High-level classification using complex networks for Autism Spectrum Disorder detection

Lucas G. T. Araújo[1], **Robinson Sabino-Silva**[2], **Murillo G. Carneiro**[1]

[1]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

lucas.teodoro@ufu.br , robinsonsabino@gmail.com , mgcarneiro@ufu.br

Abstract. *The diagnosis of Autism Spectrum Disorder (ASD) is typically based on behavioral observation, which is a process time-consuming, subjective and reliant on professional judgment. This study leverages research on salivary biomarkers to develop a tool capable of adding objectivity to this process. A high-level classifier based on complex networks was employed using different network formation methods based on Attenuated Total Reflection Fourier-Transform Infrared spectroscopy (ATR-FTIR) data from saliva samples. The results indicate the use of high-level classifiers as a promising tool for ASD detection.*

Resumo. *O diagnóstico do Transtorno do Espectro Autista (TEA) se dá por observação comportamental, em um processo demorado, subjetivo e dependente de um profissional qualificado. Este trabalho investiga biomarcadores salivares no desenvolvimento de uma ferramenta capaz de adicionar objetividade nesse processo. Para isso foi utilizado um classificador de alto nível baseado em redes complexas com diferentes métodos de formação da rede a partir de dados de espectroscopia de infravermelho (ATR-FTIR) de amostras de saliva. Os resultados apontam o uso de classificadores de alto nível como técnica promissora para a detecção do TEA.*

1. Introdução

O Transtorno do Espectro Autista (TEA) é caracterizado pelo déficit de comunicação em diferentes graus como dificuldades em compreender mensagens não verbais, fazer novos amigos ou em iniciar conversas [American Psychiatric Association 2014]. Em graus mais extremos podemos encontrar a criação de rituais ou rotinas fixas, as quais se não forem seguidas, gera muito estresse ao indivíduo[American Psychiatric Association 2014]. Atualmente o diagnóstico é realizado por especialistas mediante observação comportamental [Ministério da Saúde do Brasil 2021]. Existem, porém, diversos estudos que buscam maneiras mais efetivas de detectar o TEA, inclusive utilizando análises químicas e/ou biológicas [Michelassi et al. 2023]. Esse trabalho se propõe a analisar um desses métodos o qual considerara a análise de dados sequenciais obtidos via espectroscopia no infravermelho por transformada de Fourier com reflexão total atenuada (ATR-FTIR) de amostras de saliva [Silva 2020].

Embora a saliva seja majoritariamente composta por água nela também estão presentes secreções da gengiva, células epiteliais e imunológicas, e microrganismos, além

de grandes quantidades de proteínas, RNAs e micro RNAs, vírus, fungos e bactérias [Dawes and Wong 2019]. Dessa forma, a saliva se apresenta como um bom representante do estado fisiológico e patológico do indivíduo. Além disso, as proteínas podem atuar como biomarcadores por refletirem as dinâmicas genéticas e ambientais.

A espectroscopia ATR-FTIR é uma técnica que se vale da absorção de ondas eletromagnéticas da região do infravermelho médio pelos materiais analisados obtendo informações quantitativas e qualitativas de multicomponentes através da vibração dos átomos que compõe as moléculas da amostra [Baker et al. 2014, Morais et al. 2019]. A técnica tem eficaz aplicação na área da saúde para análise de tecidos e fluidos corporais, sendo, portanto, ferramenta útil para diagnóstico de câncer e diabetes, por exemplo [Lima-Filho and Carneiro 2023, Caixeta et al. 2023]. Na literatura, o uso de algoritmos de aprendizado de máquina para a análise dos espectros ATR-FTIR ganha cada vez mais destaque, com soluções voltadas para diagnóstico de câncer de boca [Lima-Filho and Carneiro 2023], detecção do Zika vírus [Oliveira et al. 2023] e detecção de COVID-19 [Zhang et al. 2021]. A ATR-FTIR é por sua vez uma ferramenta de análise não invasiva de fluidos salivares a qual não precisa de reagentes ou preparação especial das amostras com sensibilidade similar a métodos bioquímicos clássicos de análise [Caixeta et al. 2023].

A classificação de dados é uma das tarefas mais comuns do chamado aprendizado de máquina supervisionado, i.e., quando as classes dos objetos analisados durante o processo de aprendizado, ou treinamento, já são conhecidos a priori [Carneiro 2016]. Todavia a maioria desses métodos considera apenas as características físicas dos dados analisados, os quais são representados majoritariamente na forma de matriz ou vetor de atributos [Carneiro 2016, Carneiro and Zhao 2018a]. Podemos utilizar algoritmos de classificação tradicionais como esses para extrair padrões nos dados e detectar o TEA, utilizando métricas de (dis)similaridade para comparar cada uma das instancias de dados.

Existe ainda uma segunda categoria de algoritmos de classificação, chamados de classificadores de alto nível, os quais consideram não só características físicas dos dados, mas também atributos estruturais e/ou topológicos dos dados [Carneiro and Zhao 2018b]. Nesse sentidos os dados devem estar dispostos em uma formato de grafo ou rede, portanto os dados de entrada são submetidos a um método que os represente como uma rede antes de serem processados no treinamento do classificador [Carneiro 2016].

Este trabalho tem como objetivo avaliar a hipótese de detectar o transtorno do espectro autista utilizando dados extraídos da saliva do paciente através de espectroscopia de infravermelho, desenvolvendo, para isso, diferentes heurísticas de formação da rede para os algoritmos de classificação de alto nível. Especificamente, o presente trabalho investiga uma técnica de classificação de alto nível via caracterização de importância capaz de explorar propriedades topológicas e estruturais dos dados em rede para o processo de classificação. O trabalho também avalia a técnica investigada comparando-a com algoritmos clássicos e do estado-da-arte para classificação de dados.

O diagnóstico do TEA é feito por profissionais, os quais avaliam como os indivíduos reagem a diferentes estímulos sociais [American Psychiatric Association 2014]. Uma análise como essa pode demandar anos até que a conclusão do diagnóstico dificultando uma intervenção terapêutica direcionada ao TEA

[Ministério da Saúde do Brasil 2021]. Com isso em mente a proposta de um método de análise agnóstica ao comportamento valendo-se de atributos químicos e biológicos, como o investigado neste trabalho, representa a possibilidade de agilizar a detecção, e portanto a adoção de uma abordagem terapêutica para pacientes com TEA.

Este trabalho possui a seguinte organização. A Seção 2 apresenta os principais trabalhos relacionados. A Seção 3 descreve o desenvolvimento do trabalho, o formato da base de dados e a aplicação de cada algoritmo. A Seção 4 traz os resultados obtidos por cada uma das técnicas adotadas e compara o desempenho de cada uma. Por fim, a Seção 5 conclui o trabalho.

2. Trabalhos relacionados

Esta seção tem por objetivo apresentar alguns trabalhos relacionados à utilização de técnicas de aprendizado de máquina para detecção do TEA e diferentes utilização de classificadores de alto nível na literatura.

Em [Simeoli et al. 2024] temos uma revisão da literatura onde são destacados algumas abordagens de aprendizado de máquina na detecção do TEA. O trabalho destaca que tais técnicas podem colaborar com o diagnóstico do transtorno e mapeia os marcadores adotados na literatura tais como movimentação das mãos e dos olhos. Em sua conclusão o autor aponta que estas pesquisas contribuem na busca de biomarcadores para tarefa de diagnóstico do TEA.

O trabalho de [Abdelwahab et al. 2024] utiliza bases não clínicas de TEA, as quais possuem dados categóricos, sequenciais e binários, disponíveis publicamente e adota diferentes algoritmos tradicionais de aprendizado supervisionado na tarefa de detectar o transtorno em indivíduos de diferentes idades. Todavia, o maior foco do trabalho está em crianças visto que o diagnóstico do TEA ainda na infância é muito importante para uma melhor efetividade dos procedimentos médicos. Dentre os resultados atingidos nesse trabalho destacam-se os dos classificadores *Naive Bayes* e *Random Forest*.

Outra ferramenta para auxiliar o diagnóstico do TEA é proposta em [Michelassi et al. 2023] a qual utiliza detecção facial e classificação de imagens. Sabendo que características faciais podem estar relacionadas ao desenvolvimento cerebral, a utilização de classificadores capazes de processar uma face e inferir o diagnóstico para TEA é uma ideia promissora. Essa abordagem busca adicionar mais objetividade ao diagnóstico, destacando a relevância de escolher os algoritmos de detecção facial para otimizar os resultados.

O trabalho de [Silva 2020] para a detecção do TEA propõe, por sua vez, uma análise de biomarcadores salivares via ATR-FTIR. Esta metodologia não invasiva oferece uma assinatura espectral salivar, indicando a viabilidade de um método rápido para o diagnóstico do TEA em comparação às abordagens de análise comportamental. Os resultados preliminares sugerem a presença de potenciais biomarcadores a partir da análise estatística univariada, proporcionando uma perspectiva exploratória alternativa às abordagens de aprendizado de máquina tradicionais.

De modo similar, [Lima-Filho and Carneiro 2023] também investiga a utilização de dados de ATR-FTIR extraídos da saliva, porém para a detecção do câncer de boca. Para isso, os autores adotam classificadores de alto nível a partir de métricas de redes

complexas para classificar os espectros ATR-FTIR. A fase de treinamento é definida pela construção do grafo usando algoritmo do *KNN-Graph* (KNNG) e pela caracterização estrutural dos componentes (classes) do grafo a partir de diferentes medidas de redes complexas; e a fase de testes considera a classificação de novos dados ATR-FTIR analisando o quanto o padrão estrutural deles está em consonância ao padrão estrutural dos componentes (classes) formados pelos dados de treinamento. O modelo construído obteve bons resultados de acurácia e sensibilidade.

3. Materiais e Métodos

Nesta seção será apresentado um resumo das ferramentas e componentes utilizados nesse trabalho, incluindo o conjunto de dados adotado e a descrição do algoritmo de classificação de alto nível.

3.1. Base de dados

Neste trabalho adotamos uma base de dados composta por 159 amostras de saliva de 53 pacientes, três para cada. Os registros do *dataset* representam cada uma das coletas de saliva submetidas à ATR-FTIR, pré-processadas e expressas no formato de vetor de atributos, somando 34 pacientes neurotípicos e 19 pacientes com TEA, dispostos em um vetor de atributos com 159 registros (três por paciente) e 1868 atributos. Os dados adotados neste trabalho foram coletados em [Silva 2020] com aprovação do Comitê de Ética em Pesquisa (CEP) com seres humanos da Universidade Federal de Uberlândia, e disponibilizado para nosso uso no formato de vetor de atributos.

Como os dados extraídos via ATR-FTIR são suscetíveis a ruído, adotamos também dois pré-processamentos os quais são capazes de aumentar a estabilidade dos mesmos, uma limitando as frequências entre 900 e $1800cm^{-1}$, a qual chamaremos de truncamento, e outra baseada em normalização pela *amida I* a qual é dada por [Caixeta et al. 2023]:

$$\forall(X_i)\ X_{i_{norm}} = \frac{X_i}{\max(X_{amideI})}, \tag{1}$$

onde X_{amideI} é o subconjunto das colunas de X compreendidas entre 1630 e 1660 cm^{-1} e X é uma linha do vetor de atributos da base de dados.

3.2. Classificação de Alto Nível Baseada em Importância

A classificação baseada em importância é um algoritmo de aprendizado supervisionado o qual utiliza heurísticas baseadas em grafo no processo de aprendizado e predição sendo, portanto, um classificador de alto nível baseado em redes. O algoritmo recebe esse nome por considerar a importância individual dos nós do grafo para classificar uma nova amostra, colocando-a na classe em que maximize essa importância. O conceito de importância é baseado na heurística do *pagerank*, algoritmo baseado em grafos adotado pelo Google em seu buscador, por esse motivo também podemos chamá-lo de classificador PGR [Carneiro and Zhao 2018b]. Esse classificador, assim como outros, possui duas fases a de treinamento e de teste.

- **Fase de Treinamento:** Na fase de treinamento dado o grafo $\mathcal{G}_{v,a}$ construído de um vetor de atributos X, calculamos os padrões de eficiência E e a importância individual I das amostras.

- **Fase de Teste:** Na fase de teste um novo objeto y é apresentado ao classificador, baseado na medida de eficiência diferencial espaço-estrutural são selecionados um conjunto de vértices para se ligar a y temporariamente, calcula-se então a importância de y para cada classe, de modo que o novo objeto recebe o rótulo da classe que possuir maior importância.

  A importância I de uma instância y em relação a uma classe C é dada por:

$$I_y^{(C)} = \sum_{v_j \in \wedge_y^C} I_j, \tag{2}$$

onde v_j é um nó pertencente à base de treinamento X, ou seja, já rotulado, e $\wedge_y^C$ é o conjunto de nós da classe C aos quais y está temporariamente conectado.

O algoritmo de classificação baseado em importância utiliza uma heurística chamada de eficiência diferencial espaço-estrutural ($\mathcal{F}$) a qual mensura a eficiência do envio de uma informação entre dois nós em um componente do grafo e o fluxo de informação através de uma aresta ponderada por uma medida de similaridade entre duas instâncias. Em poucas palavras, ele considera informações físicas e topológicas da rede e do vetor de atributos original [Carneiro and Zhao 2018b] e pode ser definido como:

$$\mathcal{F}_{y,j} = \mathcal{E}_{j \in a}^{a} \cdot \gamma - S_{y,j}, \tag{3}$$

onde o primeiro termo de $\mathcal{F}$, definido pela multiplicação do fator de eficiência $\mathcal{E}$ e o parâmetro de balanceamento γ, captura as propriedades topológicas do grafo e o segundo termo caracteriza a relação espacial dos dados [Carneiro and Zhao 2018b].

A eficiência $\mathcal{E}$, por sua vez é definida como:

$$\mathcal{E}^a = \frac{1}{|N^a|} \sum_{i \in a} \xi_i^a, \tag{4}$$

onde $|N^a|$ é a quantidade de vértices no componente a e ξ é a eficiência local de um nó rotulado pertencente a a, a qual é definida por:

$$\xi_i^a = \frac{1}{N_i} \sum_{i \to j} S_{i,j}, \tag{5}$$

onde N_i é o número de ligações entre os nós i e j, e $S_{i,j}$ é o peso da aresta que liga os vértices, o qual é definido utilizando alguma métrica de (dis)similaridade dos dados. A escolha de tal métrica deve levar em consideração a natureza dos dados.

Finalmente, cabe destacar que o nó v_y apresentado ao classificador será conectado temporariamente à v_j pela seguinte regra [Carneiro and Zhao 2018b]:

$$\{v_j \in \wedge_y^C \mid \mathcal{F}_{y,j} \geq 0 \; e \; v_j \in C\}. \tag{6}$$

3.3. Construção da Rede

Como os dados estão dispostos originalmente no formato de vetor de atributos é necessário utilizar uma heurística de construção de grafo para dispô-los no formato adotado pelo classificador de alto nível. Existem, para essa tarefa, diferentes algoritmos na literatura, destacando-se entre eles o grafo *KNN-Graph* [Carneiro and Zhao 2018a, Freitas and Carneiro 2019].

Tomemos A_v como a matriz de adjacência de um grafo não direcionado $\mathcal{G}_{(v,a)}$ onde cada v_i representa um nó do grafo e o valor de $A_{i,j}$ dita se existe ou não uma conexão entre v_i e v_j, S_v como a matriz de distâncias onde o valor de $S_{i,j}$ é a distância entre as instâncias representadas pelos vértices v_i e v_j e C_i como a classe o objeto v_i. A seguir apresentamos as definições correspondentes aos algoritmos de construção da rede adotados nesse trabalho:

O método *KNN-Graph* conecta um nó a seus k nós mais próximos, segundo uma métrica de distância, através dos seguintes passos [Carneiro 2016]:

1. Define-se a quantidade k de vizinhos;
2. Calcula-se a distância entre cada par de instâncias do vetor de atributos;
3. Para cada vértice adiciona-se uma aresta entre seus k pares com maior similaridade dentro de uma mesma classe.

Dessa forma a matriz de adjacência será dada pela seguinte função:

$$A_{ij} = \begin{cases} 1, \ se\ C_j = C_i\ e\ v_j \in kNN_i \\ 0, \ caso\ contrario \end{cases} \tag{7}$$

onde A é a matriz de adjacência do *KNN-Graph*, C_i é a classe de v_i e kNN_i é a lista dos k vizinhos mais próximos de v_i.

Duas variações do *KNN-Graph* são consideradas nesta investigação:

- **SKNNG** ou *KNN-Graph Simétrico*, em que as conexões da rede são obtidas por:

$$A' = max(A, A^T). \tag{8}$$

- **MKNNG** ou *KNN-Graph Mútuo*, em que as conexões da rede são obtidas por:

$$A' = min(A, A^T). \tag{9}$$

Concluída a etapa de construção do grafo, cada espectro ATR-FTIR será representado por um vértice correspondente no grafo, cujas arestas são estabelecidas com outros vértices (espectros) a partir de algum critério de afinidade.

3.4. Medidas de dissimilaridade

Ambas os métodos de formação da rede adotados necessitam de uma heurística de (dis)similaridade para construir o grafo partindo de dados tabulares. A escolha desse critério de comparação pode influenciar diretamente nos resultados obtidos pelo classificador, por este motivo adotamos duas das principais heurísticas de distância da literatura: a distância euclidiana e a distância de cossenos definidas respectivamente como:

- Distância euclidiana:

$$dis(a,b) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}, \tag{10}$$

- Distância de Cosseno:

$$dis(a,b) = 1 - \frac{a \cdot b}{||a||_2 \cdot ||b||_2}, \tag{11}$$

onde: a e b são os vetores que representam cada um dos objetos e d é o número de dimensões geométrica dos mesmos.

3.5. Métricas de desempenho

Com objetivo de validar o desempenho do PGR iremos comparar os resultados obtidos por ele e por dois algoritmos de classificação já consolidados na literatura, o kNN e o *Random Forest*. A Tabela 1 resume a variação dos parâmetros de execução de cada algoritmo adotado neste trabalho. Em nossa comparação utilizamos o parâmetro que maximiza o resultado de cada algoritmo. Como métricas de avaliação de desempenho foram adotadas as seguintes medidas (12)-(17): Acurácia, Precisão, Sensibilidade (Recall), Especificidade, F1 e Média Harmônica (MH) entre Sensibilidade e Especificidade. A divisão da base em conjunto de treino e teste foi feito paciente a paciente, seguindo o método *hold-out*, 39 pacientes na pasta de treino e 14 pacientes na pasta de teste, de modo que as amostras de uma mesma pessoa estejam sempre no mesmo conjunto. Cada modelo foi avaliado 10 vezes, com diferentes divisões de treino e teste, e ao final feita a media aritmética de seu desempenho.

Tabela 1. variação dos argumentos de execução por algoritmo. Onde $\overline{X}$ média aritmética das distâncias dos objetos na base de dados

Parâmetro	Variação	Algoritmos
k	[1, 50]	SkNNG
k	[1, 50]	MkNNG
k	[1, 100]	kNN *classifier*
n	[1, 100]	*Random Forest classifier*

Acurácia:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}, \tag{12}$$

Precisão:

$$Prec = \frac{tp}{tp + fp}, \tag{13}$$

Sensibilidade:

$$Sens = \frac{tp}{tp + fn}, \tag{14}$$

Especificidade:

$$Esp = \frac{tn}{fn + tp}, \tag{15}$$

Medida F1:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{16}$$

Média harmônica:

$$MH = 2 \cdot \frac{E \cdot R}{E + R}. \tag{17}$$

onde tp é o numero de verdadeiros positivos, tn é o numero de verdadeiros negativos, fp é o numero de falsos positivos, fn é o número de falsos negativos.

4. Resultados

As Tabelas 2 e 3 trazem, respectivamente, os melhores resultados obtidos utilizando a classificação de alto nível baseada em importância para o *dataset* original e para o *dataset* truncado e normalizado. Pela figura, é possível perceber que o melhor resultado obtido em ambas as tabelas considera a similaridade do cosseno e o grafo $MkNNG$ como método de construção da rede. Nesse sentido, os grafos construídos com *kNN-Graph Mútuo* foram os que obtiveram maiores médias harmônica e F1, além de ser o único método de formação da rede que não performou abaixo dos 60% em nenhuma das métricas. Por outro lado, o grafo *kNN-Graph Simétrico* apresentou melhor desempenho com a similaridade do cosseno no dataset original (Tabela 2), e com a distância euclidiana no dataset truncado e normalizado (Tabela 3).

Observando as matrizes de confusão na Figura 1, construídas com a média das execuções do classificador, podemos observar que o *kNN-Graph Mútuo* possui menor número de falsos negativos, todavia é o que possui o maior número de falsos positivos.

Tabela 2. Melhores resultados obtidos por cada um dos algoritmos de construção da rede utilizando as distância euclidiana e cosseno para o *dataset* original.

Distância	Algoritmo	Acurácia	Precisão	Recall	Especificidade	F1	MH
Euclid.	$SkNN$	0.72	**0.76**	0.55	0.81	0.57	0.65
	$MkNN$	0.71	0.60	0.64	0.74	0.61	0.69
Cosseno	$SkNN$	0.70	0.61	**0.73**	0.68	0.64	0.70
	$MkNN$	**0.75**	0.65	0.69	**0.78**	**0.66**	**0.73**

Tabela 3. Melhores resultados obtidos por cada um dos algoritmos de construção da rede utilizando as distância euclidiana e cosseno para o *dataset* truncado e normalizado pela amida I.

Distância	Algoritmo	Acurácia	Precisão	Recall	Especificidade	F1	MH
Euclid.	$SkNN$	0.71	0.84	0.41	0.89	0.49	0.56
	$MkNN$	0.67	0.64	**0.71**	0.65	0.62	0.68
Cosseno	$SkNN$	0.55	0.54	0.40	0.64	0.40	0.49
	$MkNN$	**0.75**	0.68	**0.71**	0.78	**0.67**	**0.74**

Os grafos produzidos pelo *kNN-Graph Mútuo* tendem a ser mais esparsos conectando apenas nós com elevado grau de similaridade visto que preserva reciprocidade de conexões. A heurística do *pagerank* tende a favorecer nós com maior grau. Para o *kNN-Graph Mútuo* esses nós tendem a ser nós centrais com maior grau de similaridade com várias instâncias que podem ser mais representativos para a classificação. Tal característica pode explicar a superioridade dos modelos que utilizam o *kNN-Graph Mútuo*.

Na Tabela 4 temos um resumo dos melhores resultados obtidos em cada uma das técnicas adotadas para classificação dos dados. Dentre os métodos já consolidados na literatura, o que mais se destacou foi a *Random Forest* que obteve bons resultados. Já para a classificação baseada em importância o melhor método para construir a rede foi

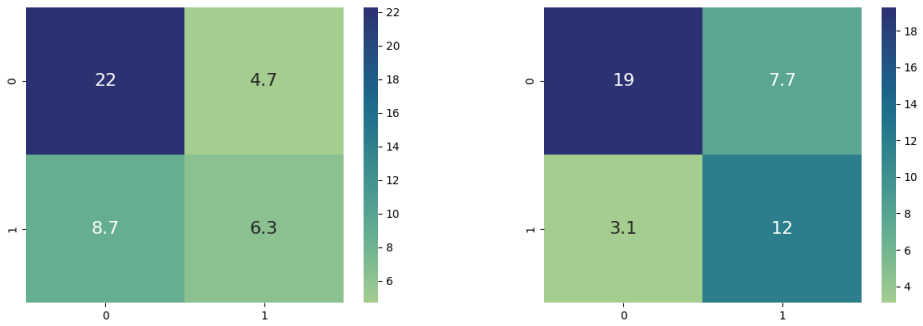

Figura 1. Matriz de confusão média do classificador de alto nível para o *kNN-Graph Simétrico* e *kNN-Graph Mútuo*, respectivamente, utilizando distância do cosseno para o *dataset* truncado e normalizado

o *kNN-Graph Mútuo* o qual performou de maneira satisfatória para ambos formatos do *dataset* e para ambas medidas de dissimilaridade.

Tabela 4. Comparação entre os melhores desempenhos dos modelos de classificação

Algoritmo	Acurácia	Precisão	Recall	Especificidade	F1	MH
KNN	0.73	0.65	0.58	**0.81**	0.60	0.67
Random Forest	**0.76**	0.67	0.69	0.80	**0.67**	**0.74**
PGR	0.75	**0.68**	**0.71**	0.78	**0.67**	**0.74**

É possível observar que os melhores resultados obtidos pelos modelos adotados foram muito próximos, sendo PGR com *kNN-Graph Mútuo* e *Random Forest* os que mais se destacaram. O provável motivo desse resultado é a resistência ao ruído e a melhor escolha dos comparadores, as instâncias com maior importância no caso do *kNN-Graph Mútuo* e as características mais significativas nas árvores que compõe o *Random Forest*.

5. Conclusão

Este trabalho apresentou uma abordagem para classificação de alto nível baseada em importância na detecção de TEA via espectros ATR-FTIR obtidos a partir de amostras de saliva. Especificamente, o estudo oferece uma análise inovadora de dados ATR-FTIR a partir de sua representação em redes e posterior análise considerando propriedades de redes complexas. Especificamente, o estudo avaliou métodos de construção de rede baseados no método *KNN-Graph* e diferentes métricas de dissimilaridade. Os experimentos considerando uma base de dados real, bem como o extenso conjunto de medidas de avaliação adotadas, evidenciam um potencial interessante da nossa abordagem na tarefa de detectar o TEA via aprendizado de máquina.

Foram avaliadas duas heurísticas de construção da rede, sendo duas variações do *KNN-Graph*baseadas em vizinhos mais próximos: o *KNN-Graph Simétrico* e o *KNN-Graph Mútuo*.A análise dos resultados demonstra grande estabilidade do *KNN-Graph Mútuo* como método de construção da rede para a tarefa de classificação de alto nível

do TEA, sendo superior aos demais métodos considerados para a tarefa e performando tão bem quanto o classificador *Random Forest*, e melhor que o classificador KNN. Observou-se também um ganho significativo na adoção da distância do cosseno como medida de dissimilaridade entre as amostras da base de dados em comparação com os resultados obtidos com a distância euclidiana.

Para trabalhos futuros sugere-se a adoção de heurísticas de construção da rede que considerem a característica sequencial dos dados, e.g., grafo de visibilidade; investigar outras medidas de rede para caracterização de importância; adotar estratégias de otimizações estruturais para o grafo; adotar um conjunto de dados com maior número de pacientes e que possa ser balanceado; e avaliar outras técnicas de classificação do estado-da-arte como redes neurais convolucionais e redes neurais de grafo.

Agradecimentos

Os autores agradecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (processos n. 402196/2021-0, 408216/2022-0 e 420212/2023-0), da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG (processo APQ-00410-21), do INCT em Teranóstica e Nanobiotecnologia (processo n. CNPq-465669/2014-0), e do INCT em Saúde Oral e Odontologia (processo n. CNPq-406840/2022-9). LGTA also thanks CNPq by his scholarship.

Referências

Abdelwahab, M. M., Al-Karawi, K. A., Hasanin, E., and Semary, H. (2024). Autism spectrum disorder prediction in children using machine learning. *Journal of Disability Research*, 3(1):20230064.

American Psychiatric Association (2014). *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Artmed Editora.

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.

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.

Carneiro, M. G. and Zhao, L. (2018a). Analysis of graph construction methods in supervised data classification. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 390–395. IEEE.

Carneiro, M. G. and Zhao, L. (2018b). Organizational data classification based on the importance concept of complex networks. *IEEE transactions on neural networks and learning systems*, 29(8):3361–3373.

Dawes, C. and Wong, D. (2019). Role of saliva and salivary diagnostics in the advancement of oral health. *Journal of dental research*, 98(2):133–141.

Freitas, L. M. and Carneiro, M. G. (2019). Community detection to invariant pattern clustering in images. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 610–615. IEEE.

Lima-Filho, R. B. and Carneiro, M. G. (2023). Diagnóstico do câncer oral através da classificação de alto nível. In *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 54–59. SBC.

Michelassi, G. C., Bortoletti, H. S., Pinheiro, T. D., Nobayashi, T., de Barros, F. R., Testa, R. L., Silva, A. F., Revers, M. C., Portolese, J., Pedrini, H., et al. (2023). Classification of facial images to assist in the diagnosis of autism spectrum disorder: A study on the effect of face detection and landmark identification algorithms. In *Brazilian Conference on Intelligent Systems*, pages 261–275. Springer.

Ministério da Saúde do Brasil (2021). Definição - Transtorno do Espectro Autista (TEA) na criança. Acessado em: 18/01/2022.

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.

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.

Silva, S. F. d. P. (2020). Avaliação de biomarcadores salivares para diagnóstico de transtorno de espectro autista por espectroscopia ATR-FTIR.

Simeoli, R., Rega, A., Cerasuolo, M., Nappo, R., and Marocco, D. (2024). Using machine learning for motion analysis to early detect autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, pages 1–20.

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-FTIR. *Analytical chemistry*, 93(4):2191–2199.