OCANSpectra: um sistema de detecção de câncer oral a partir da espectroscopia salivar ATR-FTIR

Anagê C. Mundim Filho[1], Janayna M. Fernandes[1], Robinson Sabino-Silva[2], Murillo G. Carneiro[1]

[1]PPG em Ciências da Computação, Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, MG, Brasil

[2]Departamento de Fisiologia, Instituto de Ciências Biomédicas, Universidade Federal de Uberlândia, Uberlândia, MG, Brasil

{anage.mundim, mgcarneiro}@ufu.br

{fernandesmjanayna, robinsonsabino}@gmail.com

Abstract. *Detecting oral cancer through immunohistochemical analysis is invasive, expensive, and often only detects cancer in later stages. Therefore, finding a non-invasive, sustainable, low-cost, and accurate diagnostic method is of great interest to the medical community. Attenuated total reflection infrared spectroscopy (ATR-FTIR) can provide valuable data for the detection of various diseases, including oral cancer. We investigate the use of ATR-FTIR data obtained from salivary samples to the detection of oral cancer. We evaluate five baseline correction methods and four classification techniques in order to improve respectively the spectrum quality and the predictive model. The combination of asymmetric least squares and support vector machine with gaussian kernel provided the best results with real data.*

Resumo. *Detectar câncer oral com análise imuno-histoquímica é invasivo, caro e só funciona em estágios avançados. Por isso, a comunidade médica busca um método diagnóstico não invasivo, preciso, sustentável e de baixo custo. A espectroscopia infravermelha de reflexão total atenuada (ATR-FTIR) pode ajudar a detectar várias doenças, incluindo câncer oral, usando amostras salivares. Nós testamos cinco métodos de correção de linha de base e cinco técnicas de classificação para melhorar a qualidade do espectro e o modelo preditivo. A combinação do método de mínimos quadrados assimétricos com máquina de vetores de suporte (kernel gaussiano) obteve os melhores resultados com os dados reais.*

1. Introdução

Câncer de boca (CB) é um problema de saúde global [Jubair et al. 2022], no qual 80% a 85% dos casos estão associados a fatores de estilo de vida [Kavarthapu and Gurumoorthy 2021] como por exemplo o uso de tabaco, consumo excessivo de bebidas alcoólicas, hábitos alimentares inadequados e má higiene bucal. Este tipo de câncer também está relacionado com a susceptibilidade genética, processos inflamatórios derivados de bactérias e vírus na cavidade oral, e pelo uso de próteses orais mal adaptadas [Rivera 2015].

A detecção precoce do CB é a forma mais eficiente de evitar casos de morte relacionados às metástases deste tipo de câncer ou mesmo uma cirurgia que impacta diretamente a qualidade de vida do paciente [Lobbezoo et al. 2022]. Todavia, a detecção precoce do CB requer um alto investimento em uma infraestrutura que promova a testagem frequente da população e, em caso de confirmação do diagnóstico, forneça o tratamento e demais recursos para atender a população necessitada [Neville and Day 2002].

Não obstante, estudos relacionados à espectroscopia de infravermelho de biofluídos têm chamado a atenção da comunidade científica [Ralbovsky and Lednev 2020]. Tais estudos apresentam resultados promissores para o diagnóstico de várias doenças e transtornos como COVID-19 [Zhang et al. 2021], Zika [Oliveira et al. 2023], HIV [Silva et al. 2020], câncer cerebral [Butler et al. 2019], câncer de pulmão [Peng et al. 2022] e câncer de mama[Sitnikova et al. 2020]. Uma parcela desses estudos conduz suas análises a partir de amostras de saliva, as quais possuem como vantagens coletas não invasivas bem como análises de baixo custo devido a ausência de reagentes. A espectroscopia ATR-FTIR (*Attenuated total reflectance by fourier transform infrared*) destaca-se por ser um método rápido e sustentável [Bunaciu et al. 2015]. A amostra de saliva além de fácil de ser coletada e armazenada é ideal para detecção de doenças nos seus estados iniciais [Malamud 2011]. Nesse cenário, ela pode ser de grande ajuda para a detecção precoce do CB. Contudo, duas limitações maiores precisam ser contornadas em estudos envolvendo ATR-FTIR: a presença de ruídos e outros compostos (por exemplo, água) que podem interferir na representação final do espectro infravermelho capturado das amostras [Baker et al. 2014], bem como a alta dimensionalidade dos espectros o que torna inviável em muitos casos o uso de ferramentas estatísticas convencionais [Morais et al. 2019]. Por essa razão, a preparação, pré-processamento e ajuste dos dados são etapas essenciais para sistemas baseados nesse tipo de análise.

O objetivo geral deste trabalho é desenvolver o OCANSpectra, um sistema escalável e facilmente implementável para diagnóstico de CB baseado em aprendizado de máquina capaz de discriminar pacientes com CB e indivíduos saudáveis a partir dos espectros de infravermelho com transformada de Fourier (FTIR) de amostras de saliva. A principal motivação para a pesquisa é a possibilidade de um diagnóstico precoce e rápido, que promova a detecção precoce da doença. Dessa forma, a hipótese investigada afirma que técnicas de classificação quando aplicadas para análise de espectros de saliva podem prover diagnóstico acurado de CB.

Entre as contribuições do sistema, destacam-se a avaliação de diferentes métodos de preparação dos dados (também conhecida por correção de baseline), bem como de diferentes técnicas de classificação para o diagnóstico de CB. Ademais, diferente de outros estudos da literatura, OCANSpectra é capaz de analisar e selecionar automaticamente a melhor combinação entre os métodos de correção de baseline e as técnicas de classificação. Outra contribuição importante diz respeito ao estudo de caso adotado para validação do estudo, o qual considerou uma base de dados reais coletadas em um hospital público. A seguir apresenta-se uma breve descrição sobre tais contribuições:

- O sistema apresentado neste artigo permite a análise dos espectros tanto em sua forma bruta (sem tratamento - Raw) quanto a partir de cinco métodos de correção de baseline: polinomial, *asymmetric least squares* (als), *asymmetrically reweighted penalized least squares* (arPLS), *doubly reweighted partial least squares* (dr-

PLS) e rubberband.
- OCANSpectra permite a análise de padrões espectrais correspondentes aos grupos de câncer de boca e controle a partir de cinco modelos de aprendizado de máquina: análise discriminante linear (LDA) [Balakrishnama and Ganapathiraju 1998]; SVM linear (SVM) [Noble 2006]; SVM com kernel função de base radial (RBF) [Janik and Lobos 2006]; floresta aleatória (RF) [Liaw et al. 2002], e redes neurais convolucionais [Krizhevsky et al. 2012].
- Validação do sistema considerando seu pipeline completo, desde a coleta dos dados, preparação e treinamento dos modelos até o diagnóstico a partir de novas amostras de saliva. Destaque aqui para os bons resultados obtidos pelo OCANSpectra, o qual além de possibilitar avaliação de diferentes combinações de técnicas de pré-processamento e aprendizado, também possui mecanismos para visualização e análise dos resultados.

É importante destacar o impacto positivo de OCANSpectra tanto no campo computacional quanto na vida das pessoas, auxiliando no diagnóstico regular, não invasivo e financeiramente acessível para a população, além de abrir o caminho para o desenvolvimento de sistemas para diagnóstico de outras doenças com essas mesmas vantagens. O sistema abre novas oportunidades para o mercado e traz a possibilidade de descentralizar a realização análises e diagnósticos, logo qualquer unidade de saúde, por menor que seja, poderia realizar a coleta e o diagnóstico localmente desde que possuam acesso à Internet e o equipamento de espectroscopia ATR-FTIR, o qual tem um custo muito baixo comparado a outros equipamentos e procedimentos adotados para diagnóstico de câncer de boca. Os resultados rápidos e precisos alcançados por OCANSpectra, podem tornar o sistema adequado para uso em triagem e exames de rotina, por exemplo, contribuindo na promoção e realização de testes regularmente na população. Ademais, diferente de outros estudos da literatura, OCANSpectra é capaz de analisar e selecionar automaticamente a melhor combinação entre os métodos de correção de baseline e aprendizado de máquina.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta uma revisão relacionada ao uso do espectrograma ATR-FTIR para amostras de CB utilizando amostras de saliva. A Seção 3 apresenta o método de pesquisa adotado neste artigo, além de descrever a base de dados utilizada e os métodos de correção de baseline e classificação de dados investigados. A Seção 4 apresenta o ambiente experimental e os resultados das simulações, além de discutir análises e achados da pesquisa. Finalmente, a Seção 5 relata as contribuições do artigo e as próximas etapas da pesquisa.

2. Estudos Relacionados

A utilização da espectroscopia ATR-FTIR com amostras biológicas tem sido utilizada na triagem diagnostica de doenças [Sala et al. 2020]. Basicamente, a análise de absorbância de várias bandas de infravermelho permite a identificação de componentes e estruturas moleculares [Giamougiannis et al. 2021]. Em comparação com outras plataformas de diagnóstico, as principais vantagens das plataformas ATR-FTIR que utilizam amostras de saliva estão relacionadas à coleta não invasiva, análise sustentável, baixo custo e resultados rápidos. ATR-FTIR tem sido empregado para o diagnóstico de diversas doenças como diabetes, cânceres e também COVID-19 [Caixeta et al. 2023, Bozkurt et al. 2010, Sala et al. 2020, Martinez-Cuazitl et al. 2021]. No entanto, para aplicação da espectroscopia ATR-FTIR clínica faz-se necessária uma validação em estudos de larga-escala e

uma evolução em algoritmos que permitam reduzir diferenças da aquisição do espectro infravermelho por diferentes operadores e diferentes aparelhos. Neste sentido, o desenvolvimento de algoritmos que permitam a correção de baseline do espectro é de fundamental relevância [Morais et al. 2019].

A correção de baseline é importante para remover interferências de fatores externos à coleta de amostras, como componentes do ar atmosférico na aquisição de dados pelo ATR-FTIR. A aplicação deste método permite uma otimização da análise mantendo o mesmo perfil espectral e deve ser utilizada para todas as amostras. A aplicação da correção de baseline no espectro ocorre normalmente após etapas de truncamento da faixa espectral analisada e da eliminação ou suavização de ruídos e previamente à aplicação de ferramentas de normalização de dados. Entre os métodos baseados em teorias da física para correção de baseline estão automatic weighted least squares e rubberband [Morais et al. 2019].

Utilizando as bases de dados MEDLINE, Scopus, Google Scholar e SciELO com as palavras-chave "oral cancer detection" ou "oral cancer diagnosis" ou "oral cancer" e "FTIR" ou "ATR-FTIR" e "saliva" foram localizados 3 artigos científicos de relevância nesta área. [Mikkonen et al. 2016] detectaram a presença de ânions tiocianato (OSCN-) no modo vibracional 2050 cm^{-1}, o que foi indiretamente associado ao câncer de boca embora a saliva tenha sido coletada apenas em indivíduos controles saudáveis. No trabalho a correção do baseline foi realizada de forma polinomial ancorada em 4 modos vibracionais. A utilização do espectro infravermelho adquirido por espectroscopia ATR-FTIR foi capaz de discriminar com 82% de acurácia amostras de saliva de pacientes com câncer de boca e de orofaringe (n=19) em comparação com sujeitos controles (n=13). No estudo apresentado em [Falamas et al. 2021], dados ATR-FTIR em conjunto com análise de componente principal (PCA) são usados para detecção de CB; a correção de baseline foi aplicada por ajuste e subtração baseado em polinômio de quinta ordem do conjunto de dados, seguido pela normalização dos espectros para o espectro bruto de cada espectro e alcançaram 82% de acurácia. Por fim, o estudo de [Shaikh et al. 2021], em que a saliva foi coletada de sujeitos com fibrose submucosa oral (n=15), um tipo de condição pré-cancerígena de câncer de boca com alto potencial de malignidade, apresentou a acurácia baseada na área sobre a curva de 69% em comparação com sujeitos controles. Como correção de baseline foi apresentado apenas o método rubberband.

Apesar dos resultados promissores dos trabalhos relacionados, eles apresentam apenas um método de correção de baseline e um método para análise ou aprendizado dos dados. Ademais, na própria literatura ATR-FTIR, o processo de análise e seleção automática desses métodos têm sido pouco explorado [Morais et al. 2019]. Neste sentido, OCANSpectra contribui para a literatura por prover esse tipo de recurso na direção de modelos de aprendizado de máquina automático.

3. Materiais e Métodos

Esta seção descreve as principais etapas funcionais do sistema desenvolvido para detecção de câncer de boca a partir da análise de amostras de saliva em equipamento de espectroscopia infravermelho ATR-FTIR, a saber: base de dados, preparação e pré-processa-mento dos dados, aprendizado de máquina e avaliação de resultados.

A Figura 1 apresenta uma visão geral sobre o sistema desenvolvido. Pela figura, é

possível perceber que OCANSpectra é composto de três componentes principais: coleta, armazenamento e processamento de dados. O componente de coleta de dados refere-se à coleta de amostras de saliva dos pacientes e ao processamento dessas amostras pelo equipamento ATR-FTIR, resultando em uma representação espectral do composto biológico. O componente de armazenamento de dados refere-se tanto ao armazenamento dos dados coletados a partir do componente anterior, quanto do pipeline de processamento com melhor desempenho preditivo (chamado pipeline referência), os quais serão utilizados no diagnóstico de amostras cujo grupo (câncer ou controle) é desconhecido. O componente de processamento de dados é responsável pelo ajuste de pipelines de classificação a partir da base de treinamento e considerando diferentes métodos de correção de baseline e de aprendizado de máquina, sendo que o pipeline de referência é armazenado no componente anterior.

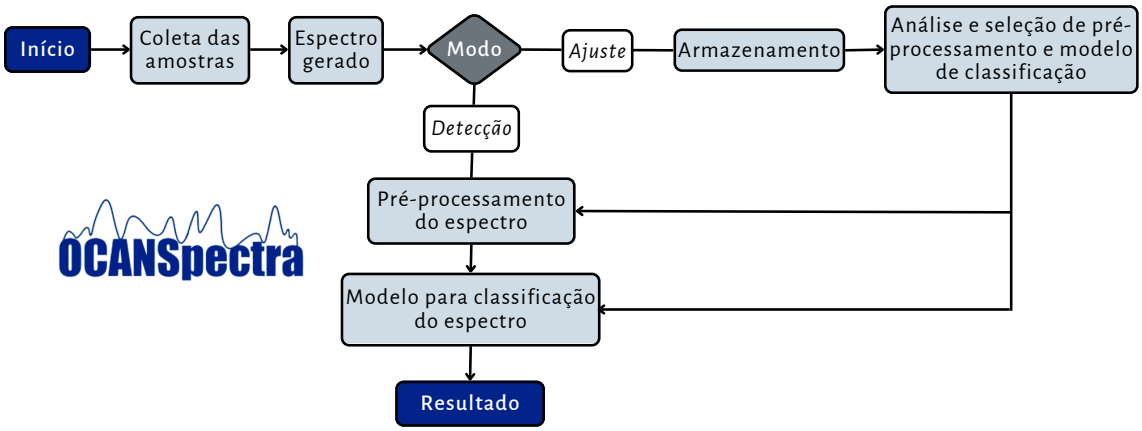


Figura 1. Funcionamento do OCANSpectra

A maior vantagem da arquitetura apresentada na Figura 1 é que uma vez ajustado o pipeline de referência e disponibilizado através de um servidor Web, o componente de coleta de dados pode ser realizado localmente, por qualquer unidade de saúde que dispõe de equipamento ATR-FTIR, enquanto que o envio dos espectros e o recebimento dos resultados da análise pelo pipeline de referência podem ser realizados pela própria Internet. Nas subseções a seguir são detalhados a base de dados adotada no estudo de caso, bem como o componente de processamento dos dados pelo sistema, além de questões relacionadas à avaliação de resultados.

3.1. Base de dados

A base de dados foi obtida a partir de amostras de saliva de um total de 65 pacientes, dos quais 39 tiveram diagnóstico positivo para câncer de boca (grupo alvo) e 26 tiveram diagnóstico negativo (grupo controle). As amostras foram processadas por um espectrômetro ATR-FTIR, responsável por medir variações vibracionais das moléculas de saliva para diferentes bandas de infravermelho entre 4000cm^{-1} e 400cm^{-1}. Para cada amostra, o equipamento é responsável por gerar um espectro correspondente.

A base de dados foi obtida com aprovação do Comitê de Ética em Pesquisas da Universidade Federal de Uberlândia, sob protocolo 249.200.9. Para diagnosticar os pacientes do grupo câncer de boca foi usada a classificação TNM de tumores malignos da União Internacional para Controle do Câncer. Por outro lado, pacientes do grupo controle

foram selecionados de modo a terem idade e sexo equiparados ao grupo alvo, além de apresentarem histórico negativo para outros tipos de câncer.

3.2. Preparação dos espectros

Nesta seção serão apresentadas as três principais etapas para tratamento dos espectros ATR-FTIR, a saber: correção de baseline, truncamento do espectro e normalização do espectro pela região de amida I.

3.2.1. Correção de baseline

Correção de baseline, também denominada subtração de baseline, é uma importante etapa de preparação de amostras em equipamentos suscetíveis à interferência de materiais e do próprio espectrômetro. O propósito desta etapa é o tratamento de artefatos indesejáveis, incluindo deslocamento e inclinação de baseline, os quais causam desvios de intensidade que podem atrapalhar o aprendizado e detecção de padrões nas etapas de análise seguintes [Peng et al. 2011]. Neste estudo, os seguintes métodos de correção de baseline foram investigados:

- Polinomial (poly): Ajuste polinomial para correção de baseline [Losq 2018].
- *Asymmetric least squares* (als): Ajuste de mínimos quadrados para correção de baseline [Boelens et al. 2005];
- *Asymmetrically reweighted penalized least squares* (arPLS): Ajuste de mínimos quadrados ponderado assimétricamente para correção de baseline [Baek et al. 2015].
- *Doubly reweighted partial least squares* (drPLS): Correção de baseline baseada em ajuste de mínimos quadrados duplamente reponderados [Xu et al. 2019].
- Rubberband: Correção de baseline baseada na interpolação linear dos pontos que formam um fecho convexo sobre cada espectro [Losq 2018].

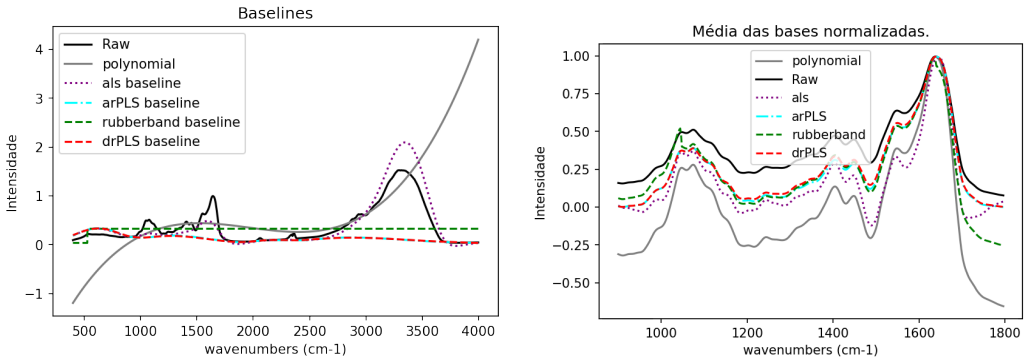
A Figura 2(a) apresenta as baselines obtidas pelos métodos de correção de baseline para o espectro médio da base de dados, e a Figura 2(b) apresenta o espectro médio após o pré-processamento completo das amostras da base de dados, o qual além da correção de baseline envolve o truncamento e a normalização descritos a seguir.

3.2.2. Truncamento do espectro

O espectro correspondente a cada amostra é truncado na região entre 1800cm^{-1} e 900cm^{-1}, sendo que as demais regiões são desconsideradas. O objetivo principal é evitar regiões de muito ruído que podem dificultar uma boa capacidade de generalização por parte dos modelos aprendidos.

3.2.3. Normalização pela Amida I

A última etapa de pré-processa-mento é a normalização pela amida I, uma região com elevado potencial de caracterização de proteínas. Nesse método, os valores de cada espectro são normalizados pelo seu pico de amida I (maior valor na região entre 1630cm^{-1} e 1660cm^{-1}).



(a) Baselines

(b) Média dos espectros após pré-processamento.

Figura 2. Métodos de correção de baseline aplicados nos espectros.

3.3. Aprendizado dos padrões espectrais

A seguir são introduzidos os cinco métodos de classificação de dados considerados neste estudo a partir de trabalhos importantes da literatura [Morais et al. 2019, Carneiro 2016], sendo dois deles lineares e três não lineares:

- Análise discriminante linear (LDA): É uma técnica de classificação com um limite de decisão linear, ajustando densidades condicionais de classe aos dados e usando a regra de Bayes. O modelo ajusta uma densidade gaussiana para cada classe, assumindo que todas as classes compartilham a mesma matriz de covariância.
- SVM linear (SVM-linear): Encontra um hiperplano com fronteira máxima de separação projetando uma linha nos objetos da borda. Para amostras não-linearmente separáveis, os dados de entrada são transformados em um vetor de características multidimensional, para então ser possível separar os atributos linearmente no espaço, esta transformação é realizada através de uma representação de kernel.
- SVM com função de base radial (SVM-RBF): É otimizado de acordo com o parâmetro gamma em conjunto com o parâmetro C. Gamma é um parâmetro de função do kernel que lida com o padrão não linear, pode ser visto como o inverso do raio de influência das amostras selecionadas pelo modelo como vetores de suporte e quanto menor menor o viés. C é o parâmetro de penalidade que manipula o hiperplano, quanto maior menor o viés, enquanto o ajuste excessivo pode ocorrer.
- Random Forest (RF): É uma técnica que ajusta vários classificadores de árvore de decisão em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo (overfitting).
- Rede neural convolucional de 1 dimensão (1D CNN): A 1D CNN é um tipo de arquitetura de rede neural projetada especificamente para trabalhar com dados unidimensionais. Uma Rede Neural Convolucional 1 dimensão é um algoritmo de Aprendizado Profundo que pode processar sequências de dados unidimensionais, como séries temporais ou sinais de áudio. Assim como em uma CNN convencional, a CNN 1D atribui importância aos diferentes aspectos ou padrões presentes nos dados de entrada, utilizando pesos e vieses que podem ser aprendidos durante o treinamento.

3.4. Avaliação dos resultados

Como medidas de avaliação dos resultados foram consideradas a acurácia (quantidade de acerto do algoritmo correspondente às duas classes, positiva e controle), a sensibilidade (corresponde ao percentual de resultados positivos dentre as pessoas que realmente foram diagnosticadas com a doença), a especificidade (percentual de pessoas que testaram negativo que não apresentam diagnóstico da doença), e a média entre sensibilidade e especificidade. As medidas de sensibilidade e especificidade são bastante utilizadas no contexto de diagnóstico de doenças por conta de informar o desempenho do classificador para cada grupo, evitando erros comuns associados à acurácia.

4. Resultados experimentais

Para a realização das simulações com a base de dados foi utilizada a validação cruzada estratificada com 10 subconjuntos. Nesse método, nove subconjuntos são usados para treinamento do modelo e um subconjunto para teste, de modo que a cada iteração um subconjunto diferente seja testado, resultando em um total de 10 simulações. Esse procedimento é repetido três vezes, alterando a configuração dos subconjuntos em cada uma delas, de modo a ter uma estimativa média mais aproximada do erro real [Berrar 2019], totalizando 30 simulações. Tais simulações foram projetadas para permitir uma estimativa mais próxima do erro real do sistema, o qual apesar de ainda contar com uma base de tamanho razoável, foi avaliado de forma rigorosa. Ademais, os experimentos buscam investigar a relevância da análise e seleção de vários métodos de preparação dos dados e de classificação para o desempenho preditivo na detecção de câncer oral.

Além da análise do espectro bruto (Raw), foram analisados os seguintes métodos de correção de baseline: poly, als, arPLS, drPLS e rubberband (vide Subseção 3.2). Para a discriminação dos espectros do grupo controle e alvo, foram considerados duas técnicas lineares, LDA e SVM-linear, e três técnicas não lineares: SVM-RBF, RF e redes neurais convolucionais (vide Subseção 3.3). Os parâmetros dessas técnicas foram ajustados de modo que o número de árvores em RF foi definido a partir de $\{2^1, 2^2, \ldots, 2^{10}\}$, e a função de custo em SVM-linear e SVM-RBF a partir de $\{2^0, 2^2, \ldots, 2^{14}\}$. Em relação aos parâmetros da CNN, os filtros foram definidos para $\{32, 64, 128\}$, tamanhos de kernel $\{3, 5\}$, neurônios na camada totalmente conectada $\{32, 64, 128\}$ e épocas $\{20, 100, 200\}$.

A Tabela 1 apresenta o resultado obtido pelos algoritmos de classificação considerando os diferentes métodos de correção de baseline. Para cada algoritmo, é possível perceber melhorias significativas nos resultados a depender do método de correção considerado. Por exemplo, SVM-RBF alcançou o melhor desempenho preditivo dentre todos os algoritmos com als. A tabela permite ressaltar a relevância da correção de baseline, uma vez que a ausência desse tipo de tratamento (Raw) implicou geralmente nos piores resultados para cada algoritmo. Por outro lado, merece destaque os resultados obtidos pelo método arPLS quando considerando os dois modelos de SVM bem como RF, os quais estão entre os melhores resultados obtidos por tais algoritmos. Em relação aos classificadores, os melhores resultados foram obtidos pelos modelos SVM não linear, especialmente o modelo não linear.

Uma análise complementar é conduzida através da Figura 3 (a), na qual é possível notar que as baselines geradas pelos métodos polinomial e als são de maior variação do

Algoritmo	Pré-processamento	Acurácia	Sensibilidade	Especificidade	Média(S,E)
LDA	Raw	0,56 ± 0,044	0,57 ± 0,037	0,54 ± 0,067	0,55
	Poly	0,53 ± 0,06	0,59 ± 0,038	0,43 ± 0,011	0,51
	als	0,51 ± 0,015	0,52 ± 0,014	0,50 ± 0,059	0,51
	arPLS	0,59 ± 0,036	0,59 ± 0,060	0,59 ± 0,015	0,59
	drPLS	0,64 ± 0,044	0,66 ± 0,047	0,63 ± 0,059	0,64
	rubberband	0,60 ± 0,022	0,58 ± 0,030	0,63 ± 0,034	0,61
SVM- Linear	Raw	0,54 ± 0,022	0,46 ± 0,017	0,68 ± 0,067	0,57
	Poly	0,66 ± 0,024	0,69 ± 0,010	0,63 ± 0,054	0,66
	als	0,67 ± 0,028	0,66 ± 0,010	0,70 ± 0,027	0,68
	arPLS	0,59 ± 0,037	0,66 ± 0,046	0,50 ± 0,036	0,58
	drPLS	0,61 ± 0,034	0,64 ± 0,044	0,56 ± 0,034	0,60
	rubberband	0,64 ± 0,016	0,68 ± 0,007	0,60 ± 0,023	0,64
SVM- RBF	Raw	0,60 ± 0,044	0,79 ± 0,071	0,29 ± 0,087	0,54
	Poly	0,62 ± 0,016	0,66 ± 0,010	0,58 ± 0,043	0,62
	als	**0,71 ± 0,030**	**0,69 ± 0,020**	**0,74 ± 0,061**	**0,72**
	arPLS	0,64 ± 0,004	0,64 ± 0,027	0,65± 0,061	0,65
	drPLS	0,65 ± 0,042	0,80 ± 0,045	0,43 ± 0,049	0,58
	rubberband	0,61 ± 0,026	0,69 ± 0,044	0,66 ± 0,040	0,68
RF	Raw	0,51 ± 0,044	0,52 ± 0,071	0,52 ± 0,049	0,52
	Poly	0,59 ± 0,039	0,71 ± 0,063	0,41 ± 0,023	0,53
	als	0,65 ± 0,005	0,76 ± 0,063	0,48 ± 0,023	0,52
	arPLS	0,56 ± 0,074	0,57 ± 0,125	0,56 ± 0,051	0,56
	drPLS	0,63 ± 0,062	0,74 ± 0,088	0,49 ± 0,020	0,62
	rubberband	0,55 ± 0,018	0,71 ± 0,023	0,30 ± 0,040	0,51
CNN	Raw	0,55 ± 0,023	0,61 ± 0,030	0,32 ± 0,072	0,47
	Poly	0,56 ± 0,033	0,67 ± 0,051	0,42 ± 0,072	0,55
	als	0,57 ± 0,023	0,62 ± 0,041	0,51 ± 0,007	0,57
	arPLS	0,57 ± 0,007	0,70 ± 0,031	0,36 ± 0,028	0,59
	drPLS	0,61 ± 0,019	0,76 ± 0,042	0,41 ± 0,054	0,59
	rubberband	0,52 ± 0,017	0,71 ± 0,054	0,24 ± 0,051	0,48

Tabela 1. Comparação dos resultados em termos de sensibilidade, especificidade, média(S,E) e acurácia para as cinco técnicas de classificação sob os diferentes métodos de correção de baseline.

que aquelas geradas pelos demais métodos de correção. A consequência desse comportamento pode ser vista na própria figura. Na Figura 3 (b) podemos observar a média dos espectros obtidos para cada método de correção, em que os dois métodos destoam dos demais, especialmente o método polinomial no começo do espectro. Essa variação em função da base de dados original parece dificultar o processo de treinamento dos modelos SVM, sendo que tal método foi inclusive superado pelo Raw (sem correção de baseline). O principal uso destes métodos de correção são a normalização do espectro para evitar ruídos e colaborar com os algoritmos de aprendizado de máquina para ter um melhor treinamento e conseguir diferenciar espectros de cada grupo.

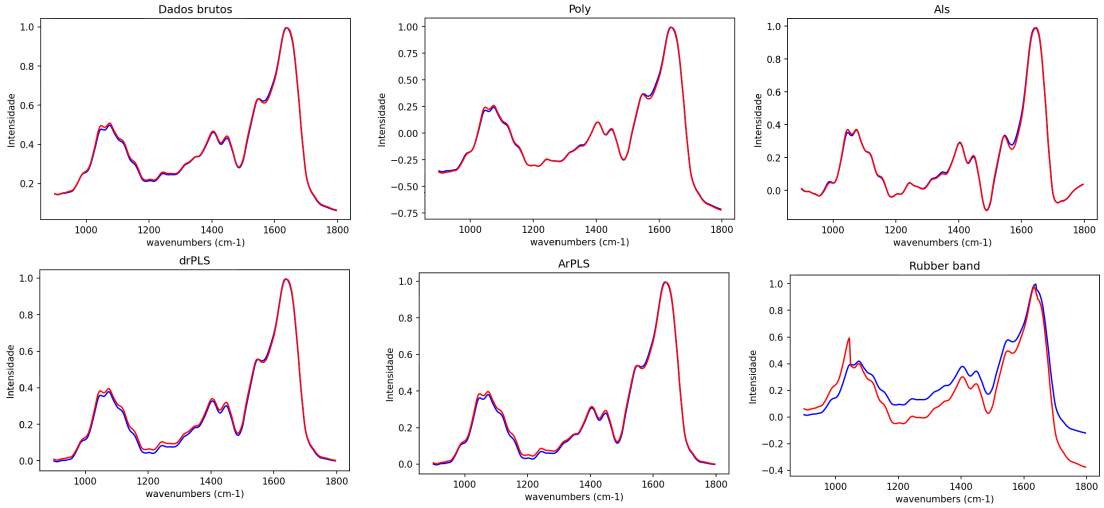


Figura 3. Média dos espectros do grupo alvo e controle após o pré-processamento das baselines

Finalmente, a Figura 3 apresenta como ficaram a média dos espectros do grupo controle e BD após a aplicação dos métodos de correção de baseline, truncamento e normalização pela amida I. Pela figura, é possível ver que o rubberband obteve maior diferenciação entre as médias, seguido pelos métodos arPLS e drPLS. Se por um lado esse tipo de análise média pode ser suscetível a variações e ruídos, por outro lado ele fornece evidências que podem ajudar a explicar o melhor desempenho preditivo de alguns métodos de correção de baseline e classificação de dados em relação a outros.

5. Conclusão

Este artigo aborda o tema de um novo método de diagnóstico para câncer de boca e outras doenças utilizando a espectroscopia ATR-FTIR. Para isso foi desenvolvido o sistema OCANSpectra para facilitar a análise de amostras coletadas e também fornecer resultados que possam servir de suporte para decisões médicas. Também, foi investigado métodos de correção de baseline e aprendizado de máquina para o diagnóstico de câncer de boca a partir de espectroscopia ATR-FTIR de amostras de saliva. Basicamente, cinco métodos de correção de baseline e cinco técnicas de classificação de dados foram consideradas para o problema em questão. Os resultados experimentais demonstraram que os modelos SVM, tanto linear quanto não linear, obtiveram os melhores desempenhos preditivos, especialmente através dos métodos de correção als e rubberband. Estas técnicas são importantes para a preparação do espectro, deixando-o com menos ruídos e influências exter-

nas do ambiente de coleta. Especificamente, a combinação de als e SVM-RBF alcançou sensibilidade de 0,69 e especificidade de 0,74. Esse é um resultado atrativo, uma vez que a base de dados ainda possui tamanho limitado (65 pacientes) e que outros métodos de pré-processamento e classificação de dados ainda serão investigados. Para trabalhos futuros, gostaríamos de expandir o sistema, pois um fator importante é a inclusão de métodos *explainable AI* como o SHapley Additive exPlanations (SHAP) para auxiliar na interpretação e análise dos resultados, deixando mais acessível o sistema de análise para ser utilizado por equipes médicas. Além disso, será possível expandir o banco de dados, assim conseguindo melhorar o treinamento dos dados e experimentar novos algoritmos para aprimorar o processamento do diagnóstico e o sistema como um todo.

Agradecimentos

Os autores agradecem o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG (processo APQ-00410-21), do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES. Os autores também agradecem o apoio da NVIDIA Corporation através da GPU Titan V usada nesta pesquisa.

Referências

Baek, S.-J., Park, A., Ahn, Y.-J., and Choo, J. (2015). Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, 140(1):250–257.

Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., et al. (2014). Using fourier transform ir spectroscopy to analyze biological materials. *Nature protocols*, 9(8):1771–1791.

Balakrishnama, S. and Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8.

Berrar, D. (2019). Cross-validation.

Boelens, H. F., Eilers, P. H., and Hankemeier, T. (2005). Sign constraints improve the detection of differences between complex spectral data sets: Lc- ir as an example. *Analytical chemistry*, 77(24):7998–8007.

Bozkurt, O., Severcan, M., and Severcan, F. (2010). Diabetes induces compositional, structural and functional alterations on rat skeletal soleus muscle revealed by ftir spectroscopy: a comparative study with edl muscle. *Analyst*, 135(12):3110–3119.

Bunaciu, A. A., Hoang, V. D., and Aboul-Enein, H. Y. (2015). Applications of ft-ir spectrophotometry in cancer diagnostics. *Critical reviews in analytical chemistry*, 45(2):156–165.

Butler, H. J., Brennan, P. M., Cameron, J. M., Finlayson, D., Hegarty, M. G., Jenkinson, M. D., Palmer, D. S., Smith, B. R., and Baker, M. J. (2019). Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nature communications*, 10(1):1–9.

Caixeta, D. C., Carneiro, M. G., Rodrigues, R., Alves, D. C. T., Goulart, L. R., Cunha, T. M., Espindola, F. S., Vitorino, R., and Sabino-Silva, R. (2023). Salivary ATR-FTIR

spectroscopy coupled with support vector machine classification for screening of type 2 diabetes mellitus. *Diagnostics*, 13(8):1396.

Carneiro, M. G. (2016). *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. PhD thesis, Universidade de São Paulo.

Falamas, A., Faur, C., Ciupe, S., Chirila, M., Rotaru, H., Hedesiu, M., and Pinzaru, S. C. (2021). Rapid and noninvasive diagnosis of oral and oropharyngeal cancer based on micro-raman and ft-ir spectra of saliva. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 252:119477.

Giamougiannis, P., Morais, C. L., Rodriguez, B., Wood, N. J., Martin-Hirsch, P. L., and Martin, F. L. (2021). Detection of ovarian cancer (±neo-adjuvant chemotherapy effects) via ATR-FTIR spectroscopy: comparative analysis of blood and urine biofluids in a large patient cohort. *Analytical and bioanalytical chemistry*, 413(20):5095–5107.

Janik, P. and Lobos, T. (2006). Automated classification of power-quality disturbances using svm and rbf networks. *IEEE Transactions on Power Delivery*, 21(3):1663–1669.

Jubair, F., Al-karadsheh, O., Malamos, D., Al Mahdi, S., Saad, Y., and Hassona, Y. (2022). A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Diseases*, 28(4):1123–1130.

Kavarthapu, A. and Gurumoorthy, K. (2021). Linking chronic periodontitis and oral cancer: A review. *Oral Oncology*, 121:105375.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lobbezoo, F., Aarab, G., Verhoeff, M. C., and Volgenant, C. M. (2022). The value of oral care in dying and death. *The Lancet*, 399(10342):2187–2188.

Losq, C. L. (2018). Rampy: a Python library for processing spectroscopic (IR, Raman, XAS...) data.

Malamud, D. (2011). Saliva as a diagnostic fluid. *Dental Clinics*, 55(1):159–178.

Martinez-Cuazitl, A., Vazquez-Zapien, G. J., Sanchez-Brito, M., Limon-Pacheco, J. H., Guerrero-Ruiz, M., Garibay-Gonzalez, F., Delgado-Macuil, R. J., de Jesus, M. G. G., Corona-Perezgrovas, M. A., Pereyra-Talamantes, A., et al. (2021). ATR-FTIR spectrum analysis of saliva samples from covid-19 positive patients. *Scientific Reports*, 11(1):1–14.

Mikkonen, J. J., Raittila, J., Rieppo, L., Lappalainen, R., Kullaa, A. M., and Myllymaa, S. (2016). Fourier transform infrared spectroscopy and photoacoustic spectroscopy for saliva analysis. *Applied Spectroscopy*, 70(9):1502–1510.

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

Neville, B. W. and Day, T. A. (2002). Oral cancer and precancerous lesions. *CA: a cancer journal for clinicians*, 52(4):195–215.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

Oliveira, S. W., Cardoso-Sousa, L., Georjutti, R. P., Shimizu, J. F., Silva, S., Caixeta, D. C., Guevara-Vega, M., Cunha, T. M., Carneiro, M. G., Goulart, L. R., et al. (2023). Salivary detection of zika virus infection using ATR-FTIR spectroscopy coupled with machine learning algorithms and univariate analysis: A proof-of-concept animal study. *Diagnostics*, 13(8):1443.

Peng, J., Peng, S., Xie, Q., and Wei, J. (2011). Baseline correction combined partial least squares algorithm and its application in on-line fourier transform infrared quantitative analysis. *Analytica chimica acta*, 690(2):162–168.

Peng, X., Dai, R., Ma, Y., Lin, B., Hui, X., Chen, X., and Lv, R. (2022). Early diagnosis and bioimaging of lung adenocarcinoma cells/organs based on spectroscopy machine learning. *Journal of Innovative Optical Health Sciences*, 15(02):2250011.

Ralbovsky, N. M. and Lednev, I. K. (2020). Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chemical Society Reviews*, 49(20):7428–7453.

Rivera, C. (2015). Essentials of oral cancer. *International journal of clinical and experimental pathology*, 8(9):11884.

Sala, A., Anderson, D. J., Brennan, P. M., Butler, H. J., Cameron, J. M., Jenkinson, M. D., Rinaldi, C., Theakstone, A. G., and Baker, M. J. (2020). Biofluid diagnostics by ftir spectroscopy: A platform technology for cancer detection. *Cancer letters*, 477:122–130.

Shaikh, S., Yadav, D. K., and Rawal, R. (2021). Saliva based non invasive screening of oral submucous fibrosis using ATR-FTIR spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 203:114202.

Silva, L. G., Péres, A. F., Freitas, D. L., Morais, C. L., Martin, F. L., Crispim, J. C., and Lima, K. M. (2020). ATR-FTIR spectroscopy in blood plasma combined with multivariate analysis to detect hiv infection in pregnant women. *Scientific reports*, 10(1):1–7.

Sitnikova, V. E., Kotkova, M. A., Nosenko, T. N., Kotkova, T. N., Martynova, D. M., and Uspenskaya, M. V. (2020). Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis. *Talanta*, 214:120857.

Xu, D., Liu, S., Cai, Y., and Yang, C. (2019). Baseline correction method based on doubly reweighted penalized least squares. *Applied optics*, 58(14):3913–3920.

Zhang, L., Xiao, M., Wang, Y., Peng, S., Chen, Y., Zhang, D., Zhang, D., Guo, Y., Wang, X., Luo, H., et al. (2021). Fast screening and primary diagnosis of covid-19 by atr–ft-ir. *Analytical chemistry*, 93(4):2191–2199.