Diagnóstico do câncer oral através da classificação de alto nível

Ricardo B. Lima Filho[1], **Murillo G. Carneiro**[1]

[1]Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brasil

{ricardo.filho, mgcarneiro}@ufu.br

Abstract. *This work investigates high-level classification techniques derived from properties and measures of complex networks for the salivary detection of oral cancer using Total Attenuated Reflectance by Fourier Transform Infrared Spectroscopy (ATR-FTIR). ATR-FTIR is a sustainable, fast and non-invasive platform capable of contributing to the detection of several diseases. Among the several network measures evaluated in this study, our results indicate clustering coefficient as the most satisfactory one with* 71% *and* 81% *of accuracy and sensitivity, respectively. Moreover, the high-level technique outperformed several other classifiers used for spectral analysis, including state-of-the-art ones like support vector machine and convolutional neural networks.*

Resumo. *Este trabalho investiga técnicas de classificação de alto nível baseadas em propriedades e medidas de redes complexas para a detecção salivar de câncer de boca a partir da Reflectância Total Atenuada por Espectroscopia de Infravermelho por Transformada de Fourier (ATR-FTIR). ATR-FTIR é uma plataforma sustentável, rápida e não invasiva capaz de contribuir para a detecção de diversas doenças. Dentre as diversas medidas de rede avaliadas neste estudo, nossos resultados indicam o coeficiente de agrupamento como o mais satisfatório com* 71% *e* 81% *de acurácia e sensibilidade respectivamente. Além disso, a técnica de alto nível superou vários outros classificadores usados para análise espectral, incluindo os de última geração, como máquina de vetores de suporte e redes neurais convolucionais.*

1. Introdução

De acordo com o Instituto Nacional do Câncer (Inca), que atua no desenvolvimento e coordenação de ações integradas de prevenção e controle do câncer, o diagnóstico do câncer de boca é geralmente feito por exame clínico associado à biópsia invasiva do tecido, além disso, a detecção tardia da doença muitas vezes requer cirurgias invasivas que impactam na qualidade de vida desses pacientes [Freitas et al. 2016]. Assim, a saúde pública seria amplamente beneficiada com o desenvolvimento de soluções tecnológicas capazes de realizar o diagnóstico em estágio inicial da doença.

A espectroscopia infravermelha de biofluidos chama a atenção para um diagnóstico livre de reagentes de várias doenças e distúrbios, como COVID-19 e Diabetes [Barauna et al. 2021, Caixeta et al. 2023]. A coleta de saliva pode ser aplicada em plataformas de triagem e diagnóstico por meio desse método não invasivo, de baixo custo, indolor e de auto-coleta [Ferreira et al. 2020].

O termo Redes Complexas refere-se a um grafo que possui uma estrutura composta por um conjunto de vértices (nós) que são interligados através de arestas (links). Portanto, diferentes situações e aspectos do mundo real podem ser expressos por meio de redes complexas para resolver problemas específicos [Carneiro 2017]. Uma das vantagens da rede é a possibilidade de analisar padrões de alto nível dos dados da rede, ao invés de analisar apenas suas características físicas [Fernandes et al. 2023]. Técnicas de classificação que analisam padrões de formação da rede são conhecidas na literatura como classificadores de alto nível [Carneiro and Zhao 2018].

Neste artigo, é dado mais um passo na investigação de classificadores de alto nível, considerando sua investigação para o diagnóstico molecular do câncer de boca. Até onde sabemos, este é o primeiro estudo na literatura que avalia medidas e propriedades de redes complexas na análise de dados de espectroscopia ATR-FTIR. O principal objetivo é desenvolver um modelo de classificação de alto nível capaz de detectar pacientes com câncer de boca em estágio inicial. Ademais, as seguintes contribuições também são desdobramentos do estudo:

- Avaliação comparativa do desempenho preditivo obtido através de duas medidas de proximidade amplamente adotadas na literatura para a construção da rede de espectros ATR-FTIR;
- Avaliação comparativa do desempenho de seis medidas de redes complexas na classificação de alto nível de dados ATR-FTIR voltados para a detecção do câncer de boca;
- Comparação dos resultados obtidos com a classificação de alto nível em relação a técnicas conhecidas da literatura, incluindo algoritmos do estado-da-arte para ATR-FTIR.

2. Descrição do Classificador de alto nível para dados ATR-FTIR

Este trabalho propõe uma técnica de classificação de alto nível baseada em rede para o problema de diagnóstico do câncer de boca. A técnica pode ser dividida em duas grandes fases: fase de treinamento e fase de teste. A Figura 1 mostra os principais passos de cada fase da nossa técnica de classificação. As setas inferiores (vermelha) representam as etapas da fase de treinamento (fluxo representado na cor azul), enquanto as setas superiores (verde) representam as etapas da fase de teste (fluxo representado na cor verde).

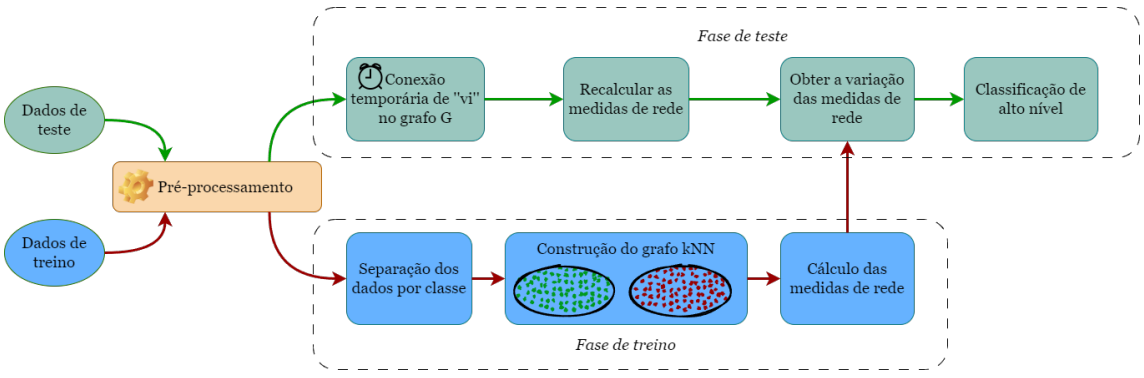

Figure 1. Estrutura principal do algoritmo de classificação de alto nível proposto

2.1. Etapa de Treinamento

A primeira etapa do classificador de alto nível baseado em redes complexas é construir a rede a partir dos dados de entrada. Neste cenário, considere X um conjunto de dados de treinamento na forma de $X_i = (x_i, y_i)$, em que x_i denota um dado na forma de vetor com m atributos e y_i a classe de destino associado a x_i. Formalmente, uma rede $G = \{V, E\}$ é obtida de X, na qual $v_i \in V$ representa um item (espectro) x_i e cada aresta $e_{ij} \in E$ refere-se a uma conexão entre os vértices v_i e v_j de acordo com algum critério de afinidade. Particularmente, são investigados dois critérios de afinidade nesta iniciação científica, selecionados pela grande representatividade que possuem na área: distância Euclidiana e similaridade do cosseno. Além disso, o método *Mutual k-Nearest Neighbors Graph* (MKNNG) [Carneiro 2017] foi adotado como a heurística responsável por transformar os dados vetoriais de características em uma rede. MKNNG cria uma aresta entre dois vértices v_i e v_j se eles tiverem a mesma classe e se ambos vértices estiverem entre os k vizinhos mais próximos um do outro.

Na classificação de alto nível, as estruturas e topologias são capturadas dos dados em rede através de medidas de rede complexas. Apesar da literatura conter uma ampla gama de medidas de rede [Carneiro 2017], neste estudo foram adotadas seis medidas que estão listados pela Tabela 1 e tiveram seu desempenho preditivo analisado individualmente.

Table 1. Medidas de redes complexas escolhidas neste estudo com suas definições e referências.

Medidas de rede (Abreviação)	Definição
Assortatividade	$\frac{L^{-1}\sum_u i_u k_u - \left[L^{-1}\sum_u \frac{1}{2}(i_u+k_u)\right]^2}{L^{-1}\sum_u \frac{1}{2}(i_u^2+k_u^2) - \left[L^{-1}\sum_u \frac{1}{2}(i_u+k_u)\right]^2}$
Coeficiente de Agrupamento	$\frac{1}{n}\sum_{i=1}^{n} \frac{\lvert e_{us}\rvert}{k_i(k_i-1)}$
Grau médio	$\frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{n} a_{ij}$
Comprimento médio do caminho mais curto (Short.Path)	$\frac{1}{n(n-1)}\sum_{i\neq j} D_{i,j}$
Closeness	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-1}{\sum_{j=1}^{n} d(i,j)}\right)$
Intermedialidade	$b_i = \sum_{u,v\in V-i} n_{uv}^i$

2.2. Etapa de teste

Na fase de teste, a instância a ser testada é temporariamente inserida em cada componente da rede (classe) considerando a heurística dos k-vizinhos mais próximos. Como os rótulos dos dados de teste são desconhecidos, é necessário avaliar o impacto causado em cada medida de rede após a inserção em cada componente da rede (classe). Para identificar a classe em que a nova instância se encaixa melhor, a classificação de alto nível utiliza medidas de rede que, quando calculadas antes e depois da inserção de uma nova instância,

podem fornecer "*insights*" úteis sobre o impacto (variação) causado naquela classe de rede.

Com a variação de cada medida da rede, é possível analisar o impacto da inserção do objeto de teste em cada classe. Quanto maior a variação da medida da rede, menor é a probabilidade do objeto de teste pertencer àquele componente da rede (classe); caso contrário, quanto menor a variação da medida de rede, maior será tal probabilidade, evidenciando que o objeto esta em maior conformidade com o padrão daquela classe [Carneiro et al. 2021].

3. Base de dados de câncer de boca e seu pré-processamento

A base de dados foi obtida com aprovação do Comitê de Ética em Pesquisas da Universidade Federal de Uberlândia (UFU), sob protocolo 249.200.9, no Hospital de Clínicas da UFU. 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. Em suma, a base de dados consiste em 65 instâncias nas quais a espectroscopia ATR-FTIR foi registrada para 1868 bandas infravermelhas (atributos). No conjunto de dados, temos 26 amostras do grupo controle (neg.) e 39 do grupo de câncer oral (pos.).

Em relação às etapas de preparação e pré-tratamento dos dados, nesta investigação é procedido da seguinte forma:

1. Correção da linha de base do espectro usando os métodos de mínimos quadrados assimétricos [Boelens et al. 2004] e elástico [Baek et al. 2015];
2. Normalização dos espectros pelo pico da amida I (região da banda do infravermelho entre 1630cm^{-1} e 1660cm^{-1});
3. Truncamento dos espectros para a região entre 1800cm^{-1} e 900cm^{-1}, a fim de considerar regiões com menos propensão a ruídos e outliers.

4. Resultados experimentais

Nos experimentos apresentados nesta seção, é adotado um processo de validação cruzada em 10 subconjuntos repetido três vezes. Nesse método, o conjunto de dados é particionado em 10 partes, das quais nove são adotadas para treinar o modelo e uma para testá-lo. Como cada parte é selecionada para o teste uma vez, temos um total de 10 simulações. Também é repetido o processo de particionamento três vezes, alterando a configuração dos subconjuntos, o que ao final representa um total de 30 simulações. Os parâmetros considerados pelo nosso classificador de alto nível foram a medida de proximidade, na qual foram investigadas a euclidiana e cosseno, bem como o parâmetro k, que é responsável por definir o número de vizinhos a serem considerados para a construção da rede usando MKNNG. Tais valores foram definidos no seguinte intervalo $k \in \{1, 2, \ldots, 15\}$.

Para fins comparativos, nosso técnica de classificação é avaliada em relação a outras técnicas amplamente adotadas para dados ATR-FTIR, como análise discriminante linear (LDA) e técnicas avançadas, como máquina de vetores de suporte com kernel gaussiano (SVM-RBF), além de técnicas tradicionais bem conhecidas, como Naive Bayes (NB) e Perceptron multicamadas (MLP), e das redes neurais convolucionais (CNN),

técnica de aprendizado profundo de última geração. Com exceção de LDA e NB, todas as outras técnicas tiveram seus parâmetros devidamente tunados através de diversas configurações de hiperparâmetros.

A Tabela 2 mostra os resultados preditivos obtidos por cada uma das técnicas sob comparação. Pode-se ver que os melhores resultados foram obtidos pela técnica proposta, cujo gráfico foi construído usando o cosseno e a análise de alto nível usando o Coeficiente de agrupamento, seguida pela técnica de alto nível com Grau médio, NB, SVM-RBF, Comprimento médio do caminho mais curto e CNN. Em relação às demais medidas de redes complexas, assortatividade e closeness tiveram resultados razoáveis em termos de sensibilidade, mas resultados baixos em termos de especificidade. O oposto aconteceu com a medida de Intermedialidade, que apresentou os maiores resultados de especificidade, mas valores intermediários de sensibilidade. Em relação às técnicas comumente adotadas na literatura ATR-FTIR, LDA apresentou os piores resultados e SVM-RBF demonstrou sua robustez ao alcançar resultados competitivos contra a classificação de alto nível com Grau médio, embora tenha sido superado nas quatro métricas preditivas pela técnica de classificação de alto nível com Coeficiente de agrupamento. Aliás, esta medida também superou a CNN, técnica considerada estado-da-arte, nas quatro métricas preditivas.

Table 2. Resultados da detecção de câncer de boca em termos de Acurácia Média (AA), Sensibilidade (SE), Especificidade (SP) e Média (SE,SP) usando a técnica de classificação de alto nível com diferentes medidas de redes complexas em comparação com outros classificadores da literatura.

Algoritmo	Medidas de rede	Proximidade	AA	SE	SP	Média
LDA	-	-	0.53	0.58	0.46	0.52
NB	-	-	0.64	0.58	**0.80**	0.67
KNN	-	Euclidiana	0.59	0.68	0.48	0.58
MLP	-	-	0.52	0.41	**0.80**	0.57
CNN	-	-	0.65	0.69	0.60	0.65
SVM-RBF	-	-	0.68	0.78	0.54	0.66
High-level (Proposto)	Assortatividade	Euclidiana	0.59	0.68	0.48	0.58
	Intermedialidade	Cosseno	0.60	0.52	0.73	0.63
	Closeness	Euclidiana	0.66	0.71	0.57	0.64
	Short.Path	Cosseno	0.67	0.75	0.55	0.65
	Grau médio	Euclidiana	0.65	0.62	0.71	0.67
	Coeficiente de agrupamento	Cosseno	**0.71**	**0.81**	0.57	**0.69***

5. Considerações Finais

Em síntese, nosso estudo apresenta contribuições sobre as perspectivas de representação de dados ATR-FTIR em rede e sobre o potencial da classificação de alto nível para o problema de detecção de câncer de boca. No primeiro ponto, o estudo fornece evidências de que algumas medidas de proximidade podem contribuir mais do que outras na construção de uma rede mais efetiva. Esse foi o caso da similaridade de cosseno, que apresentou o melhor desempenho preditivo para a detecção do câncer de boca. No segundo ponto, merece destaque a capacidade da nossa técnica de classificação de alto nível baseada em redes complexas em lidar com distribuições arbitrárias a partir de análise de aspectos estruturais dos dados em rede. De fato, a técnica possui ampla capacidade de representação

para se adaptar a diferentes problemas devido à grande quantidade de medidas de rede capazes de capturar diferentes propriedades da estrutura do grafo, demostrando a robustez do modelo. Finalmente, outras contribuições desse estudo podem ser evidenciadas pela produção bibliográfica, em que o artigo *Molecular Detection of Oral Cancer Using Complex Networks* foi gerado e se encontra em fase de revisão para publicação em veículo de boa relevância, bem como na possibilidade de depósito de patente a partir de desdobramentos originados do presente estudo.

References

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.

Barauna, V. G., Singh, M. N., Barbosa, L. L., Marcarini, W. D., Vassallo, P. F., Mill, J. G., Ribeiro-Rodrigues, R., Campos, L. C., Warnke, P. H., and Martin, F. L. (2021). Ultrarapid on-site detection of sars-cov-2 infection using simple atr-ftir spectroscopy and an analysis algorithm: high sensitivity and specificity. *Analytical Chemistry*, 93(5):2950–2958.

Boelens, H. F., Dijkstra, R. J., Eilers, P. H., Fitzpatrick, F., and Westerhuis, J. A. (2004). New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and raman spectroscopic detection. *Journal of chromatography A*, 1057(1-2):21–30.

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. (2017). *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. PhD thesis, Universidade de São Paulo.

Carneiro, M. G., Gama, B. C., and Ribeiro, O. S. (2021). Complex network measures for data classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Carneiro, M. G. and Zhao, L. (2018). Organizational data classification based on the importance concept of complex networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3361–3373.

Fernandes, J. M., Oliveira, G. M. B., and Carneiro, M. G. (2023). Network optimization based on genetic algorithm for high-level data classification. *IEEE Latin America Transactions*, 21(2):295–301.

Ferreira, I. C., Aguiar, E. M., Silva, A. T., Santos, L. L., Cardoso-Sousa, L., Araujo, T. G., Santos, D. W., Goulart, L. R., Sabino-Silva, R., and Maia, Y. C. (2020). Attenuated total reflection-fourier transform infrared (atr-ftir) spectroscopy analysis of saliva for breast cancer diagnosis. *Journal of oncology*, 2020.

Freitas, R. M., Rodrigues, A. M. X., Matos Júnior, A., and Oliveira, G. (2016). Fatores de risco e principais alterações citopatológicas do câncer bucal: uma revisão de literatura. *Rbac*, 48(1):13–8.