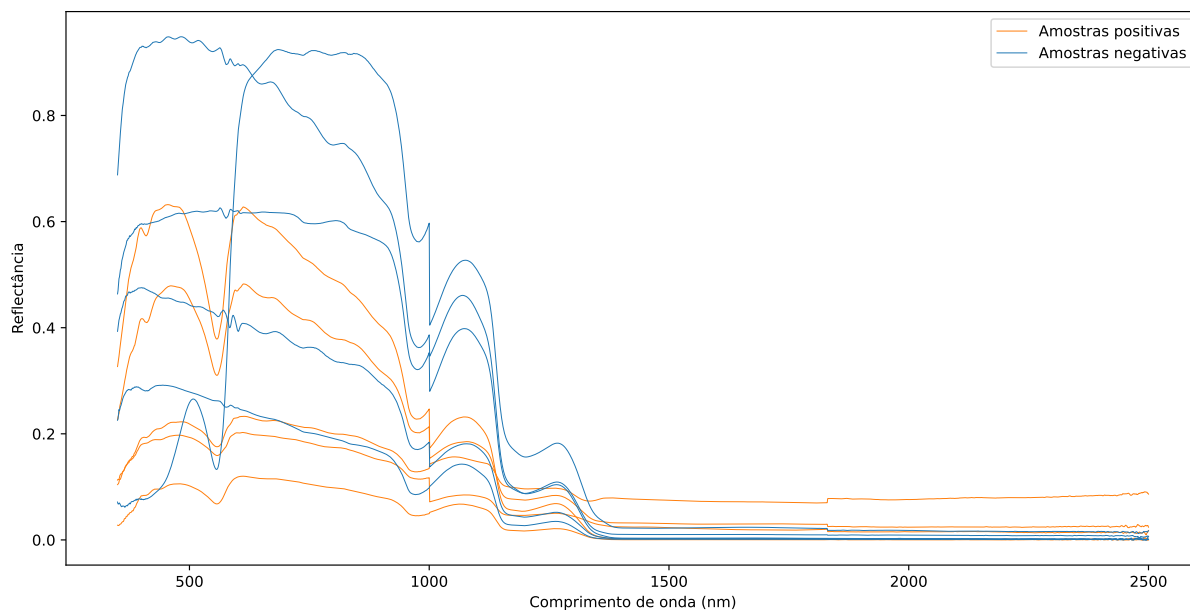


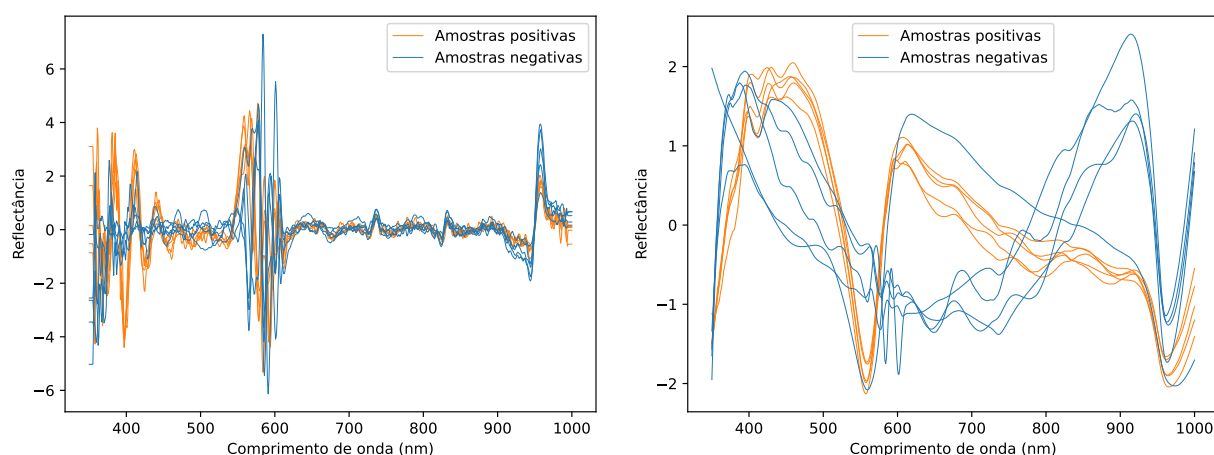
Figura 9 – Sinais coletados utilizando o FieldSpec 3. Na imagem, é possível observar uma região com pouca informação para comprimentos de onda acima de 1500 nm. Também é possível observar que o sinal apresenta uma descontinuidade em 1000 nm.

tamento oscilatório e ruidoso. Esse resultado é esperado, uma vez que o cálculo da derivada tende a amplificar variações presentes no sinal original. Ainda é possível notar que essas oscilações são mais predominantes nas regiões de 350-600 nm e acima de 900 nm. Já na Figura 10b, a remoção da linha de base possibilitou observar os picos de absorção nos sinais, que são representados pelos picos negativos de refletância. Em especial, destacam-se as regiões abaixo de 400 nm, entre 500-600 nm e entre 900-1000 nm, em que a absorção se mostrou mais intensa.

A seguir, são apresentados os resultados para quando os 10 comprimentos de onda que obtiveram o maior valor de F no teste ANOVA foram utilizados como entrada para os classificadores. A Tabela 2 e a Figura 11 mostram as medidas de desempenho extraídas da matriz de confusão e as curvas ROC para cada um dos modelos.

Apesar do alto valor de especificidade encontrado para o classificador do tipo SVM em conjunto com os comprimentos de onda extraídos do sinal cuja a linha de base foi removida por interpolação ($ESP = 0,73 \pm 0,10$), esse resultado não pode ser considerado relevante, uma vez que foi acompanhado de uma $ACC = 0,49 \pm 0,06$, indicando que as predições foram feitas de forma aleatória.

Outra metodologia de extração de características utilizada foi a energia dos coeficientes da transformada *wavelet*. A Figura 12 apresenta o comportamento dos coefi-



(a) Segunda derivada do sinal obtida através da aplicação de um filtro Savitzky-Golay.

(b) Sinal após o processo de remoção da linha de base por interpolação.

Figura 10 – Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.

Tabela 2 – Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Apesar do desempenho dos classificadores ter sido ruim, com acurácia em torno de 50%, destaca-se a especificidade média de 73% obtida com o classificador SVM para o caso em que os sinais tiveram sua linha de base removida por interpolação.

		SEN	ESP	ACC
2ª Derivada	SVM	0,44 ± 0,08	0,62 ± 0,06	0,53 ± 0,05
	RF	0,47 ± 0,07	0,63 ± 0,05	0,55 ± 0,06
	<i>Naive Bayes</i>	0,49 ± 0,06	0,56 ± 0,06	0,52 ± 0,04
Interpolação	SVM	0,27 ± 0,11	0,73 ± 0,10	0,49 ± 0,06
	RF	0,31 ± 0,08	0,60 ± 0,07	0,45 ± 0,06
	<i>Naive Bayes</i>	0,39 ± 0,10	0,57 ± 0,07	0,48 ± 0,06

cientes da transformada para ambos os casos.

Para formar os vetores de entrada dos classificadores, foi calculada a energia média dos coeficientes para cada um dos níveis de decomposição. Dessa forma, um vetor de dimensão 9 (correspondente ao número de níveis utilizado) foi utilizado como atributo de entrada. A Tabela 3 e a Figura 13 apresentam os resultados obtidos pelos 3 classificadores para cada metodologia de pré-processamento.

Outra vez, todos os modelos apresentaram resultados ruins, com acurácia em torno de 50%, indicando que a maioria das predições foram atribuídas de maneira aleatória. Entretanto, assim como no caso anterior, o classificador SVM conseguiu uma alta especificidade ($ESP = 0,71 \pm 0,09$) quando utilizados os comprimentos de onda do sinal com a linha de base removida por interpolação. Como esse valor é acompanhado por uma baixa sensibilidade ($SEN = 0,32 \pm 0,07$) e consequentemente

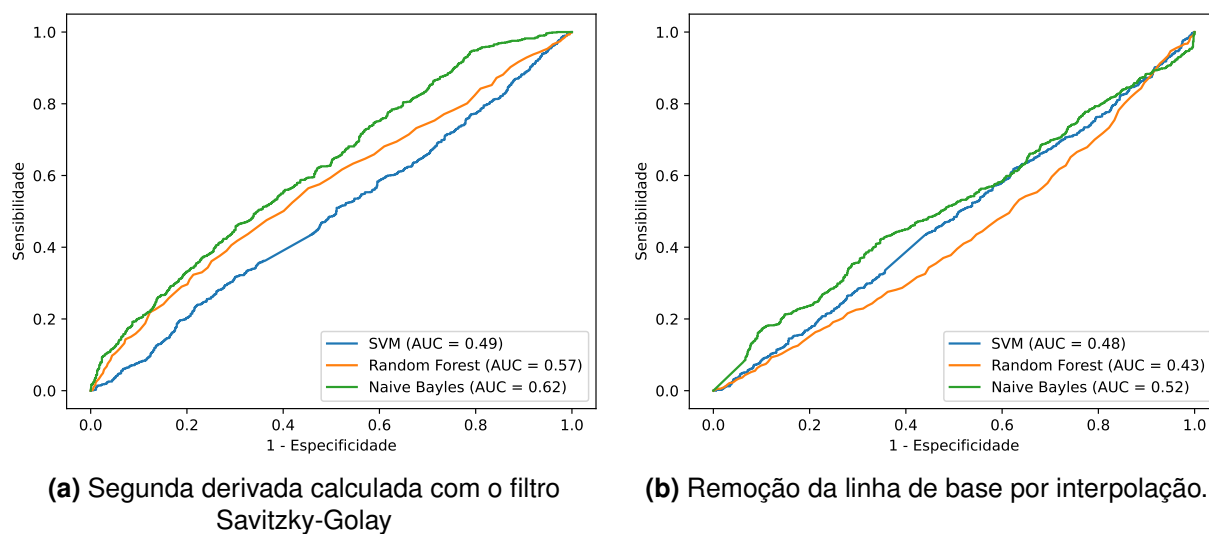


Figura 11 – Curvas ROC obtidas para os 3 classificadores utilizados. É possível observar que todos os classificadores alcançaram um baixo desempenho, com o melhor resultado sendo obtido através da segunda derivada e do classificador *Naive Bayes*, com $AUC = 0,62$.

Tabela 3 – Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Nesse caso foram utilizadas as energias médias para cada nível de decomposição da transformada *wavelet*.

		SEN	ESP	ACC
2ª Derivada	SVM	0,34 ± 0,06	0,68 ± 0,08	0,51 ± 0,06
	RF	0,43 ± 0,09	0,66 ± 0,05	0,55 ± 0,07
	<i>Naive Bayes</i>	0,46 ± 0,09	0,58 ± 0,07	0,52 ± 0,06
Interpolação	SVM	0,32 ± 0,07	0,71 ± 0,09	0,52 ± 0,05
	RF	0,48 ± 0,06	0,67 ± 0,07	0,57 ± 0,05
	<i>Naive Bayes</i>	0,49 ± 0,07	0,63 ± 0,08	0,56 ± 0,07

de uma baixa acurácia ($ACC = 0,52 \pm 0,05$), esse resultado deve ser considerado pouco relevante.

Por fim, os *scores* obtidos através do cálculo das componentes principais também foram utilizados como entrada para os classificadores. O número de componentes utilizado foi determinado através da variância explicada acumulada.

Pelo critério estabelecido, foram utilizadas como entrada para os classificadores 12 componentes no primeiro caso, e apenas 4 componentes no segundo. Esse fenômeno é explicado ao analisar o comportamento desses sinais na Figura 10. A segunda derivada (Figura 10a) apresenta um comportamento muito mais oscilatório do que o sinal cuja linha de base foi removida por interpolação (Figura 10b). Essa maior oscilação também implica uma maior 'variância', e devido a isso, mais componentes foram necessárias para explicar essa variação. A Tabela 4 e a Figura 14 apresen-

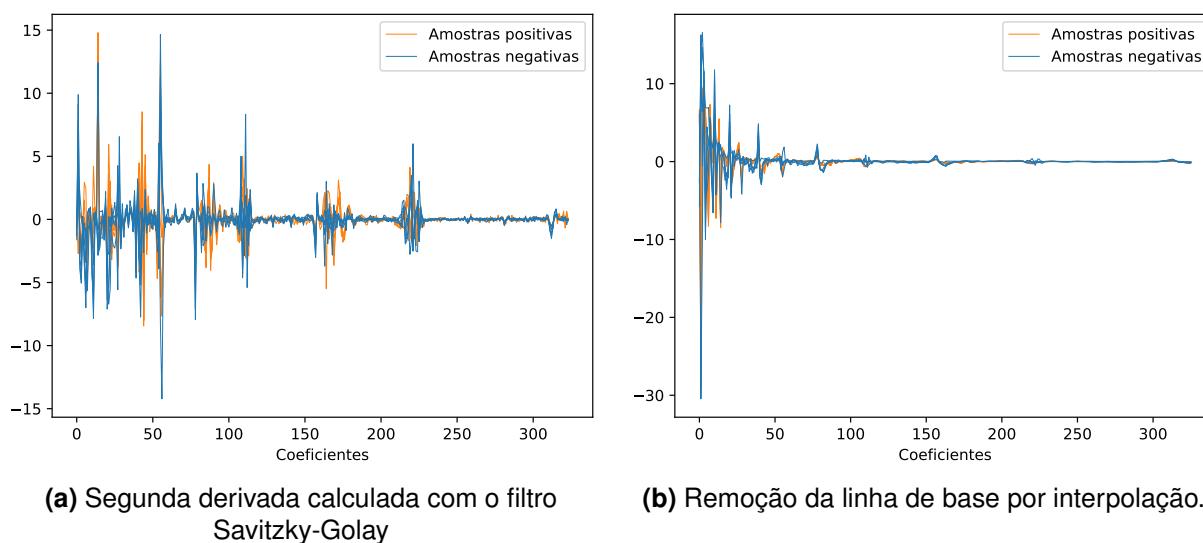


Figura 12 – Gráficos dos coeficientes da transformada *wavelet* para as duas metodologias de pré-processamento. Foram utilizados 9 níveis de decomposição e uma função *wavelet-mãe* do tipo Daubechies 1.

tam os resultados obtidos para o uso dos *scores* das componentes principais como atributos de entrada.

Tabela 4 – Resumo das métricas extraídas da matriz de confusão após as 20 iterações do processo de classificação. Os *scores* das componentes principais foram usados como parâmetros de entrada.

		SEN	ESP	ACC
2ª Derivada	SVM	0,28 ± 0,03	0,57 ± 0,07	0,42 ± 0,04
	RF	0,38 ± 0,07	0,56 ± 0,07	0,47 ± 0,04
	<i>Naive Bayes</i>	0,44 ± 0,06	0,54 ± 0,07	0,49 ± 0,05
Interpolação	SVM	0,26 ± 0,06	0,66 ± 0,11	0,47 ± 0,06
	RF	0,37 ± 0,07	0,6 ± 0,07	0,48 ± 0,07
	<i>Naive Bayes</i>	0,41 ± 0,07	0,58 ± 0,09	0,49 ± 0,07

Os resultados obtidos mostram que o processo de classificação utilizando componentes principais teve um péssimo desempenho, com acurácia sempre abaixo de 50%. Entretanto, ao analisar o comportamento da projeção dos *scores* das componentes principais dos sinais, evidências pertinentes foram encontradas, como mostra a Figura 15.

Ao observar a Figura 15a é possível notar que não há uma distinção entre as amostras positivas e negativas na projeção das componentes principais. Entretanto, por inspeção, percebe-se que existem 3 grupos distintos de amostras. Devido a isso, um classificador não supervisionado do tipo *k-means* foi utilizado para fazer o agrupamento desses dados e o resultado é mostrado na Figura 15b.

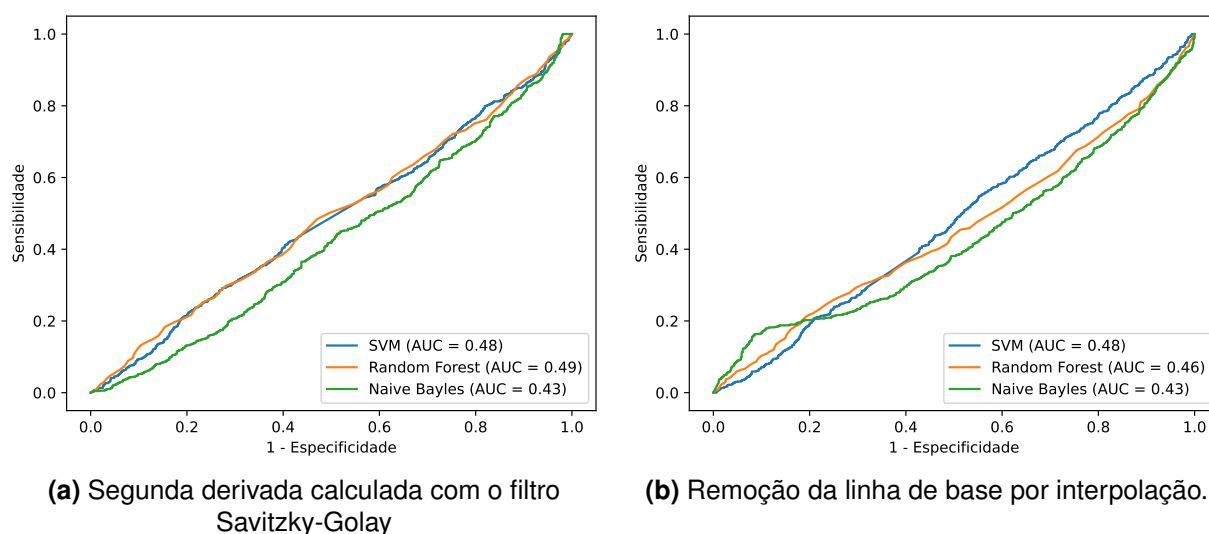


Figura 13 – Curvas ROC obtidas para os 3 classificadores utilizados. O desempenho geral foi ruim em todas as situações, com nenhum modelo alcançando uma AUC superior a 0,5.

Diante disso, foi feita uma análise para verificar quais amostras faziam parte de cada um desses grupos. Assim, percebeu-se que o Grupo 1 era formado pelas amostras que utilizaram a solução salina, enquanto o Grupo 2 representavam as amostras que utilizaram o MTV Laborclin. O Grupo 3 é formado pelas amostras coletadas nos dias 8/12/2021 e 10/12/2021 que utilizaram a solução salina. Os grupos possuem diferentes tamanhos: o Grupo 1 é formado por 17 amostras positivas e 148 negativas, o Grupo 2 por 18 positivas e 153 negativas e o Grupo 3 por 10 positivas e 25 negativas.

Esse agrupamento se deu possivelmente pelo fato dessa faixa espectral incluir os comprimentos de onda do visível, e, uma vez que o MTV Laborclin possui uma coloração avermelhada enquanto a solução salina de NaCl uma coloração transparente, a reflexão na faixa do visível era determinada majoritariamente pelo tipo de solução de transporte utilizada, e não pelo material biológico presente nas amostras. Ainda não se sabe o que ocasionou o agrupamento das amostras coletadas nos dias 8/12/2021 e 10/12/2021 que utilizaram a solução salina como solução de transporte.

Essa influência da solução de transporte também pode explicar os resultados ruins alcançados pelos classificadores. Dessa forma, com o intuito de realizar um estudo em que os efeitos da solução utilizada sejam reduzidos, as análises feitas anteriormente foram refeitas considerando os 3 grupos separadamente.

Para o Grupo 1, os melhores resultados foram encontrados utilizando o classificador *Naive Bayes* tendo como parâmetros de entrada os comprimentos de onda

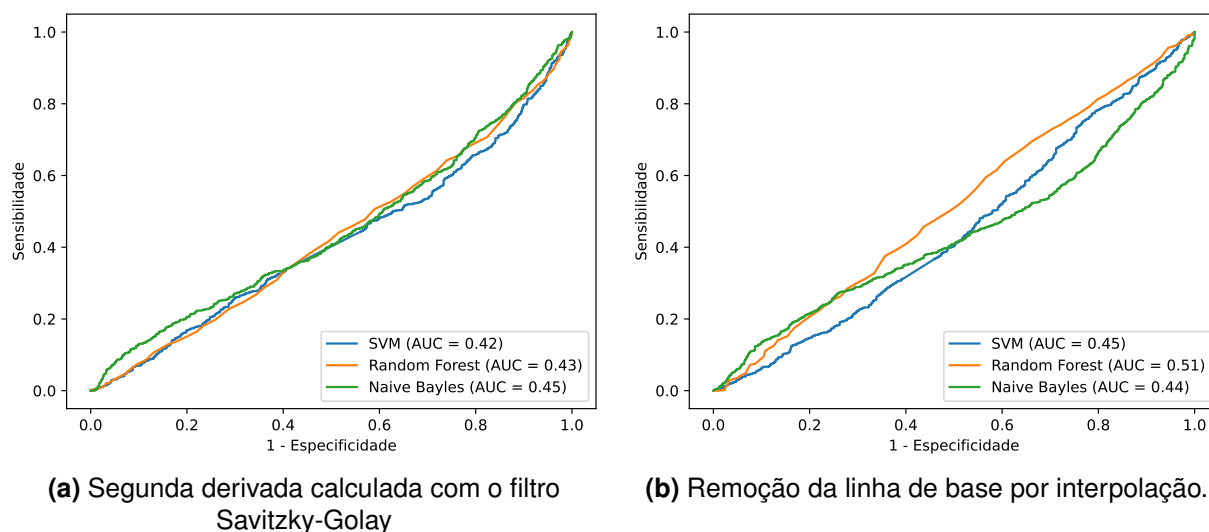


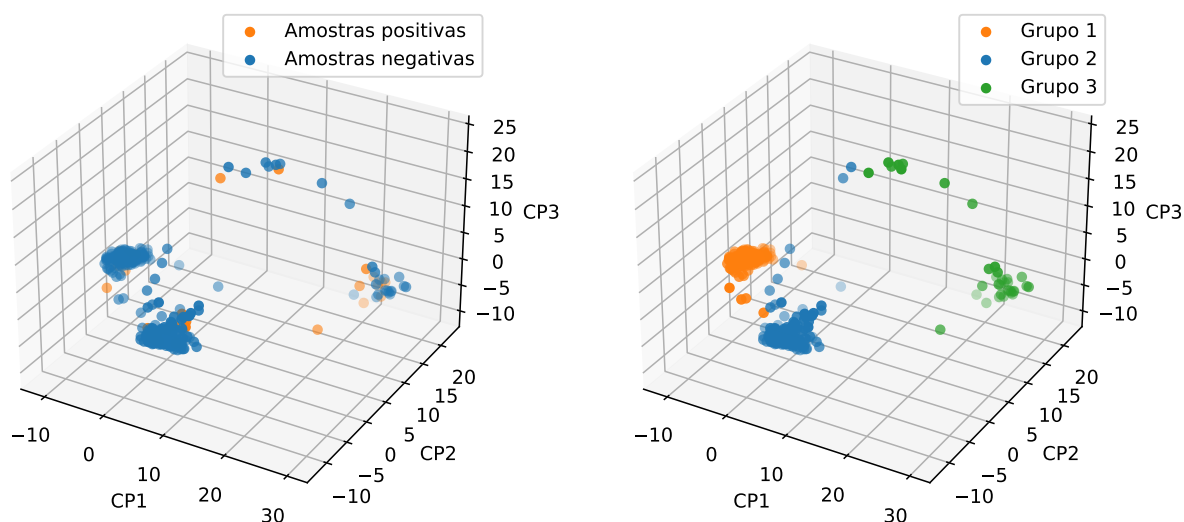
Figura 14 – Curvas ROC obtidas pelos 3 classificadores em cada caso. É possível observar que todos os classificadores atribuíram predições aleatórias.

com os maiores valores de F , calculados nos sinais cuja linha de base foi removida por interpolação. As métricas obtidas nesse cenário foram: $ACC = 0,68 \pm 0,05$; $SEN = 0,62 \pm 0,05$; $ESP = 0,73 \pm 0,08$; $AUC = 0,70$. Para o Grupo 2, o melhor desempenho foi alcançado com mesma estratégia, mas dessa vez com o classificador *Random Forest* e a segunda derivada dos sinais, alcançando as métricas: $ACC = 0,68 \pm 0,08$; $SEN = 0,64 \pm 0,09$; $ESP = 0,72 \pm 0,07$; $AUC = 0,74$.

Esses resultados corroboram com a hipótese da influência da solução de transporte na aquisição dos dados na faixa espectral do visível. É possível notar que essas acurácias indicam que os classificadores não mais atribuíram predições de maneira aleatória, como no caso em que as amostras foram analisadas sem fazer distinção entre o tipo de solução de transporte utilizada.

Os melhores resultados foram obtidos para o Grupo 3, e por isso serão apresentados em mais detalhes. A Tabela 5 e a Figura 16 apresentam os resultados para o caso em que os 10 comprimentos de onda com maior valor de F para o teste ANOVA foram utilizados como parâmetro de entrada.

A partir dos resultados alcançados para o Grupo 3, é perceptível que os classificadores são capazes de distinguir as amostras positivas das negativas com boa acurácia. Ao analisar as curvas ROC da Figura 16, nota-se que os classificadores *Naive Bayes* e *SVM* conseguiram um desempenho melhor do que o classificador *Random Forest* em ambas as situações.



(a) Amostras agrupadas entre amostras positivas e negativas. (b) Amostras agrupadas em 3 subgrupos com o auxílio de um classificador não supervisionado.

Figura 15 – Distribuição das amostras com base nos *scores* de suas componentes principais. No gráfico, são representadas as 3 componentes de maior variância explicada.

Tabela 5 – Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os 10 comprimentos de onda com maior valor de F . Houve uma melhora considerável nas acurácias obtidas, com destaque para o modelo em que a segunda derivada dos sinais foi usada em conjunto com os classificadores SVM e *Naive Bayes*, obtendo uma acurácia de 91% e 93% respectivamente. Também é possível notar que a segunda derivada obteve um melhor desempenho geral.

		SEN	ESP	ACC
2ª Derivada	SVM	0,98 ± 0,04	0,83 ± 0,09	0,91 ± 0,01
	RF	0,83 ± 0,10	0,84 ± 0,10	0,83 ± 0,09
	<i>Naive Bayes</i>	0,96 ± 0,07	0,89 ± 0,09	0,93 ± 0,07
Interpolação	SVM	0,75 ± 0,06	0,75 ± 0,13	0,75 ± 0,09
	RF	0,76 ± 0,07	0,76 ± 0,10	0,75 ± 0,10
	<i>Naive Bayes</i>	0,78 ± 0,05	0,79 ± 0,09	0,78 ± 0,07

Para a situação em que a energia dos coeficientes da transformada *wavelet* foi utilizada, os melhores resultados foram obtidos do sinal cuja linha de base foi removida por interpolação juntamente com o classificador do tipo SVM (ACC = 0,72 ± 0,07; SEN = 0,74 ± 0,07; ESP = 0,68 ± 0,12; AUC = 0,75). Essa mesma configuração também foi a que alcançou as melhores métricas quando utilizados os *scores* das componentes principais: ACC = 0,75 ± 0,10; SEN = 0,79 ± 0,11; ESP = 0,72 ± 0,12; AUC = 0,61.

Para esse grupo, foi possível notar uma melhora significativa nos resultados quando comparados ao valores obtidos para os Grupos 1 e 2. A melhor acurácia

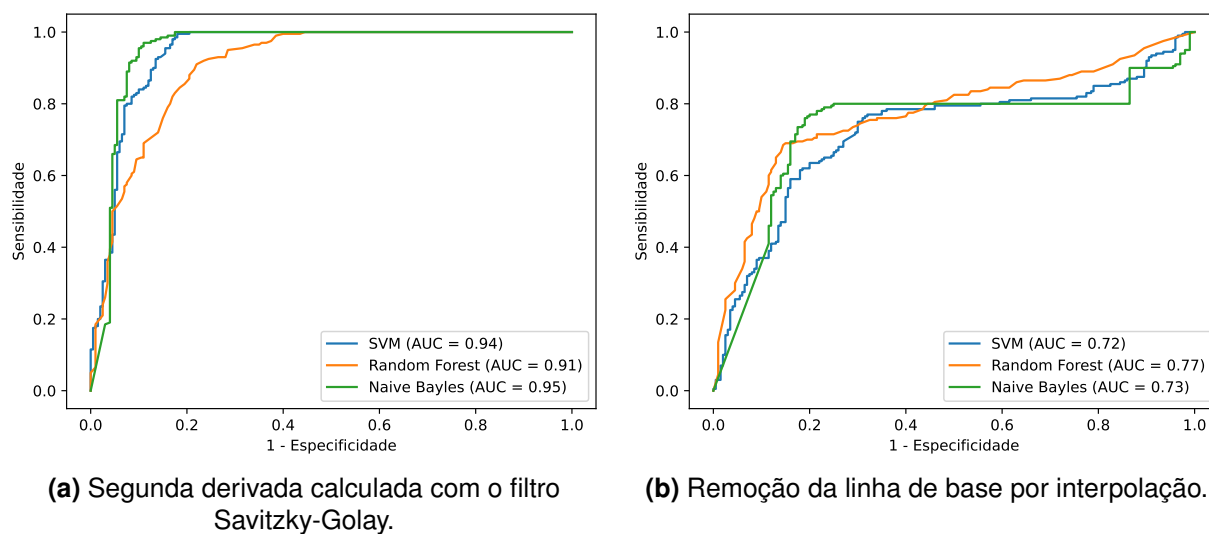


Figura 16 – Curvas ROC obtidas pelos 3 classificadores em cada caso. Os classificadores obtiveram um ótimo desempenho, principalmente quando utilizada a segunda derivada dos sinais.

obtida, de 93%, equipara-se aos valores encontrados em testes vendidos comercialmente. Vale ainda ressaltar que esses resultados foram obtidos apenas para as amostras coletadas nos dias 08/12/2021 e 10/12/2021 que utilizaram a solução salina como solução de transporte. Ainda não se sabe o porquê dessas amostras se comportarem dessa forma, uma vez que a metodologia de aquisição se manteve ao longo de todo o período de coleta.

5.1.2 Faixa de 1000-1500 nm

A Figura 17 apresenta exemplos de sinais após a etapa de pré-processamento para a faixa de 1000-1500 nm.

Da Figura 17a, é possível observar que o sinal apresenta uma alta oscilação no intervalo entre 1100 nm e 1200 nm. Esse mesmo intervalo também representa o pico de absorção percebido na Figura 17b. Os resultados obtidos na etapa de classificação utilizando todas as amostras nessa faixa espectral não foram bons, sendo o melhor desempenho alcançado pelo classificador do tipo SVM quando, mais uma vez, utilizados os comprimentos de onda da segunda derivada selecionados com o teste ANOVA. As métricas de avaliação para esse caso foram: $ACC = 0,62 \pm 0,06$; $SEN = 0,58 \pm 0,07$; $ESP = 0,66 \pm 0,06$ e $AUC = 0,63$.

Assim como na análise da faixa espectral de 350-1000 nm, também foi avaliado o desempenho dos classificadores ao considerar os grupos de amostras formados pelo algoritmo de treinamento não supervisionado. Para o Grupo 1, os melhores re-

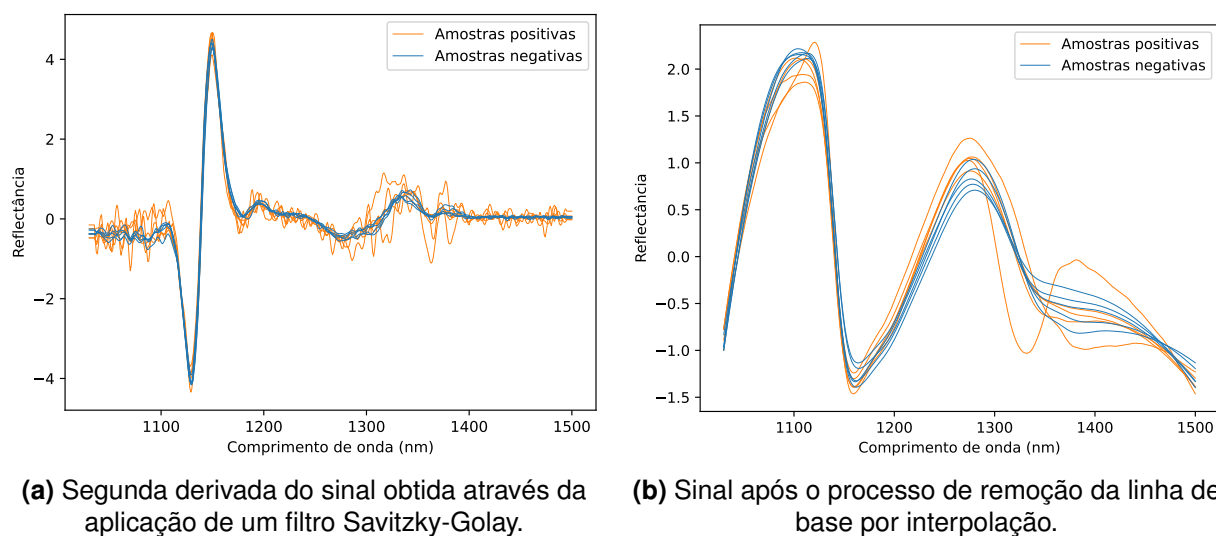


Figura 17 – Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.

sultados foram obtidos utilizando o classificador SVM em conjunto com os comprimentos de onda da segunda derivada do sinal que alcançaram os maiores valores de F . Para esse caso, as métricas obtidas foram: $ACC = 0,70 \pm 0,08$; $SEN = 0,71 \pm 0,10$; $ESP = 0,68 \pm 0,10$; $AUC = 0,76$. Para o Grupo 2, o melhor desempenho foi obtido com mesma estratégia, mas dessa vez o classificador *Random Forest* alcançou as melhores métricas: $ACC = 0,66 \pm 0,10$; $SEN = 0,68 \pm 0,16$; $ESP = 0,58 \pm 0,09$; $AUC = 0,71$.

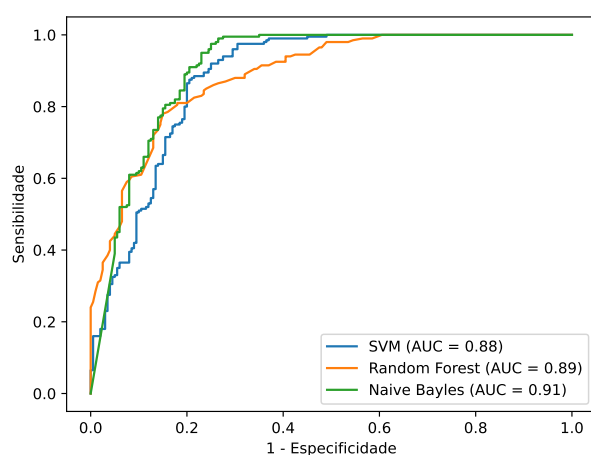
Da mesma forma que ocorreu na análise da faixa de 350-1000 nm, os classificadores deixaram de apenas atribuir previsões aleatórias e passaram a operar com desempenho razoável. Esses resultados mostram que a solução de transporte também influencia a análise dos dados fora da faixa do visível.

Mais uma vez, o melhor desempenho foi obtido na análise do Grupo 3. A Tabela 6 e a Figura 18 apresentam os resultados para o caso em que os 10 comprimentos de onda com maior valor de F para o teste ANOVA foram utilizados como parâmetros de entrada.

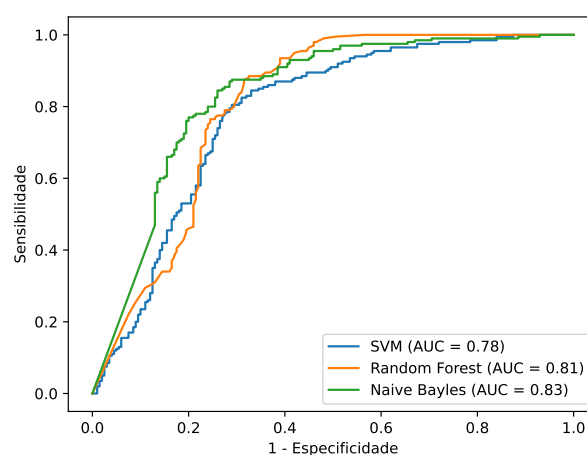
Vale observar que a separação das amostras nos 3 diferentes grupos de acordo com a solução de transporte utilizada e data da coleta não pode ser feita utilizando essa faixa espectral. A Figura 19 apresenta a projeção das componentes principais para as duas metodologias de pré-processamento, bem como os grupos formados pelo algoritmo de classificação não supervisionado.

Tabela 6 – Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os 10 comprimentos de onda com maior valor de F . Houve uma melhora considerável nas acurácias obtidas, com destaque para o modelo em que a segunda derivada dos sinais foi usada em conjunto com os classificadores SVM e *Naive Bayes*, obtendo uma acurácia de 83%. Também é possível notar que a segunda derivada obteve um melhor desempenho geral.

		SEN	ESP	ACC
2ª Derivada	SVM	$0,85 \pm 0,07$	$0,80 \pm 0,08$	$0,83 \pm 0,05$
	RF	$0,83 \pm 0,06$	$0,79 \pm 0,04$	$0,81 \pm 0,05$
	<i>Naive Bayes</i>	$0,81 \pm 0,08$	$0,84 \pm 0,07$	$0,83 \pm 0,06$
Interpolação	SVM	$0,82 \pm 0,06$	$0,76 \pm 0,10$	$0,79 \pm 0,06$
	RF	$0,72 \pm 0,11$	$0,76 \pm 0,10$	$0,75 \pm 0,10$
	<i>Naive Bayes</i>	$0,73 \pm 0,10$	$0,80 \pm 0,07$	$0,77 \pm 0,08$

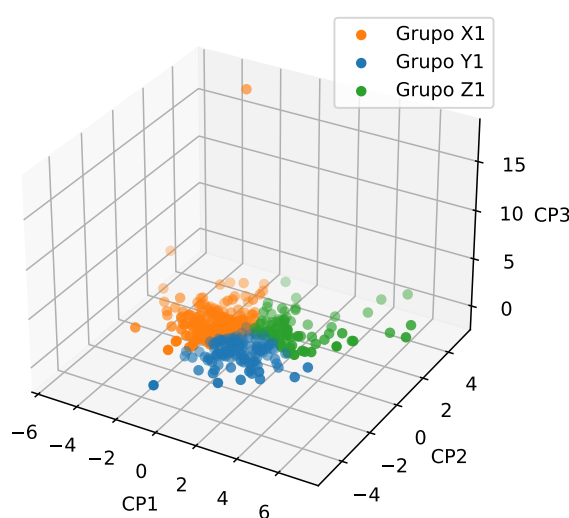


(a) Segunda derivada calculada com o filtro Savitzky-Golay.

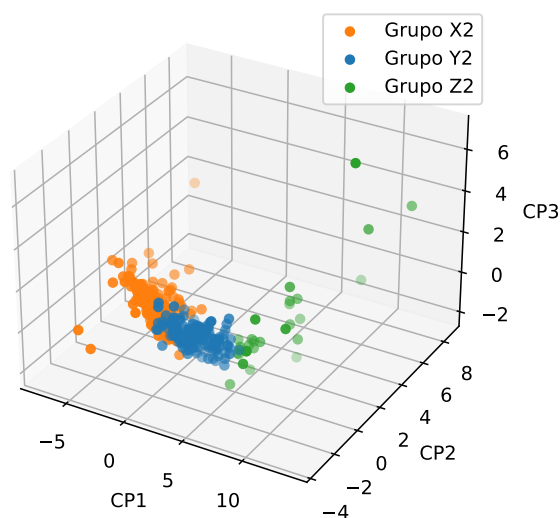


(b) Remoção da linha de base por interpolação.

Figura 18 – Curvas ROC obtidas pelos 3 classificadores em cada caso. Os classificadores obtiveram um ótimo desempenho, principalmente quando utilizada a segunda derivada dos sinais.



(a) Segunda derivada calculada com o filtro Savitzky-Golay.



(b) Remoção da linha de base por interpolação.

Figura 19 – Projeção das 3 componentes principais com maior variância explicada. Nesse caso, os grupos formados pelo classificador *k-means* não representa nenhum conjunto de dados específico.

Diferentemente da análise para a faixa de 350-1000 nm, não é possível identificar agrupamentos nos dados projetados. Dessa forma, o classificador apenas agrupou as amostras mais próximas entre si, sendo que os grupos formados não correspondem a nenhum conjunto de amostras em específico. A partir dessas observações, é possível inferir que a solução de transporte possui uma influência menor nessa faixa espectral, uma vez que não é possível separar as amostras com base na solução de transporte utilizada através de um algoritmo de classificação não supervisionado. Entretanto, essa influência ainda existe, evidenciada pelo desempenho superior dos classificadores quando analisadas amostras que utilizaram a mesma solução de transporte.

5.2 Banco de Dados 2

A Figura 20 apresenta exemplos de sinais de reflectância obtidos com o espectrômetro TellSpec Enterprise Sensor.

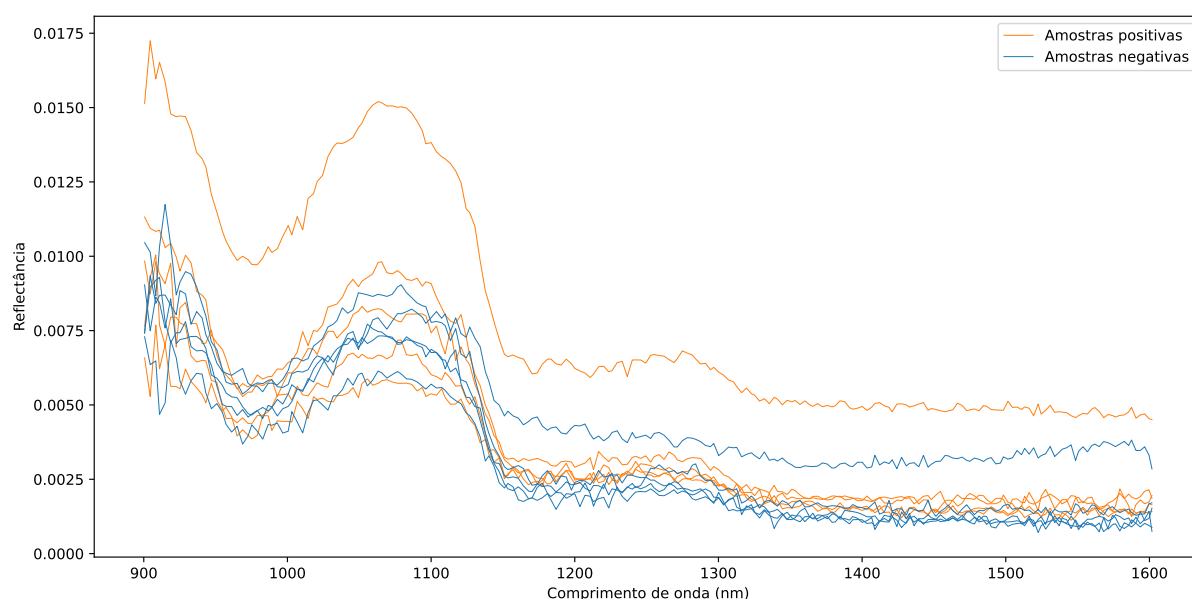


Figura 20 – Sinais coletados utilizando o TellSpec Enterprise Sensor. Diferentemente do caso em que foi utilizado o FieldSpec 3, os sinais agora não apresentam uma descontinuidade em 1000 nm.

Ao observar a Figura 20, é possível notar que a ordem de grandeza dos sinais coletados com esse equipamento é menor do que a dos sinais coletados no Banco de Dados 1. Isso ocorre não apenas devido ao uso de um equipamento diferente, mas também pela nova configuração e posicionamento dos sensores e da fonte luminosa, como mostrado anteriormente nas Figuras 6 e 7. Esses sinais também aparentam ser mais ruidosos, o que se dá por conta da resolução espectral desse

equipamento ser inferior. Na Figura 21, são apresentados os sinais após as etapas de pré-processamento e normalização.

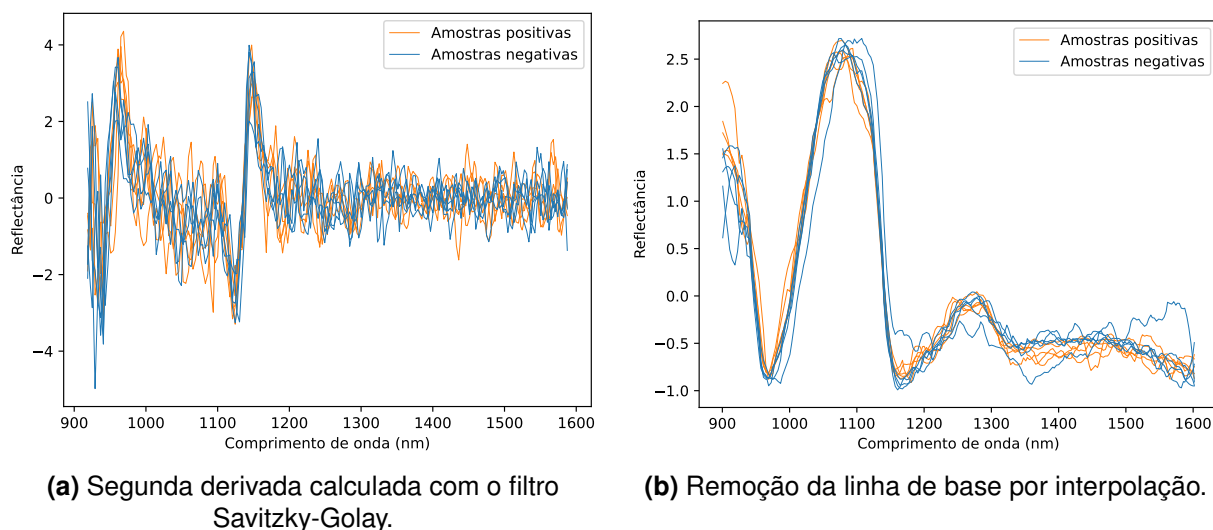


Figura 21 – Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.

Da Figura 21a, percebe-se que a segunda derivada mais uma vez amplificou os ruídos do sinal original. Nota-se também uma grande variação na região próxima a 1000 nm e na região entre 1100-1200 nm. Essas variações são observadas como picos de absorção na Figura 21b. Quando comparados os resultados da Figura 21 com os da Figura 17, é possível notar uma equivalência entre os sinais obtidos pelos dois equipamentos para a faixa entre 1000 e 1500 nm. Dessa forma, mesmo que a ordem de grandeza dos sinais coletados com o Telspec Enterprise Sensor seja menor, quando pré-processados e normalizados, os sinais obtidos com ambos os equipamentos são semelhantes, o que é esperado, uma vez que se trata da mesma faixa espectral.

Na etapa de classificação, os melhores resultados foram alcançados utilizando os comprimentos de onda da segunda derivada do sinal em conjunto com um classificador do tipo *Random Forest*. Os parâmetros de desempenho para esse caso foram: $ACC = 0,64 \pm 0,08$; $SEN = 0,68 \pm 0,12$; $ESP = 0,57 \pm 0,08$ e $AUC = 0,67$. Esses resultados são superiores aos obtidos com o espectrômetro FieldSpec 3 para a faixa espectral de 1000-1500 nm. Vale ainda ressaltar que quando a metodologia de pré-processamento foi a remoção da linha de base por interpolação, todos os classificadores tenderam para a aleatoriedade, com acurácias em torno de 50% em todas as

situações.

Ao analisar o comportamento da projeção das componentes principais, não foi possível separar os grupos de acordo com solução de transporte utilizada, como mostrado na Figura 22. Entretanto, dois grupos foram criados manualmente, um com as amostras que utilizaram a solução salina e o outro com as amostras que utilizaram o MTV Laborclin. O primeiro grupo possui 103 amostras sendo 14 positivas e 89 negativas, enquanto o segundo grupo possui 10 amostras positivas e 39 amostras negativas.

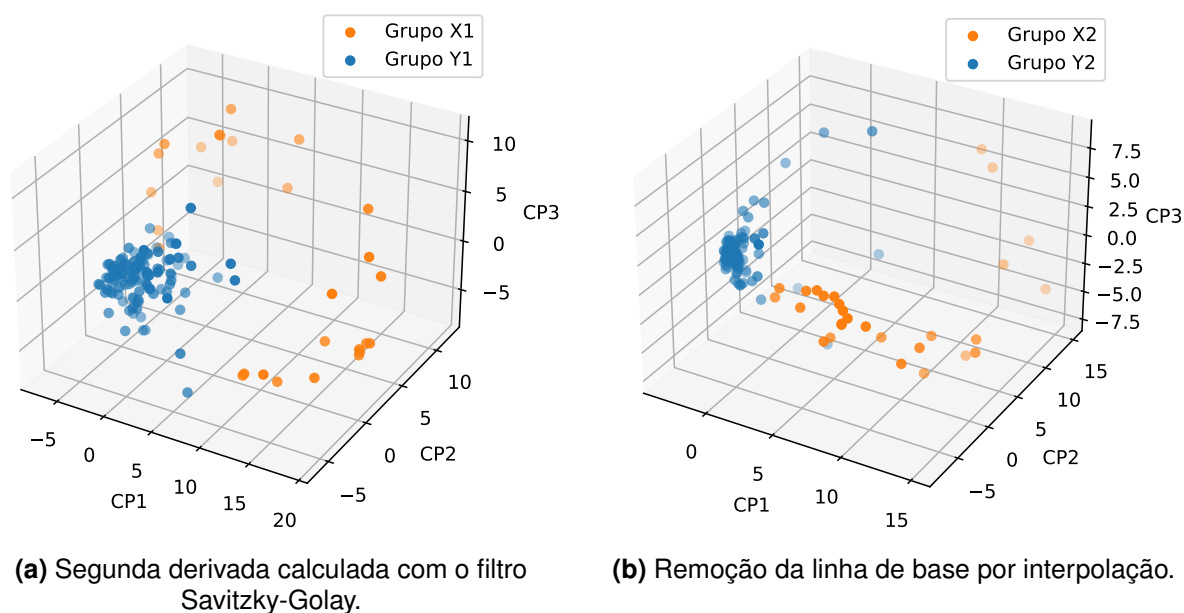


Figura 22 – Grupos formados pelo classificador K-means. Uma primeira inspeção visual identificou a possível formação de dois grupos, por isso os parâmetros do classificador foram ajustados para a identificação de apenas dois. Apesar do agrupamento, os grupos formados não correspondem à amostras com metadados em comum.

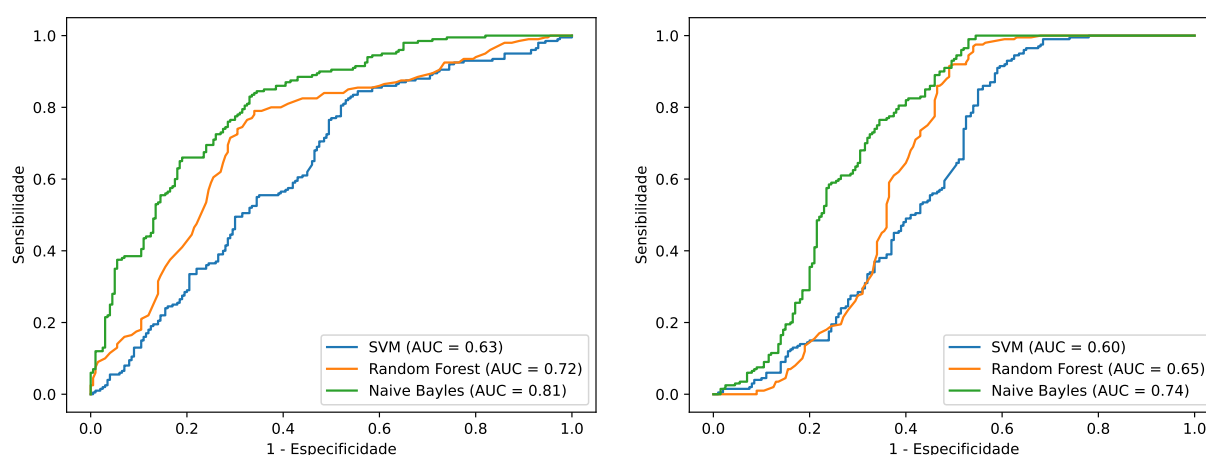
Para o grupo formado pelas amostras que utilizaram o a solução salina de NaCl, o melhor resultado foi obtido utilizando o classificador *Random Forest* em conjunto com os comprimentos de onda extraídos da segunda derivada dos sinais. As métricas para esse caso foram: $ACC = 0,67 \pm 0,14$, $SEN = 0,73 \pm 0,16$, $ESP = 0,66 \pm 0,16$ e $AUC = 0,74$.

Já o caso em que foram analisadas as amostras que utilizaram o MTV Laborclin, em especial quando utilizados os comprimentos de onda selecionados com o teste ANOVA, os resultados alcançaram um melhor desempenho e estão apresentados na Tabela 7 e na Figura 23.

Da Tabela 7, é possível notar que os modelos alcançaram bons resultados, com

Tabela 7 – Resumo das métricas extraídas da matriz de confusão após a etapa de classificação usando os comprimentos de onda com maior valor de F . Ambas as metodologias de pré-processamento alcançaram acurácias semelhantes. Entretanto, os sinais que tiveram a linha de base removida por interpolação obtiveram uma sensibilidade maior.

		SEN	ESP	ACC
2ª Derivada	SVM	$0,77 \pm 0,08$	$0,67 \pm 0,06$	$0,72 \pm 0,12$
	RF	$0,74 \pm 0,07$	$0,68 \pm 0,06$	$0,72 \pm 0,10$
	<i>Naive Bayes</i>	$0,79 \pm 0,04$	$0,70 \pm 0,06$	$0,74 \pm 0,06$
Interpolação	SVM	$0,94 \pm 0,04$	$0,53 \pm 0,07$	$0,74 \pm 0,10$
	RF	$0,69 \pm 0,08$	$0,58 \pm 0,06$	$0,64 \pm 0,12$
	<i>Naive Bayes</i>	$0,82 \pm 0,04$	$0,60 \pm 0,05$	$0,71 \pm 0,07$



(a) Segunda derivada calculada com o filtro Savitzky-Golay.

(b) Remoção da linha de base por interpolação.

Figura 23 – Curvas ROC obtidas pelos 3 classificadores em cada caso. O classificador do tipo *Naive Bayes* alcançou os maiores valores de área sob a curva em ambos os cenários, com $AUC = 0,81$ e $AUC = 0,74$.

acurácias acima de 70% na maioria dos casos. Também vale destacar a alta sensibilidade alcançada pelos classificadores SVM ($SEN = 0,94 \pm 0,04$) e *Naive Bayes* ($SEN = 0,82 \pm 0,04$) quando utilizados os sinais que tiveram a linha de base removida por interpolação. Essa alta sensibilidade veio acompanhada de uma especificidade baixa, mas sem comprometer a acurácia. Já ao observar a Figura 23, percebe-se que as áreas sobre a curva são maiores quando utilizada a segunda derivada, sugerindo que esse método de pré-processamento pode ser o mais eficiente para esse banco de dados.

5.3 Banco de Dados 3

Esse banco de dados foi montado com o intuito de avaliar a influência do vidro dos tubos de ensaio na coleta dos dados espectroscópicos. As janelas de cristal utilizadas se apresentam como uma solução ideal para esse propósito, uma vez que possuem uma transmitância total nessa faixa espectral, ou seja, não refletem nem absorvem esses comprimentos de onda.

Além disso, o volume de líquido (*swab* + solução de transporte) utilizado é padronizado ($5 \mu L$), evitando interferências que possam vir a acontecer devido a variação de volume de líquido nas amostras. A Figura 24 apresenta exemplos de sinais espectrais coletados para o Banco de Dados 3 antes das etapas de pré-processamento.

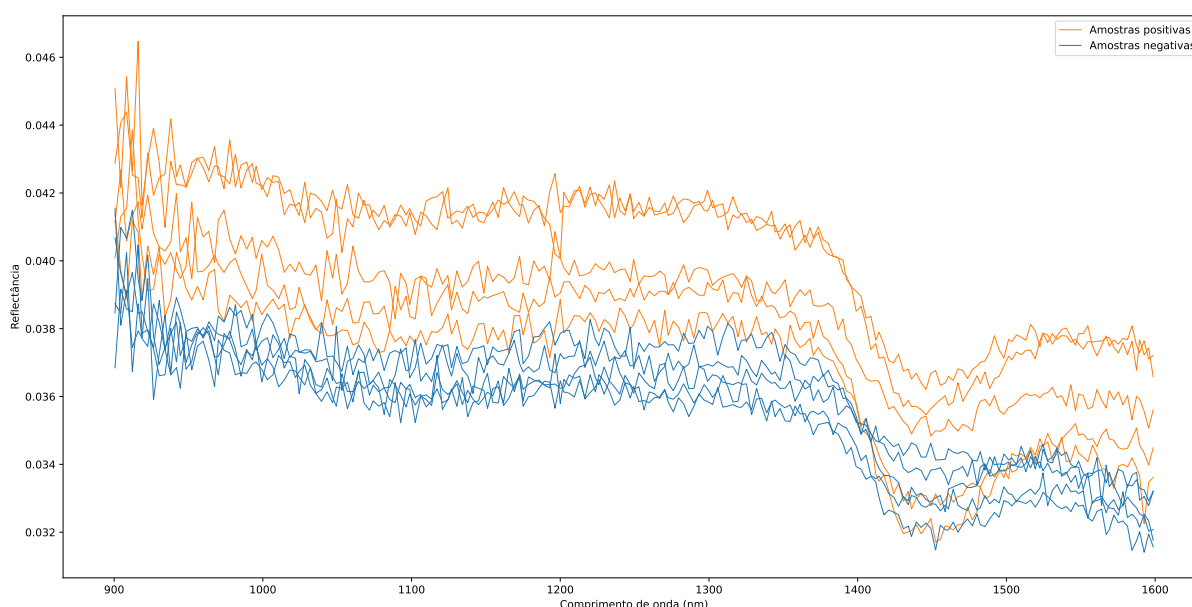
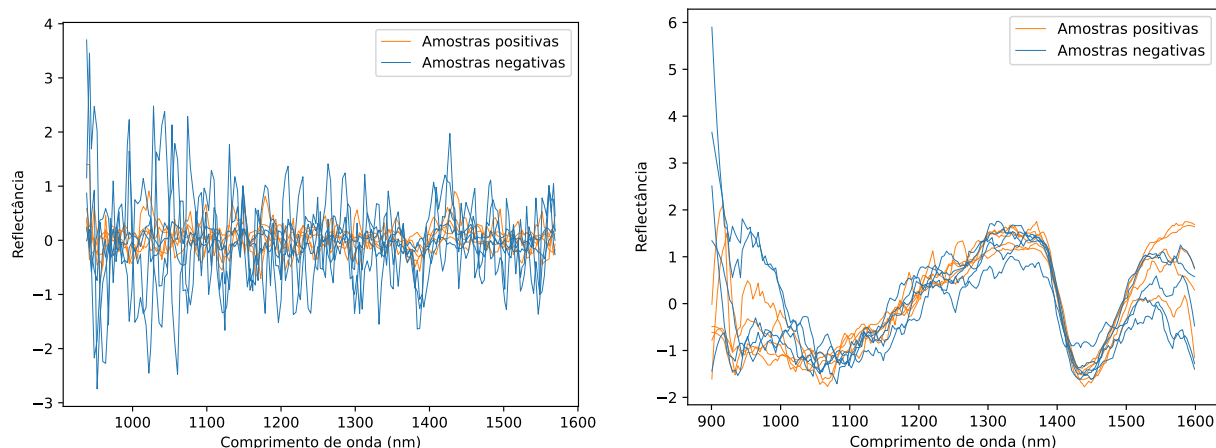


Figura 24 – Sinais coletados utilizando o Telspec Enterprise Sensor em conjunto com as janelas de cristal. Mais uma vez os sinais apresentam um aspecto ruidoso devido à baixa resolução do espectrômetro.

Comparando-se as Figuras 24 e 20, percebe-se uma grande diferença entre os sinais coletados, o que evidencia a influência do vidro dos tubos de ensaio na coleta. Também é possível notar que a ordem de grandeza da reflectância dos sinais coletados no Banco de Dados 3 é maior quando comparada à do Banco de Dados 2. A Figura 25 apresenta os sinais após as etapas de pré-processamento.

Da Figura 25a, percebe-se que a segunda derivada amplificou os ruídos dos sinais originais, e nenhum padrão específico pôde ser observado. Já na Figura 25b, a remoção da linha de base tornou evidente a presença duas zonas de alta de absorção,



(a) Segunda derivada do sinal obtida através da aplicação de um filtro Savitzky-Golay.

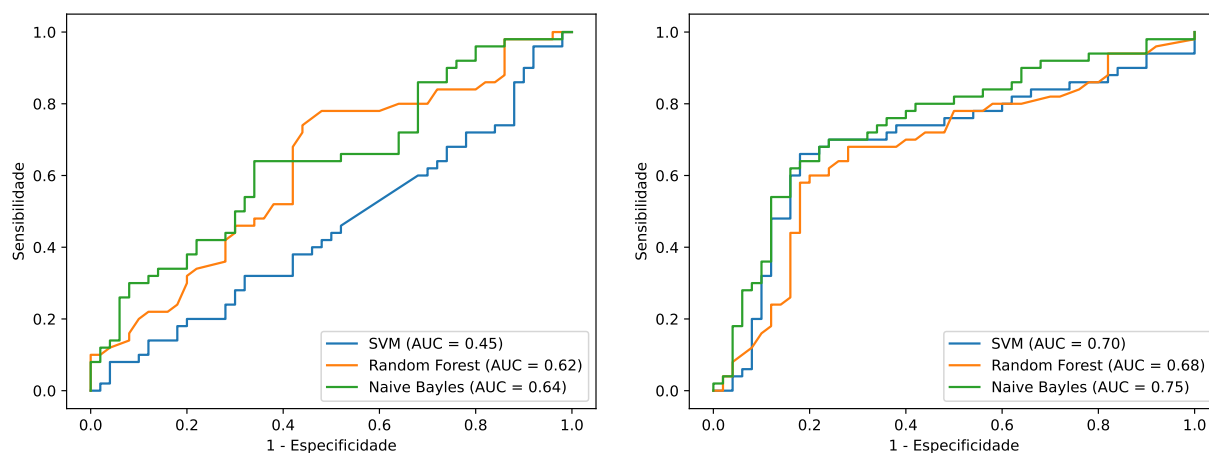
(b) Sinais após a remoção da linha de base por interpolação.

Figura 25 – Sinais após a etapa de pré-processamento. Todos os sinais foram normalizados para média e variância nulas.

a primeira entre os comprimentos de onda de 1000-1100 nm, e a segunda entre 1400-1600 nm.

Na etapa de classificação, os melhores resultados foram obtidos para o caso em que foram utilizados os comprimentos de onda que obtiveram os maiores valores de F extraídos do sinal cuja linha de base foi removida por interpolação. O classificador do tipo SVM foi o que alcançou as melhores métricas: $ACC = 0,72 \pm 0,05$, $SEN = 0,70 \pm 0,10$, $ESP = 0,74 \pm 0,13$ e $AUC = 0,70$. A Figura 26 apresenta as curvas ROC dos três classificadores utilizados nessa situação.

O desempenho dos modelos quando utilizada a energia média dos coeficientes da transformada *wavelet*, calculados no sinal sem a linha de base, que foi removida por interpolação, se mostrou razoável, obtendo uma acurácia de 71% para o classificador *Naive Bayes*. Já o uso dos *scores* das componentes principais obteve uma acurácia máxima de 65% com o modelo SVM. Além disso, o desempenho dos modelos quando utilizadas métricas extraídas da segunda derivada dos sinais foi inferior, chegando a uma acurácia máxima de 61% com o algoritmo *Naive Bayes*. Vale ainda ressaltar que o classificador *Random Forest* obteve as piores métricas na análise desse banco de dados.



(a) Segunda derivada do sinal obtida através da aplicação de um filtro Savitzky-Golay.

(b) Sinais após a remoção da linha de base por interpolação.

Figura 26 – Curva ROC para os 3 classificadores quando utilizados os comprimentos de onda selecionados com o teste F . É possível notar que as medidas extraídas da segunda derivada não apresentaram um bom desempenho. Já quando extraídas dos sinais cuja linha de base foi removida por interpolação, os classificadores tiveram uma boa performance, sendo o modelo do tipo *Naive Bayes* aquele com a maior área sobre a curva $AUC = 0,75$. Entretanto, a acurácia alcançada pelo SVM (72%) foi superior a dos algoritmos *Random Forest* (69%) e *Naive Bayes* (69%).

6 DISCUSSÃO

Este trabalho buscou utilizar os fundamentos da espectroscopia Vis-NIR e do aprendizado de máquina para distinguir amostras de *swab* nasofaríngeo com a presença do vírus SARS-CoV-2. Foram utilizadas amostras coletadas em um laboratório parceiro, que posteriormente foram submetidas à análise espectral através do uso de dois diferentes espectrômetros, o FieldSpec 3 e o TellSpec Enterprise Sensor.

Os resultados sugerem uma forte influência da solução de transporte na aquisição dos dados espectrais, em especial na faixa de 350-1000 nm, em que foi possível separar as amostras com diferentes soluções de transporte utilizando um algoritmo não supervisionado. Além disso, um novo grupo, formado pelas amostras coletadas nos dias 08/12/2021 e 10/12/2021 que utilizaram a solução salina como solução de transporte, pôde ser identificado. Ainda não se sabe o que ocasionou esse agrupamento, uma vez que o método de aquisição dos dados se manteve durante todo o período de coleta. Também foi nesse grupo que se obtiveram os melhores resultados, com uma acurácia em torno de 93% quando utilizada a faixa de 350-1000 nm.

Já para os outros dois grupos, os melhores resultados foram obtidos utilizando a faixa fora do visível. Para as amostras que utilizaram a solução salina, as melhores

métricas foram alcançadas usando a faixa de 1000-1500 nm do FieldSpec 3, com uma acurácia de 70%. Para o grupo formado pelas amostras que utilizaram o MTV Laborclin, o melhor desempenho se deu com o uso da faixa de 900-1600 nm do Telspec Enterprise Sensor, com uma acurácia de 74%. Esse espectrômetro também se mostrou mais eficiente na situação em que as amostras não foram separadas de acordo com a solução de transporte utilizada, obtendo uma acurácia de 64%.

A performance do TES para a base de dados em que se utilizaram as janelas de cristal apresentou bom desempenho, com uma acurácia em torno de 72%. Uma vez que esse desempenho é equiparável aos casos em que se utilizaram tubos de ensaio, esse resultado sugere que, apesar da sua influência na coleta dos dados espectrais, o vidro não compromete a performance do sistema. Logo, para esse tipo de estudo, o uso das janelas de cristal pode ser dispensado.

Na literatura, poucos trabalhos buscaram avaliar o uso da espectroscopia para o diagnóstico da Covid-19. No trabalho desenvolvido em (36), os autores utilizaram um espectrômetro do tipo ATR-FTIR (do inglês, *Attenuated total reflectance Fourier-transform infrared spectroscopy*) com uma faixa espectral de 2200 nm a 16600 nm para coletar os espectros de amostras de RNA extraídos de *swab* nasofaríngeo.

Foram utilizados 280 espécimens nesse estudo, sendo 100 testados positivo para Covid-19 e 180 testados negativo através do método RT-qPCR. Os autores utilizaram um filtro Savitzky-Golay para calcular a segunda derivada dos sinais e em seguida aplicaram um modelo de classificação esparsa para extrair características dos 280 sinais coletados. Foram ainda aplicados modelos PCA e PLS (do inglês, *Partial least squares*) para reduzir a dimensionalidade dessas características extraídas. Por fim, foram utilizados modelos de aprendizado supervisionado, como o SVM, para realizar a classificação. Os autores conseguiram alcançar uma acurácia próxima a 98% com a metodologia proposta.

Apesar dos bons resultados, o método proposto em (36) possui algumas desvantagens. Uma delas diz respeito ao custo da implementação, uma vez que um espectrômetro do tipo ATR-FTIR é consideravelmente mais caro do que os utilizados neste trabalho. Além disso, os autores coletam os dados espectrais em amostras de RNA, sendo a extração de RNA um processo custoso, demorado e que exige mão de obra especializada. Já o trabalho aqui proposto coleta os espectros diretamente das

amostras de *swab* nos tubos de ensaio ou em janelas de cristal.

Já no trabalho (37), os autores também utilizaram um espectrômetro do tipo ATR-FTIR com faixa espectral de 2200 nm a 16600 nm para obter o espectro de 243 amostras de *swab* nasofaríngeo, diagnosticados através do método RT-qPCR. Os espécimens nesse trabalho foram acondicionados em duas soluções de transporte distintas (chamadas pelos autores de líquido 1 e líquido 2), e por isso as análises foram feitas separadamente para cada líquido. Cinco microlitros de líquido foram colocados em um papel filme e secados ao ar livre por duas horas. Na sequência, o material foi colocado no espectrômetro e os dados espectrais foram coletados.

Na etapa de processamento e classificação, foi utilizado um filtro Savitzky-Golay para extrair a segunda derivada dos sinais e um modelo PLS para extração de características e redução da dimensionalidade. Foi então utilizado um classificador do tipo KNN (do inglês, *k nearest neighbors*) para realizar a identificação das amostras. Seguindo esse método, foi possível obter uma acurácia de 76,9% e 78,4% para os líquidos 1 e 2, respectivamente.

Os resultados encontrados em (37) são condizentes com os apresentados nesse trabalho, uma vez que as acurácias obtidas em ambos os trabalhos são próximas. Além disso, foi demonstrado que as soluções de transporte utilizadas interferem diretamente nos resultados dos classificadores. As desvantagens do trabalho proposto por (37) dizem respeito ao custo do equipamento, uma vez que se trata de um espectrômetro ATR-FTIR, e do processo de secagem das amostras, que leva um tempo considerável de duas horas.

Até onde vai o conhecimento do autor, não existem na literatura trabalhos que utilizem a faixa espectral Vis-NIR no diagnóstico da Covid-19, o que torna difícil uma comparação direta dos resultados com outros trabalhos. Entretanto, o método aqui proposto se apresenta como uma alternativa promissora e que corrobora com a literatura sobre o estudo da viabilidade da espectroscopia na identificação de amostras de *swab* infectadas pelo SARS-CoV-2.

7 CONCLUSÃO E TRABALHOS FUTUROS

Esse trabalho estudou a viabilidade do uso da espectroscopia Vis-NIR, em conjunto com técnicas de processamento de sinais e aprendizado de máquina, para a

identificação de amostras de *swab* nasofaríngeo infectadas pelo vírus SARS-CoV-2. Com o auxílio dos espectrômetros comerciais FieldSpec 3 e Tallspec Enterprise Sensor, foram coletados e organizados três bancos de dados espectrais obtidos de amostras de *swab* nasofaríngeo previamente analisadas para detecção do SARS-CoV-2 através do método RT-qPCR. Os sinais coletados foram então pré-processados utilizando técnicas de filtragem e remoção da linha de base. Na sequência, foram utilizadas métricas como a energia média dos coeficientes da transformada *Wavelet*, os *scores* das componentes principais e o valor de F do teste ANOVA, para extrair características que serviram de entrada para classificadores supervisionados. Os classificadores utilizados foram do tipo SVM, *Random Forest* e *Naive Bayes*

Para as amostras que utilizaram a solução salina como solução de transporte, o melhor resultado foi uma acurácia de 70%, alcançada com a faixa de 1000-1500 nm do espectrômetro FieldSpec 3. Já o espectrômetro Tallspec Enterprise Sensor, foi mais eficiente para identificar as amostras que utilizaram o MTV Laborclin, alcançando uma acurácia de 74%.

Através de um algoritmo de aprendizado não supervisionado, foi possível identificar um novo grupo de amostras, formado pelos exemplares coletados nos dias 08/12/2021 e 10/12/2021 que utilizaram a solução salina de NaCl como solução de transporte. Nesse grupo, o desempenho obtido se equipara a de testes comerciais, com uma acurácia de 93%. Além disso, foi investigada a necessidade do uso de janelas de cristal durante a etapa de aquisição dos espectros. Para o estudo proposto, o uso das janelas se mostrou dispensável, pois não teve um grande impacto no desempenho do sistema.

Com esses resultados, foi possível concluir que a espectroscopia Vis-NIR é uma técnica promissora para o diagnóstico do SARS-CoV-2. Também se pôde observar a influência da solução de transporte na aquisição dos dados espectrais e conseqüentemente o seu impacto na acurácia do sistema. Foi demonstrado que a configuração montada com o FieldSpec 3 é mais eficiente para amostras que utilizaram o NaCl como solução de transporte, enquanto a configuração com o Tallspec Enterprise Sensor é mais eficiente para as amostras que utilizaram o MTV Laborclin. Ademais, para esse estudo, o uso das janelas de cristal é dispensável.

Como sugestões para trabalhos futuros, é sugerida a ampliação do banco de

dados e padronização da solução de transporte utiliza, tanto para aqueles bancos em que se utilizaram os tubos de ensaio, como os que utilizaram as janelas de cristal. Dessa forma, será possível avaliar com uma maior confiabilidade os resultados encontrados. Também é sugerida a testagem dupla das amostras utilizadas, ou seja, submeter a mesma amostra duas vezes ao teste RT-qPCR e verificar se o resultado do teste se mantém o mesmo. Assim, os possíveis erros de diagnóstico relacionados com imprecisão intrínseca ao RT-qPCR serão reduzidos. Ademais, é sugerida a filtragem das amostras de *swab*. Essa filtragem irá eliminar dos espécimens elementos biológicos que não são de interesse, como coriza e sangue.

REFERÊNCIAS

- 1 SHI, F. et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for Covid-19. *IEEE reviews in biomedical engineering*, IEEE, 2020.
- 2 UDUGAMA, B. et al. Diagnosing COVID-19: the disease and tools for detection. *ACS nano*, ACS Publications, v. 14, n. 4, p. 3822–3835, 2020.
- 3 VASHIST, S. K. In vitro diagnostic assays for COVID-19: recent advances and emerging trends. *Diagnostics*, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 202, 2020.
- 4 LI, Y.; XIA, L. Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management. *American Journal of Roentgenology*, Am Roentgen Ray Soc, v. 214, n. 6, p. 1280–1286, 2020.
- 5 ANDREY, D. O. et al. Diagnostic accuracy of Augurix COVID-19 IgG serology rapid test. *European journal of clinical investigation*, Wiley Online Library, v. 50, n. 10, p. e13357, 2020.
- 6 SANTOS, M. C. et al. Spectroscopy with computational analysis in virological studies: A decade (2006–2016). *TrAC Trends in Analytical Chemistry*, Elsevier, v. 97, p. 244–256, 2017.
- 7 FERNANDES, J. N. et al. Rapid, noninvasive detection of Zika virus in *Aedes aegypti* mosquitoes by near-infrared spectroscopy. *Science advances*, American Association for the Advancement of Science, v. 4, n. 5, p. eaat0496, 2018.
- 8 SAKUDO, A.; BABA, K.; IKUTA, K. Discrimination of influenza virus-infected nasal fluids by Vis-NIR spectroscopy. *Clinica Chimica Acta*, Elsevier, v. 414, p. 130–134, 2012.
- 9 TONG, D. et al. Application of raman spectroscopy in the detection of hepatitis b virus infection. *Photodiagnosis and photodynamic therapy*, Elsevier, v. 28, p. 248–252, 2019.
- 10 MAIER, H. J.; BICKERTON, E.; BRITTON, P. Coronaviruses. *Methods and protocols*, Springer, 2015.
- 11 RABAAN, A. A. et al. SARS-CoV-2, SARS-CoV, and MERS-COV: a comparative overview. *Infez Med*, v. 28, n. 2, p. 174–184, 2020.
- 12 TESINI, B. Coronavírus e síndromes respiratórias agudas (Covid-19, Mers e Sars). *Manual MSD para profissionais da saúde*, 2020.
- 13 MIDDLE East respiratory syndrome coronavirus (MERS-CoV). [S.I.]: World Health Organization, 2021. Disponível em: <<https://www.who.int/emergencies/disease-outbreak-news/item/2021-DON317>>. Acesso em: 26 de maio 2021.

- 14 CHAFEKAR, A.; FIELDING, B. C. Mers-cov: Understanding the latest human coronavirus threat. *Viruses*, v. 10, n. 2, 2018. ISSN 1999-4915. Disponível em: <<https://www.mdpi.com/1999-4915/10/2/93>>.
- 15 WHO Coronavirus (COVID-19) Dashboard. [S.l.]: World Health Organization, 2021. Disponível em: <<https://covid19.who.int/>>. Acesso em: 26 de maio 2021.
- 16 CHENG, Z. J.; SHAN, J. 2019 novel coronavirus: where we are and what we know. *Infection*, Springer, v. 48, n. 2, p. 155–163, 2020.
- 17 IANNARELLA, R. et al. Coronavirus infections in children: from SARS and MERS to COVID-19, a narrative review of epidemiological and clinical features. *Acta Bio Medica: Atenei Parmensis*, Mattioli 1885, v. 91, n. 3, p. e2020032, 2020.
- 18 PETROSILLO, N. et al. COVID-19, SARS and MERS: are they closely related? *Clinical Microbiology and Infection*, Elsevier, v. 26, n. 6, p. 729–734, 2020.
- 19 WU, D. et al. The SARS-CoV-2 outbreak: what we know. *International Journal of Infectious Diseases*, Elsevier, v. 94, p. 44–48, 2020.
- 20 CASCELLA, M. et al. Features, evaluation, and treatment of coronavirus (COVID-19). *StatPearls*, 2021.
- 21 GRAYBEAL, J. D. et al. *Spectroscopy*. [S.l.]: Encyclopedia Britannica, inc., 2021. Disponível em: <<https://www.britannica.com/science/spectroscopy>>. Acesso em: 26 de maio 2021.
- 22 EINSTEIN, A. On a heuristic point of view toward the emission and transformation of light. *Ann. Phys*, v. 17, p. 132, 1905.
- 23 PLANCK, M. *The theory of heat radiation*. [S.l.]: Courier Corporation, 2013.
- 24 PAVIA, D. L. et al. Introdução à espectroscopia: Tradução da 4ª edição norte-americana. *São Paulo: Cengage Learning*, 2010.
- 25 COSTA, D. dos S. et al. Development of predictive models for quality and maturation stage attributes of wine grapes using Vis-NIR reflectance spectroscopy. *Postharvest biology and technology*, Elsevier, v. 150, p. 166–178, 2019.
- 26 STENBERG, B. et al. Visible and near infrared spectroscopy in soil science. *Advances in agronomy*, Elsevier, v. 107, p. 163–215, 2010.
- 27 SAVITZKY, A.; GOLAY, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, ACS Publications, v. 36, n. 8, p. 1627–1639, 1964.
- 28 LIEBER, C. A.; MAHADEVAN-JANSEN, A. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied spectroscopy*, Society for Applied Spectroscopy, v. 57, n. 11, p. 1363–1367, 2003.
- 29 DEVORE, J. L. *Probabilidade e estatística: para engenharia e ciências*. [S.l.]: Cengage Learning Edições Ltda., 2014.

- 30 SIFUZZAMAN, M.; ISLAM, M. R.; ALI, M. Application of wavelet transform and its advantages compared to Fourier transform. Vidyasagar University, Midnapore, West-Bengal, India, 2009.
- 31 KURITA, T. Principal component analysis (PCA). In: _____. *Computer Vision: A Reference Guide*. Cham: Springer International Publishing, 2019. p. 1–4. ISBN 978-3-030-03243-2. Disponível em: <https://doi.org/10.1007/978-3-030-03243-2_649-1>.
- 32 LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.
- 33 HARTIGAN, J. A.; WONG, M. A. et al. A k-means clustering algorithm. *Applied statistics*, USA, v. 28, n. 1, p. 100–108, 1979.
- 34 WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.
- 35 FAWCETT, T. An introduction to ROC analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.
- 36 KITANE, D. L. et al. A simple and fast spectroscopy-based technique for covid-19 diagnosis. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2021.
- 37 NOGUEIRA, M. S. et al. Rapid diagnosis of covid-19 using ft-ir atr spectroscopy and machine learning. *Scientific reports*, Nature Publishing Group, v. 11, n. 1, p. 1–13, 2021.