Redes Complexas para Classificação de Alto Nível de Assinaturas Vibracionais Moleculares

Anagê Calixto Mundim Filho

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Anagê Calixto Mundim Filho

Redes Complexas para Classificação de Alto Nível de Assinaturas Vibracionais Moleculares

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Murillo Guimarães Carneiro

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

M965 Mundim Filho, Anagê Calixto, 1999-
2025 Redes Complexas para Classificação de Alto Nível de Assinaturas Vibracionais Moleculares [recurso eletrônico] / Anagê Calixto Mundim Filho. - 2025.

Orientador: Murillo Guimarães Carneiro.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2025.599
Inclui bibliografia.

1. Computação. I. Carneiro, Murillo Guimarães,1988-, (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 34/2025, PPGCO				
Data:	25 de Setembro de 2025	Hora de início:	14:00	Hora de encerramento:	16:07
Matrícula do Discente:	12222CCP002				
Nome do Discente:	Anagê Calixto Mundim Filho				
Título do Trabalho:	Redes Complexas para Classificação de Alto Nível de Assinaturas Vibracionais Moleculares				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	• Aprendizado em redes para diagnóstico molecular da COVID-19 através da saliva (FAPEMIG APQ-00410-21) • Arquiteturas híbridas de aprendizado em redes para problemas de classificação mono e multi sequenciais com aplicações em dados de espectroscopia no infravermelho e de Eletroencefalograma (CNPQ 420212/2023-0)				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Bruno Augusto Nassif Travençolo - FACOM/UFU, Luis Paulo Faina Garcia - CIC/UnB e Murillo Guimarães Carneiro - FACOM/UFU, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Luis Paulo Faina Garcia - Brasilia/DF. Os outros membros da banca participaram da cidade de Uberlândia e o(a) aluno(a) Anagê Calixto Mundim Filho participou da cidade de Monte Carmelo/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Murillo Guimarães Carneiro, apresentou a Comissão Examinadora e o(á) candidato(a), agradeceu a presença do público, e concedeu ao(á) Discente a palavra para a exposição do seu trabalho

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(á) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos,

conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Luís Paulo Faina Garcia**, **Usuário Externo**, em 26/09/2025, às 09:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 28/09/2025, às 21:18, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 03/10/2025, às 14:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6671183** e o código CRC **6CAA45C1**.

Referência: Processo nº 23117.063431/2025-31 SEI nº 6671183

Para todos aqueles que acreditam em um mundo melhor a cada dia.

Agradecimentos

Primeiramente agradeço a oportunidade dada pelo meu Orientador Murillo Guimarães Carneiro, que está comigo junto desde a graduação e viu um potencial em mim para a carreira acadêmica, sempre me guiando e me dando boas oportunidades desde então. Agradeço aos amigos, colegas e professores que conheci nesse meio tempo, que sempre estiveram dispostos a me ajudar nessa área e deixá-la menos complexa.

Aos professores do Mestrado que deixaram as disciplinas mais divertidas. Prof. Robinson Sabino pelas nossas colaborações interdisciplinares e ensinamentos. Deixo meus agradecimentos a banca de defesa Profs Bruno Travençolo, Luis Paulo Garcia e Leandro Nogueira Couto pelo tempo de ler meu trabalho e participação na defesa do mesmo.

Agradeço a UFU pela oportunidade da vaga desde graduação até mestrado, contando com bolsas de projeto, estudo e iniciação científica que me fez chegarem até aqui. Aos órgãos de fomento FAPEMIG pela bolsa de projeto, CNPQ e CAPES pelos apoio.

Agradeço a Deus por estar sempre presente e me dar forças quando eu precisei para conseguir concluir este trabalho, minha Mãe Ana Paula Nunes e Avó Edna Montes Mundim e amigos próximos como o Gabriel Henrique por estarem do meu lado me dando apoio quando preciso. Meu Pai Anagê Calixto Mundim e avô Gregoriano Calixto Alves que não estão mais presentes porém me fazem sonhar e querer chegar mais longe, os quais ensinaram diversas lições que levo comigo.

Gostaria de agradecer principalmente a rede de apoio próxima que convive comigo todos os dias, meu padrasto Yann Nathan Silva e meu irmão Nathan Antônio, amigos, avôs, avós, tios e tias, amigos do grupo Deish, gatinho Totoro que adotei no início do mestrado e está comigo até hoje, passou pelas viagens e idas de Uberlândia a Monte carmelo! Cachorros e animaizinhos que foram presentes em diversos momentos importantes do mestrado e da minha vida

"Se eu vi mais longe, foi por estar sobre ombros de gigantes."
(Isaac Newton)

Resumo

A espectroscopia de infravermelho com transformada de Fourier por Refletância Total Atenuada (ATR-FTIR) tem se destacado como uma técnica promissora para diagnósticos rápidos, precisos e não invasivos de diversas condições patológicas. No entanto, a análise de seus dados espectrais ainda enfrenta desafios significativos devido à alta complexidade, dimensionalidade e sobreposição de grupos alvo e controle em aplicações reais, o que dificulta a distinção entre diferentes condições clínicas. Métodos tradicionais de classificação e até mesmo modelos de aprendizado profundo têm se mostrado limitados diante dessas dificuldades. Neste contexto, apresentamos o ATLA (ATR-FTIR Topological Learning Analysis), uma técnica inovadora de classificação espectral que integra informações topológicas, estruturais e físicas dos dados. Nosso método combina as avançadas capacidades topológicas das medidas de Redes Complexas com algoritmos de classificação tradicionais (Máquinas de Vetores de Suporte e Análise Discriminante Linear) e de aprendizado profundo moderno (Redes Neurais Convolucionais-BiLSTM). A eficácia do ATLA foi avaliada em bases de dados reais de espectros ATR-FTIR para detecção de pacientes com COVID-19, Hepatite Delta B e D e Doença de Chagas, todas com características espectrais desafiadoras. Os resultados mostraram que o ATLA contribuí na melhoria de desempenho preditivo dos modelos de aprendizado de máquina em termos de métricas como acurácia, precisão, sensibilidade, especificidade e F1-score. Além disso, o ATLA também pode oferecer maior interpretabilidade ao destacar as bandas e atributos topológicos mais relevantes para a classificação através de explicações SHAP (SHapley Additive exPlanations). Em síntese, a técnica ATLA se coloca como uma abordagem promissora e eficaz para elevar o diagnóstico espectral a um novo patamar de precisão e interpretabilidade.

Palavras-chave: Classificação Híbrida. Redes Complexas. Aprendizado Profundo. Aprendizado de Máquina. Classificação de dados. Espectroscopia ATR-FTIR. Diagnóstico Biomédico. Interpretabilidade. SHAP.

Abstract

Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy is a promising technique for rapid and non-invasive diagnostics, yet its clinical application is often constrained by significant analytical challenges. The high dimensionality, complexity, and spectral overlap inherent in real-world data make it difficult to distinguish between different pathological conditions, a task where both conventional classifiers and deep learning models have shown limitations. To address these challenges, we propose ATLA (ATR-FTIR Topological Learning Analysis), a novel framework that integrates topological, structural, and physical information for enhanced spectral classification. Our methodology leverages the feature extraction capabilities of Complex Networks in synergy with traditional classifiers (Support Vector Machines and Linear Discriminant Analysis) and a state-of-the-art deep learning architecture (Convolutional Neural Network-BiLSTM). We assessed ATLA's performance on challenging real-world ATR-FTIR datasets for the diagnosis of COVID-19, Hepatitis B and D, and Chagas Disease. Our findings indicate that ATLA significantly improves predictive outcomes across key metrics, including accuracy, precision, sensitivity, specificity, and F1-score. Beyond predictive power, ATLA also enhances model interpretability by using SHAP (SHapley Additive exPlanations) to highlight the most relevant spectral bands and topological features driving the classification. Ultimately, ATLA emerges as a robust and effective approach, poised to elevate spectral diagnostics to a new level of accuracy and interpretability.

Keywords: Hybrid Classification. Complex Networks. Deep Learning. Machine Learning. Data Classification. ATR-FTIR Spectroscopy. Biomedical Diagnosis. Interpretability. SHAP.

Lista de ilustrações

Lista de tabelas

Lista de siglas

AC	Acurácia
ADG	Grau médio, *Average Degree*
AM	Aprendizado de Máquina
ATLA	ATR-FTIR Topological Learning Analysis
ASP	Menor caminho médio - *Average shortest path*
ASS	Assortatividade, *Assortativity*
ATR	Reflectância Total Atenuada
BET	Intermedialidade, *Betwenness*
CNN	Redes neurais convolucionais
CCF	Coeficiente de Agrupamento - Cluster Coefficient
CLO	Proximidade, *Closeness*
ES	Especificidade
ATR-FTIR	Espectroscopia de Infravermelho por Transformada de Fourier com Reflectância Total Atenuada
IA	Inteligência Artificial
kNN	K-Vizinhos mais próximos, *K-nearest neighboors*
LDA	Análise Discriminante Linear
PR	Precisão
RCs	Redes complexas
SE	Sensibilidade
SHAP	SHapley Additive exPlanations
SVM	Support vector machine - Máquina de vetores de suporte

Sumário

Capítulo 1

Introdução

O aprimoramento de tecnologias analíticas de precisão, aliado aos avanços em Inteligência Artificial (IA), tem viabilizado novas abordagens para triagem clínica de doenças, oferecendo soluções promissoras para a triagem e o diagnóstico precoce em contextos médicos cada vez mais desafiadores. A dinâmica da sociedade contemporânea, marcada por uma hiperconectividade e mobilidade global sem precedentes, evidenciou, de forma paradoxal, a vulnerabilidade dos sistemas de saúde diante de crises como a pandemia de COVID-19 (CIOTTI et al., 2020; SURYASA; Rodríguez-Gámez; KOLDORIS, 2021).

A sobrecarga imposta aos sistemas de saúde e a urgência por métodos diagnósticos rápidos, precisos e não invasivos revelaram uma lacuna crítica no enfrentamento de doenças infecciosas e outras patologias (ALYASSERI et al., 2022). O diagnóstico de COVID-19, por exemplo, demonstrou ser um desafio devido à similaridade de sintomas com outras doenças respiratórias e à limitação de técnicas como o RT-PCR, que, embora padrão-ouro, pode ser demorado, podendo levar até 24 horas e sendo invasivo (MENEZES; LIMA; MARTINELLO, 2020), dificultando o diagnóstico.

Assim, a busca por técnicas de triagem mais rápidas e acessíveis, sem a dependência de reagentes e infraestrutura complexa, tornou-se imperativa para mitigar a propagação em massa e prevenir futuros cenários de pandemia. Nesse contexto, a espectroscopia de infravermelho com transformada de Fourier com Refletância Total Atenuada (ATR-FTIR) emerge como uma ferramenta promissora para a identificação molecular e diagnóstico biomédico (MOVASAGHI; REHMAN; REHMAN, 2008; SANTOS; MORAIS; LIMA, 2020; BUNACIU; HOANG; ABOUL-ENEIN, 2015).

Ao analisar a absorção da radiação infravermelha de uma amostra, o ATR-FTIR permite identificar e quantificar ligações químicas específicas, gerando "impressões digitais" moleculares detalhadas (MOVASAGHI; REHMAN; REHMAN, 2008). Esta técnica se destaca por sua versatilidade, alta sensibilidade e rapidez, possibilitando a análise de uma ampla variedade de amostras biológicas em minutos. A utilização de biofluídos, como a saliva, no ATR-FTIR é particularmente vantajosa por ser um método não invasivo, de baixo custo e com coleta simplificada, sendo amplamente investigado para o diagnóstico

precoce de diversas patologias, incluindo infecções virais como Zika (OLIVEIRA et al., 2023) e HIV (SILVA et al., 2020), e condições oncológicas como câncer cerebral (BUTLER et al., 2019), câncer de pulmão (PENG et al., 2022) e câncer de mama (SITNIKOVA et al., 2020).

Apesar do potencial do ATR-FTIR, a interpretação de seus dados espectrais apresenta desafios significativos devido à sua complexidade e à sobreposição de bandas, bem como à alta similaridade espectral entre diferentes estados de doença (ALYASSERI et al., 2022). Isso limita a eficácia de métodos estatísticos convencionais e exige abordagens computacionais avançadas para a detecção de padrões sutis e discriminativos (**??**). O Aprendizado de Máquina (AM), uma área da IA focada no desenvolvimento de técnicas computacionais que aprendem a partir de dados (ALPAYDIN, 2020; MONARD; BARANAUSKAS, 2003), e, mais especificamente, o aprendizado supervisionado, têm se mostrado paradigmas poderosos para a classificação de dados complexos, como os espectros de biofluídos (CUNNINGHAM; CORD; DELANY, 2008; ZHAO; SILVA; CARNEIRO, 2021).

Contudo, mesmo algoritmos de aprendizado profundo, como as Redes Neurais Convolucionais (CNNs), que são capazes de extrair características complexas e hierárquicas, podem falhar na captura de relações estruturais globais entre as amostras, um aspecto crucial em sistemas biológicos. Nesse cenário, as Redes complexas (RCs) emergem como um arcabouço complementar, capazes de modelar sistemas por meio de grafos com padrões de interconexão não triviais, revelando características topológicas e funcionais que não são evidentes em análises tradicionais (PÓSFAI; BARABASI, 2016; BOCCALETTI et al., 2006; CARNEIRO, 2017). O estudo das RCs para classificação de dados, observando a topologia e estrutura em dados em rede, tem demonstrado resultados promissores ao classificar corretamente amostras que outros algoritmos não conseguem (CARNEIRO; ZHAO, 2018).

A presente dissertação propõe o desenvolvimento de um método inovador de classificação espectral híbrida de alto nível (ATLA) que integra sinergicamente o poder de abstração de características de algoritmos de aprendizado profundo (como CNNs-BiLSTM) e algoritmos tradicionais da literatura como Support vector machine - Máquina de vetores de suporte (SVM) e Análise Discriminante Linear (LDA), com a capacidade de análise estrutural das RCs (SILVA; ZHAO, 2012). O ATLA visa superar as limitações de modelos baseados exclusivamente em características espectrais tradicionais, aprimorando significativamente o desempenho na classificação de espectros ATR-FTIR em cenários com alta sobreposição entre classes. Um dos principais desafios é a criação das redes de grafo utilizando os espectros ATR-FTIR, para isso, foi utilizado métodos de criação de rede K-Vizinhos mais próximos, *K-nearest neighboors* (kNN) e outros métodos de construção como skNN e smkNN (CARNEIRO, 2017).

Como principal contribuição, este método híbrido foi aplicado e validado extensivamente em múltiplas bases de dados espectrais biomédicas, representando diversos e de-

safiadores contextos clínicos e infecciosos, incluindo Hepatite Delta (HDV), COVID-19 e doença de Chagas. Além disso, a proposta utiliza a aplicação de técnicas de pré-processamento espectral, como correção de *baseline* e normalização dos espectros, para melhorar o aprendizado dos espectros. Este trabalho contribui para o avanço das técnicas de análise espectral e de aprendizado de máquina, e implementa uma abordagem interpretável utilizando SHAP, permitindo a análise do impacto de cada atributo espectral e topológico nas decisões dos classificadores. Esta interpretabilidade apresenta *insights* clínicos valiosos, aumentando a transparência e a aplicabilidade prática das metodologias computacionais voltadas ao diagnóstico assistido por espectroscopia, sendo uma condição fundamental para aceitação das soluções por especialistas de outras áreas.

1.1 Motivação

A demanda por ferramentas diagnósticas rápidas, precisas e tem sido acentuada pelos desafios de saúde global, como a recente pandemia de COVID-19 (ALYASSERI et al., 2022). Nesse cenário, a espectroscopia de infravermelho de biofluídos emergiu como um campo promissor para o diagnóstico de diversas doenças e transtornos (RALBOVSKY; LEDNEV, 2020). Esta técnica é particularmente valorizada por sua natureza não invasiva, baixo custo e ausência de dor na coleta de amostras, oferecendo vantagens significativas sobre os métodos diagnósticos tradicionais (MALAMUD, 2011; BUNACIU; HOANG; ABOUL-ENEIN, 2015). A espectroscopia ATR-FTIR, em particular, destaca-se como uma solução rápida, livre de reagentes e com alta reprodutibilidade (SANTOS; MORAIS; LIMA, 2020). Notavelmente, trabalhos gerados dentro da Universidade Federal de Uberlândia obtiveram destaque, sendo premiados pelo Google em 2022. [1]

A eficácia do ATR-FTIR salivar tem sido amplamente demonstrada no contexto de diversas patologias, além da COVID-19 (RALBOVSKY; LEDNEV, 2020). Incluem-se aqui outras infecções virais como Zika (OLIVEIRA et al., 2023) e HIV (SILVA et al., 2020), e condições oncológicas como câncer cerebral (BUTLER et al., 2019), câncer de pulmão (PENG et al., 2022) e câncer de mama (SITNIKOVA et al., 2020). Para esta dissertação, um dos banco de dados classificados, foi construído a partir de amostras de saliva coletadas de pacientes com COVID-19 e um grupo controle. A principal dificuldade reside na diferenciação sutil das classes, dada a alta semelhança dos espectros gerados, o que torna desafiador separar pacientes infectados de indivíduos saudáveis apenas com base na análise visual do espectro (ALYASSERI et al., 2022). Regiões específicas, como a da Amida I e II, podem indicar diferenças, auxiliando algoritmos de reconhecimento de padrões a otimizar o desempenho preditivo (BYLER; SUSI, 1986).

[1] www.comunica.ufu.ufu.br/noticia/2022/02/pesquisadores-da-ufu-ganham-premio-do-google-pela-segunda-vez

O espectro ATR-FTIR é caracterizado por *wavenumbers* e suas intensidades correspondentes (Figura 1), onde podemos ver a média dos espectros da base de dados, sendo o espectro vermelho representando o grupo alvo (pessoas infectadas com COVID-19) e o espectro azul representando o grupo controle (pessoas sintomáticas não diagnosticadas com COVID-19). Para otimizar a análise e mitigar ruídos ou influências de concentrações de água, é comum truncar o *wavenumber* em regiões específicas, como entre $1800\,\mathrm{cm}^{-1}$ e $900\,\mathrm{cm}^{-1}$ (MARTINEZ-CUAZITL et al., 2021). Abordagens similares de truncamento em outras regiões, como a região de lipídeos entre $3050 - 2800\,\mathrm{cm}^{-1}$, têm sido exploradas em estudos prévios (CAIXETA et al., 2023). Na presente dissertação, essa ideia será aproveitada para otimizar a seleção das regiões de *wavenumber* no modelo que utilizará RCs. Além disso, a correção de *baseline* e a normalização espectral são etapas cruciais de pré-processamento (MORAIS et al., 2019). A aplicação de técnicas como o filtro Savitzky-Golay e a normalização baseada no pico Amida I ($1660\,\mathrm{cm}^{-1}$) são essenciais para atenuar ruídos instrumentais, compensar variações de concentração de proteínas e otimizar os espectros, permitindo que os algoritmos de *machine learning* extraiam padrões de forma mais eficaz (BYLER; SUSI, 1986). Tal desafio pode ser observado na Figura 1, onde podemos observar a média dos espectros da base de dados de COVID-19, onde os espectros geralmente são próximos uns dos outros. Nesse cenário, as RCs demonstram um grande potencial para classificação, ao permitir observar a topologia e características não físicas dos dados, o que pode aprimorar significativamente os resultados de classificação (CARNEIRO; ZHAO, 2018). Diferentemente de métodos que analisam apenas *features* diretamente extraídas, as RCs conseguem capturar padrões intrínsecos e relacionamentos estruturais que são cruciais, especialmente em casos onde os padrões espectrais são sutis e sobrepostos (CARNEIRO; ZHAO, 2018).

O método proposto nesta dissertação, uma **abordagem híbrida** que integra RCs com técnicas de aprendizado profundo (como CNN-BiLSTM) e algoritmos convencionais (como SVM e LDA), visa combinar o melhor de ambos os mundos. Por exemplo, enquanto os modelos de aprendizado profundo são robustos na extração de características locais e dependências temporais dos espectros, as RCs são especializadas em modelar as relações estruturais globais entre as amostras. Essa sinergia permite que o algoritmo de classificação discirna com maior precisão as classes, superando as limitações que cada abordagem teria isoladamente. A variedade de métodos de construção e medidas de rede complexa oferece diversas opções para configurar o modelo e extrair informações topológicas relevantes.

1.2 Objetivos e Desafios da Pesquisa

O objetivo principal desta dissertação é **desenvolver, implementar e validar** o ATLA (ATR-FTIR Topological Learning Analysis), um método inovador de classificação

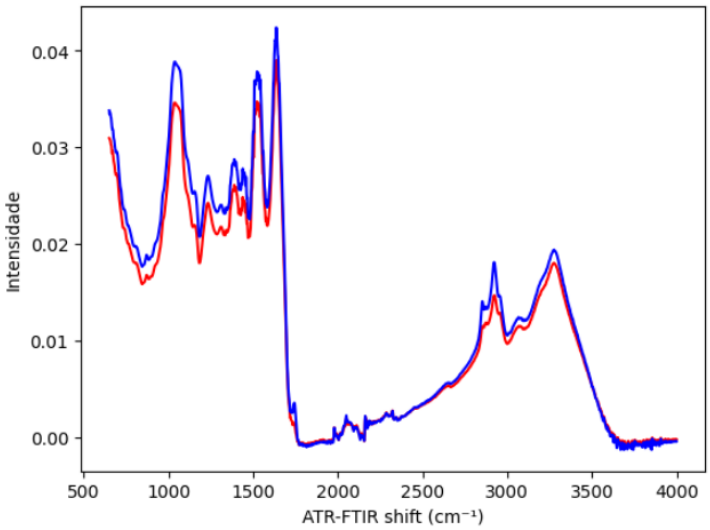


Figura 1 – Espectro ATR-FTIR mostrando a média dos espectros presentes na base de COVID-19

espectral híbrida que combina representações topológicas derivadas de RCs com algoritmos de aprendizado supervisionado. O ATLA visa aprimorar significativamente o desempenho na classificação de dados espectrais ATR-FTIR aplicados a diversos contextos biomédicos utilizando as RCs, de modo a superar as limitações de abordagens tradicionais e monolíticas na identificação de padrões espectrais.

Os objetivos específicos para alcançar este propósito são:

- **Modelar Espectros ATR-FTIR com Redes Complexas e Extrair Atributos Topológicos:** Converter os espectros pré-processados em representações de RCs utilizando o método kNN, com variações como SkNN e mKNN. Extrair um conjunto diversificado de atributos topológicos (grau médio, comprimento médio do caminho, assortatividade, coeficiente de agrupamento, centralidades) que capturem as propriedades estruturais dos sinais espectrais, sendo utilizadas para classificação das redes.

- **Propor e Validar o Algoritmo Híbrido ATLA:** Desenvolver e validar o ATLA, um algoritmo híbrido que combina as probabilidades de classificação obtidas a partir dos atributos espectrais e topológicos com algoritmos base. O ATLA visa explorar a sinergia entre as características locais extraídas por modelos de aprendizado profundo e as relações estruturais globais modeladas pelas RCs.

- **Experimentação do ATLA com classificadores de aprendizado supervisionado:** Implementar e avaliar modelos de classificação supervisionada, incluindo tanto abordagens tradicionais na área de ATR-FTIR, modelos como SVM, LDA, quanto de aprendizado profundo, CNN-1D, BiLSTM, entre outros. Estes algoritmos serão utilizados com o ATLA para o treinamento e validação do mesmo utilizando-

os de forma híbrida com RCs. Onde, os algoritmos serão comparados individuais em um extenso conjunto de simulações.

- ❑ **Interpretar os Resultados com Análise SHAP e Avaliar o Desempenho**: Avaliar quantitativamente o desempenho do ATLA e dos modelos de base utilizando métricas como acurácia, sensibilidade, especificidade e F1-score. Empregar a análise de interpretabilidade SHAP para quantificar a importância de cada atributo (espectral e topológico) nas decisões de classificação, buscando insights clínicos e aumentando a transparência do modelo.

1.2.1 Hipótese

A hipótese de pesquisa investigada nessa dissertação afirma que a integração sinérgica de atributos topológicos extraídos de representações em RCs com classificadores de aprendizado supervisionado (incluindo modelos tradicionais como SVM e LDA) e Deep Learning (CNN-1D BiLSTM) resulta em uma abordagem híbrida capaz de melhorar significativamente o desempenho da classificação de dados espectrais ATR-FTIR. Essa melhoria será investigada em diversas bases de dados biomédicas desafiadoras, englobando doenças infecciosas como Hepatite Delta (HDV), COVID-19 e doença de Chagas quando comparada a modelos de tradicionais de aprendizado profundo.

1.2.2 Questões de Pesquisa

Para investigar e comprovar a hipótese central desta dissertação, as seguintes perguntas de pesquisa foram formuladas e serão respondidas ao longo dos capítulos subsequentes:

- ❑ O método, ATLA, é capaz de melhorar significativamente o desempenho da classificação de dados espectrais ATR-FTIR em comparação com abordagens de classificação tradicionais ou monolíticas (SVM, LDA, CNN), especialmente em cenários com alta similaridade espectral entre as classes? A integração sinérgica de atributos topológicos de RCs com classificadores de aprendizado profundo (CNN-BiLSTM) e tradicional (SVM, LDA) é o fator chave para o aprimoramento do desempenho do ATLA em múltiplas bases de dados biomédicas desafiadoras?

- ❑ O desempenho aprimorado do ATLA é generalizável e robusto em diversas bases de dados biomédicas, como COVID-19, Hepatite Delta (HDV) e doença de Chagas?

- ❑ A análise de interpretabilidade do SHAP, pode fornecer insights valiosos sobre as bandas de infravermelho e atributos topológicos mais relevantes para as decisões de classificação do ATLA, aumentando a transparência e a aplicabilidade prática do modelo?

1.3 Contribuições

Esta dissertação apresenta um conjunto relevante de contribuições científicas e técnicas no campo da análise espectral aplicada à saúde, com ênfase no uso de espectroscopia ATR-FTIR e no desenvolvimento de modelos computacionais híbridos. As principais contribuições podem ser destacadas da seguinte forma:

1. **Desenvolvimento e Validação do ATLA:** Proposição e implementação de um método inovador de classificação espectral híbrida de alto nível, baseado na integração sinérgica de representações topológicas de RCs com algoritmos de aprendizado supervisionado (como SVM e LDA) e de aprendizado profundo (CNN-BiLSTM). O ATLA eleva a capacidade discriminativa na análise de espectros ATR-FTIR ao combinar os pontos fortes de cada paradigma, superando as limitações de modelos monolíticos em capturar informações globais e locais.

2. **Validação Extensiva em Múltiplas Bases de Dados Biomédicas:** Aplicação e validação do ATLA em diversas bases de dados espectrais ATR-FTIR, contemplando condições clínicas e infecciosas desafiadoras de alta relevância: COVID-19, Hepatite Delta (HDV) e doença de Chagas. A análise comparativa entre essas bases demonstrou a robustez e a capacidade de generalização do ATLA em diferentes domínios biológicos e cenários de alta similaridade espectral.

3. **Protocolo Otimizado de Pré-processamento Espectral:** Investigação e aplicação de um protocolo robusto e adaptável de pré-processamento espectral, que inclui filtro Savitzky-Golay (para suavização e diferenciação), múltiplos métodos de correção de *baseline* (como poly, als, arPLS, drPLS, rubberband) e normalização pela Amida I. Este protocolo foi fundamental para otimizar a qualidade dos sinais espectrais, resultando em dados mais limpos e padrões mais evidentes para os classificadores.

4. **Implementação de Abordagem Interpretável com SHAP:** Desenvolvimento e aplicação de uma metodologia de interpretabilidade utilizando SHAP, permitindo a análise do impacto de cada atributo, espectral e topológico, nas decisões dos classificadores. Essa etapa trouxe maior transparência aos modelos e possibilitou a obtenção de *insights* clínicos valiosos sobre as regiões espectrais e atributos mais relevantes para a diferenciação entre classes, aumentando a confiança e a aplicabilidade prática em diagnósticos assistidos por IA.

5. **Potencial para Inovação Tecnológica e Contribuição Multidisciplinar:** O trabalho reforça o caráter inovador da proposta e seu potencial para aplicação prática em diagnósticos assistidos por espectroscopia. Além disso, contribui significativamente para o conhecimento multidisciplinar nas áreas de espectroscopia, ciência

de dados e inteligência artificial aplicada à saúde, unindo conceitos da medicina, estatística, aprendizado de máquina e RCs em uma abordagem validada empiricamente.

1.4 Organização da Dissertação

Esta dissertação está organizada em cinco capítulos, seguidos pelas referências, apêndices e anexos. A estrutura a seguir detalha o conteúdo de cada seção, provendo um guia para a compreensão do trabalho:

- **Capítulo 1 – Introdução:** Este capítulo contextualiza a problemática da pesquisa, destacando a urgência por métodos diagnósticos rápidos e sustentáveis. Apresenta os objetivos gerais e específicos do trabalho, a hipótese de pesquisa, a motivação por trás do estudo – com foco na espectroscopia ATR-FTIR e nos desafios de classificação de biofluídos –, e as principais contribuições desta dissertação para a área.

- **Capítulo 2 – Fundamentação Teórica:** Este capítulo aborda os conceitos teóricos fundamentais para o desenvolvimento da pesquisa. Serão detalhados os princípios da espectroscopia ATR-FTIR, e detalhando as principais etapas de pré-processamento. A teoria de RCs, incluindo suas medidas e métodos de construção de redes, e as bases de algoritmos de aprendizado profundo, como as Redes Neurais Convolucionais (CNNs) e algoritmos clássicos da literatura como SVM e LDA. Adicionalmente, será apresentada uma revisão de trabalhos relacionados que utilizam essas técnicas para diagnóstico em saúde e também uma abordagem do algoritmo híbrido utilizando RCs.

- **Capítulo 3 – Proposta:** Neste capítulo, é apresentada a proposta de um método híbrido inovador para a classificação espectral, detalhando como a integração de representações topológicas de RCs com algoritmos de aprendizado supervisionado (de baixo e alto nível) visa aprimorar o desempenho diagnóstico. Foi desenvolvido o método híbrido ATLA utilizando a classificação de redes complexas e algoritmos bases focados em espectroscopia infravermelha, foi testado em diversas bases de dados médicas reais ATR-FTIR. No capítulo, a metodologia adotada para o desenvolvimento, implementação e validação do método é descrita em profundidade, incluindo as abordagens de pré-processamento, construção de redes, seleção de características, e o método híbrido da ATLA.

- **Capítulo 4 – Experimentos e Análise dos Resultados:** Este capítulo descreve o ambiente experimental e apresenta os resultados obtidos com a aplicação do método híbrido. Detalha as bases de dados biomédicas utilizadas (COVID-19, Chagas e HDV), os métodos de avaliação de desempenho (como acurácia, precisão,

sensibilidade, especificidade e F1-Score), e a análise dos resultados comparativos com as linhas de base. Será explorada a interpretabilidade dos modelos utilizando a metodologia SHAP para identificar as *features* mais importantes.

- **Capítulo 5 – Conclusão:** O último capítulo da dissertação resume as principais descobertas e discute como os objetivos de pesquisa foram alcançados. Destaca as contribuições científicas e técnicas do trabalho, avaliando a hipótese proposta com base nos resultados experimentais. Por fim, são apresentadas sugestões para trabalhos futuros e as contribuições em produção bibliográfica decorrentes desta pesquisa.

Capítulo 2

Fundamentação Teórica e Literatura

Neste capítulo, será estabelecida a fundamentação teórica para o desenvolvimento da dissertação, abordando os conceitos essenciais e revisando a literatura pertinente. Primeiramente, detalhar-se-ão os princípios da espectroscopia por infravermelho com transformada de Fourier por refletância total atenuada ATR-FTIR e a importância de suas etapas de pré-processamento, incluindo correção de *baseline* e suavização para otimização dos espectros. Em seguida, serão apresentados os fundamentos do Aprendizado de Máquina e do Aprendizado Profundo. A teoria de Redes Complexas, com sua construção, medidas topológicas e aplicação em classificação, será amplamente discutida. Serão também explicados os algoritmos de classificação base que compõem o ATLA – Redes neurais convolucionais (CNN), Máquinas de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA) e explicações sobre o papel da interpretabilidade via SHAP. Por fim, esta seção abordará trabalhos relacionados que exploram a aplicação da espectroscopia ATR-FTIR em conjunto com aprendizado de máquina para diagnósticos, incluindo uma revisão sobre os métodos híbridos de classificação.

2.1 Espectroscopia no Infravermelho por Transformada de Fourier

É uma técnica analítica usada para identificar e quantificar diferentes tipos de ligações químicas em uma amostra. A espectroscopia FTIR é baseada na absorção da radiação infravermelha pela amostra, e as diferentes ligações químicas presentes na amostra absorvem energia em frequências específicas (MOVASAGHI; REHMAN; REHMAN, 2008).

Essa técnica é amplamente utilizada em diversas áreas, como química, farmácia, ciência dos materiais e indústria, para análise e caracterização de substâncias. A espectroscopia FTIR permite identificar grupos funcionais, compostos orgânicos e inorgânicos, polímeros, entre outros (MANSUR et al., 2008; SCHMITT; FLEMMING, 1998; MADEJOVÁ, 2003; MOVASAGHI; REHMAN; REHMAN, 2008), auxiliando na determinação

da composição química de uma amostra.

A Reflectância Total Atenuada (Reflectância Total Atenuada (ATR)) é um método de amostragem que introduz luz em uma amostra para adquirir informações estruturais e de composição. A ATR é uma das tecnologias de amostragem mais utilizadas para a Espectroscopia FTIR. A razão para a utilização onipresente da ATR é que permite a análise de amostras sólidas e líquidas de forma limpa, o que simplifica a medição de praticamente todas as substâncias. Se uma amostra de fluido ou pasta fluida estiver em contato com a superfície de ATR, o espectro infravermelho da porção líquida será registrado. É este fenômeno que torna a ATR tão poderosa para o monitoramento de reação química (KAZARIAN; CHAN, 2013).

A utilização da espectroscopia ATR-FTIR com amostras biológicas tem sido utilizada na triagem diagnóstica 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 (BOZKURT; SEVERCAN; SEVERCAN, 2010; SALA et al., 2020; MARTINEZ-CUAZITL et al., 2021; NOGUEIRA et al., 2022). Na Figura 2 é mostrado um exemplo de coleta de saliva e como ela é transformada no espectro que utilizaremos para aplicação dos algoritmos para aplicação dos algoritmos de classificação com finalidade diagnóstica.

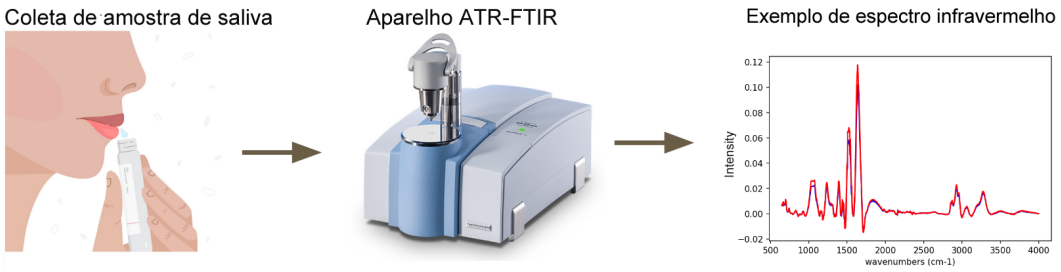


Figura 2 – Como as amostras foram coletadas utilizando o espectrômetro ATR-FTIR. Fonte: Bruker.

Estudos relacionados à espectroscopia de infravermelho de bio-fluídos têm chamado a atenção da comunidade científica (RALBOVSKY; LEDNEV, 2020). Tais estudos apresentam resultados promissores para o diagnóstico de várias doenças e transtornos como COVID-19 (ZHANG et al., 2021), 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 destaca-se por ser um

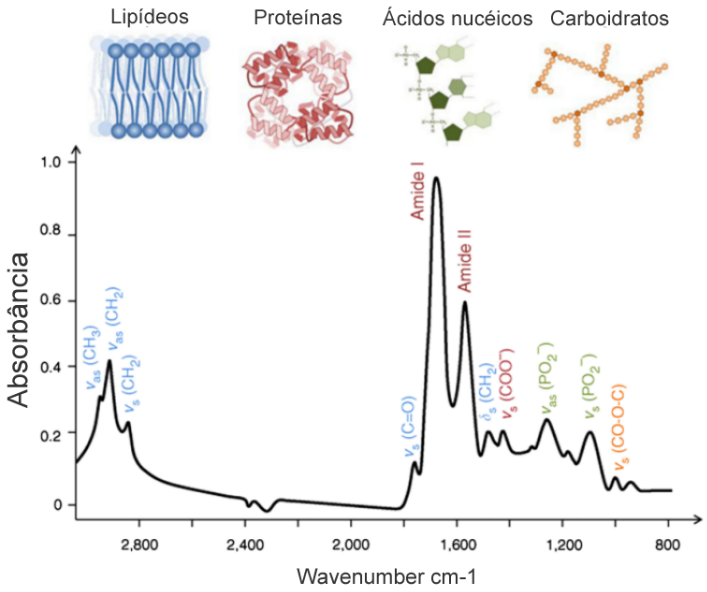


Figura 3 – Espectros FTIR típicos de biomoléculas em biofluidos em número de onda de 3.000˘800cm^1 Fonte: (ROHMAN et al., 2019)

método rápido e sustentável (BUNACIU; HOANG; ABOUL-ENEIN, 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). Na Figura 3 podemos observar o quanto de informações uma amostra de biofluidos é capaz de armazenar, as proteínas na região da Amida I e Amida II são bem importantes pois carregam informações relevantes sobre proteínas que auxiliam algoritmos em classificações (ROHMAN et al., 2019). 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.

A seguir, serão apresentadas as quatro principais etapas para tratamento dos espectros ATR-FTIR, a saber: correção de *baseline*, truncamento do espectro e normalização do espectro pela região da Amida I. Essas etapas são cruciais para otimizar a qualidade dos dados e maximizar o desempenho dos modelos de classificação.

2.1.1 Correção de *Baseline*

A análise de espectros ATR-FTIR, especialmente em contextos biológicos e clínicos, apresenta desafios intrínsecos relacionados à complexidade das amostras e às condições de aquisição. Variações indesejadas na linha de base espectral, decorrentes de fatores como dispersão da luz, impurezas na amostra, absorção por água residual, ou heterogeneidades

no contato da amostra com o cristal do ATR, podem distorcer o sinal de interesse e comprometer a acurácia das análises quantitativas e qualitativas (BAKER et al., 2014).

A correção de *baseline*, também denominada subtração de *baseline*, é uma etapa fundamental 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 prejudicar o aprendizado e a detecção de padrões nas etapas de análise subsequentes (PENG et al., 2011). Para a aplicação clínica da espectroscopia ATR-FTIR, faz-se necessária uma validação em estudos de larga escala e uma evolução em algoritmos que permitam reduzir as diferenças na aquisição do espectro infravermelho por diferentes operadores e equipamentos. Neste sentido, o desenvolvimento de algoritmos que permitam a correção da linha de base (baseline) do espectro é de fundamental relevância para garantir a reprodutibilidade e a confiabilidade dos dados (MORAIS et al., 2019).

A aplicação da correção de *baseline* permite uma otimização da análise, mantendo o perfil espectral autêntico da amostra e deve ser utilizada de forma consistente em todas as amostras para garantir a comparabilidade. Este processo 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. 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; EILERS; HANKEMEIER, 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 4 a), apresenta as *baselines* obtidas pelos métodos de correção para o espectro médio da base de dados, e a Figura 4 b), ilustra 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.

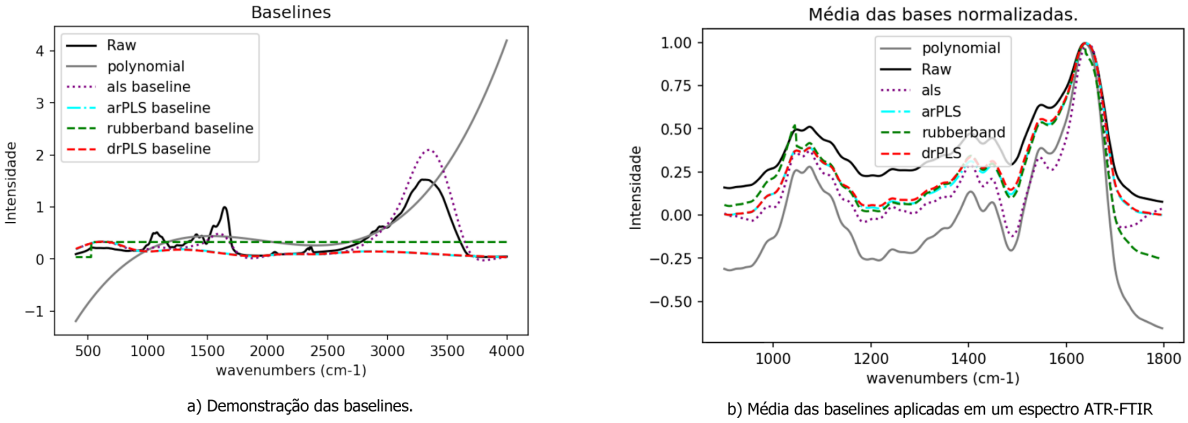


Figura 4 – Baselines demonstradas e aplicadas em espectros ATR-FTIR

2.1.2 Truncamento do Espectro

O truncamento do espectro é uma etapa vital no pré-processamento, onde a faixa espectral correspondente a cada amostra é selecionada para análise, enquanto as demais regiões são desconsideradas. Neste estudo, o espectro é truncado nas regiões entre $900\,\mathrm{cm}^{-1}$ e $1800\,\mathrm{cm}^{-1}$ e entre $2800\,\mathrm{cm}^{-1}$ e $3050\,\mathrm{cm}^{-1}$. O objetivo principal desse procedimento é evitar regiões de elevado ruído ou que contenham interferências significativas (como as de água), que poderiam dificultar a capacidade de generalização dos modelos de aprendizado (MARTINEZ-CUAZITL et al., 2021; CAIXETA et al., 2023). A relevância dessas regiões para a classificação será investigada e confirmada por meio da análise de interpretabilidade utilizando o método SHAP, explicado no final do capítulo. A escolha dessas faixas espectrais é estratégica, pois elas contêm bandas importantes, como a Amida I, que são cruciais para a classificação e para o aprendizado eficaz dos algoritmos.

2.1.3 Suavização e Diferenciação com Filtro Savitzky-Golay

O filtro Savitzky-Golay é empregado para suavizar o sinal e, neste estudo, para calcular a *segunda derivada* do espectro, o que é útil para realçar pequenos picos e remover tendências de linha de base de baixa frequência (PRESS; TEUKOLSKY, 1990). É observado na Figura 5 a suavização do espectro como resultado após a aplicação do filtro.

A seguir é mostrado o algoritmo da suavização e explicado como ele funciona e como seus parâmetros são escolhidos.

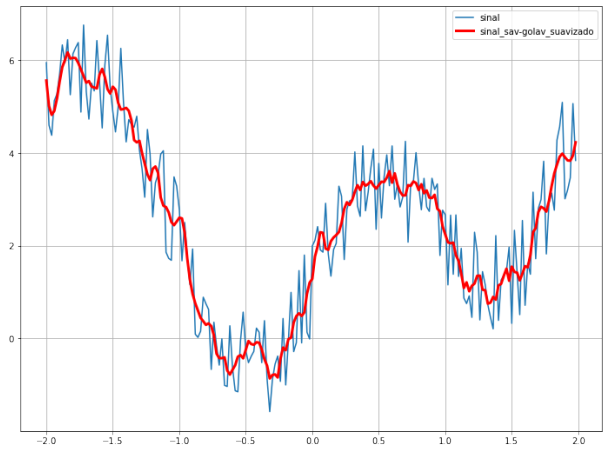


Figura 5 – Aplicação do filtro Savitzky-golay em um espectro ATR-FTIR

Algoritmo 1 Função de Suavização, Diferenciação e Normalização L2 do Espectro

Input: `espectro`: Matriz de espectros brutos (n_amostras × n_wavenumbers)
Output: `espectro_processado`: Matriz de espectros suavizados e diferenciados, normalizados pela norma L2

Função `processar_espectro(`*espectro*`):`
  `espectro_suavizado_diferenciado` ← aplicar_savgol_filter(`espectro`, `tamanho_janela`=25, `ordem_polinomial`=4, `ordem_derivada`=2, `eixo`=1)
  `norma_l2` ← calcular_norma_L2(`espectro_suavizado_diferenciado`, `eixo`=1)
  `espectro_processado` ←
  `espectro_suavizado_diferenciado` / `norma_l2`
  return *espectro_processado*
fim

No pseudocódigo apresentado, a função **aplicar savgol filter** utiliza os seguintes parâmetros:

- **tamanho_janela = 25:** Representa o número de pontos incluídos na janela deslizante usada para o ajuste polinomial. Um valor de 25 pontos é um número ímpar, o que é um requisito para o filtro. A escolha desse tamanho influencia o grau de suavização; janelas maiores resultam em maior suavização.

- **ordem_polinomial = 4:** Indica o grau do polinômio que é ajustado aos dados dentro de cada janela. Uma ordem de 4 permite um bom ajuste à curvatura do sinal sem capturar ruídos de alta frequência.

- **ordem_derivada = 2:** Este é um parâmetro crucial. Ele especifica que o filtro não está apenas suavizando, mas também calculando a **segunda derivada** do espectro. A segunda derivada é particularmente eficaz para resolver picos sobrepostos, remover tendências de linha de base e realçar características espectrais sutis que podem ser importantes para a classificação (SAVITZKY; GOLAY, 1964).

- **eixo = 1:** Indica que a operação deve ser aplicada ao longo do segundo eixo da matriz de entrada (normalmente as colunas, que representam os wavenumbers de cada amostra).

A normalização subsequente pela norma L2 (`np.linalg.norm`) é uma etapa comum para padronizar a magnitude dos espectros após o processamento, tornando-os comparáveis independentemente da concentração total ou do tamanho da amostra.

2.1.4 Normalização pelo Pico da Amida I

Além da normalização pela aplicação do filtro savitzky-golay, uma normalização específica pelo pico da Amida I é aplicada para compensar variações na concentração de proteínas entre as amostras, um fator comum em biofluídos. A Amida I é uma banda de absorção proeminente em espectros de proteínas, localizada principalmente na região de $1630\,\mathrm{cm}^{-1}$ a $1660\,\mathrm{cm}^{-1}$. A normalização por este pico visa isolar as mudanças na estrutura das proteínas, que podem ser indicativas de estados de doença.

Algoritmo 2 Função de Normalização pelo Pico da Amida I

Input: `espectro`: Matriz de espectros (n_amostras × n_wavenumbers)
`indices_amida1`: Lista de índices de *wavenumbers* correspondentes à região da Amida I
Output: `espectro_normalizado`: Matriz de espectros normalizados pela Amida I

Função *1espectro, indices_amida1*:
    `valores_amida1` ← selecionar_valores(`espectro`, `indices_amida1`)
    `pico_max_amida1` ← encontrar_maximo(`valores_amida1`, `eixo`=1)
    `espectro_normalizado` ← `espectro` / `pico_max_amida1`
    return *espectro_normalizado*
fim

Esta função opera identificando o valor de intensidade máxima na região da Amida I para cada espectro individual. Em seguida, cada espectro é dividido por esse valor máximo específico, escalando todas as suas intensidades de forma que o pico da Amida I daquela amostra se torne 1. Esse método é eficaz para minimizar as variações na concentração total da amostra, permitindo que as diferenças na composição e estrutura molecular (que são mais relevantes para o diagnóstico) se tornem mais evidentes.

Após preparados e pré-processados, as principais abordagens da literatura fazem uso de aprendizado de máquina para classificar de maneira mais efetiva as assinaturas espectrais, utilizando algoritmos como LDA, SVM e também algoritmos como CNN-1D, Naive bayes e estatísticos como PCA (TEKLEMARIAM et al., 2024; DELRUE; BRUYNE; SPEECKAERT, 2025; OLIVEIRA et al., 2023; BUTLER et al., 2019; HONÓRIO-SILVA et al., 2024; GARCIA-JUNIOR et al., 2025).

2.2 Aprendizado de Máquina e Aprendizado Profundo

No início da era computacional, algoritmos eram criados para a execução de tarefas bem definidas. No entanto, certas tarefas complexas não possuem um algoritmo explícito que possa ser facilmente programado. Exemplos notáveis incluem reconhecimento facial, filtragem de e-mails legítimos versus *spam*, e a classificação de imagens (ALPAYDIN, 2020).

O Aprendizado de Máquina (AM) é uma subárea da Inteligência Artificial que se dedica ao desenvolvimento de técnicas computacionais capazes de aprender e adquirir conhecimento a partir de dados de forma automática. Essencialmente, um sistema de aprendizado é um programa de computador que aprimora seu desempenho em uma tarefa específica por meio da experiência, baseado na solução bem-sucedida de problemas anteriores ou na identificação de padrões em grandes volumes de dados (MONARD; BARANAUSKAS, 2003).

Existem alguns paradigmas para aprendizado de máquina quanto a classificação de dados envolvendo a sua fonte e o quão contextualizados estão. O algoritmo de aprendizado é fornecido com um conjunto de exemplos de treinamento para os quais os "rótulos" da classe associada são conhecidos. Cada exemplo é tipicamente descrito por um vetor de valores de características (ou atributos), e o objetivo do algoritmo é construir um modelo (classificador ou regressor) que possa prever corretamente o rótulo de novos exemplos ainda não rotulados. Para rótulos de classe discretos, esse problema é conhecido como **classificação** (MONARD; BARANAUSKAS, 2003), sendo o tipo de aprendizado predominante nesta dissertação. O processo envolve o treinamento do modelo com dados rotulados, permitindo que ele aprenda a mapear características de entrada para saídas desejadas (MITCHELL; MITCHELL, 1997). Temos a aprendizagem supervisionada, onde, geralmente um especialista indica qual classe é cada amostra e com isso novas amostras de entrada são classificadas com conhecimento das anteriores. A semi-supervisionada é indicado algumas amostras de uma classe específica e as mais próximas se adaptam à ela, como podemos ver na Figura 6.

No aprendizado não supervisionado, os dados de entrada do modelo são descritos exclusivamente por suas características, sem rótulos associados. O próprio modelo analisa as informações da base de dados para identificar estruturas, padrões ou agrupar os dados em grupos (clusters) de acordo com algum critério de similaridade (MONARD; BARANAUSKAS, 2003). Na Figura 6 este paradigma é frequentemente empregado para exploração de dados, redução de dimensionalidade ou segmentação.

O aprendizado por reforço é um paradigma que envolve o treinamento de modelos de AM para tomar uma sequência de decisões em um ambiente. Um "agente" aprende a atingir uma meta por meio de interações, recebendo "recompensas" ou "penalidades" por suas ações (SUTTON; BARTO et al., 1998). O objetivo é que o agente aprenda a política ótima para maximizar a recompensa acumulada ao longo do tempo, podemos ver um

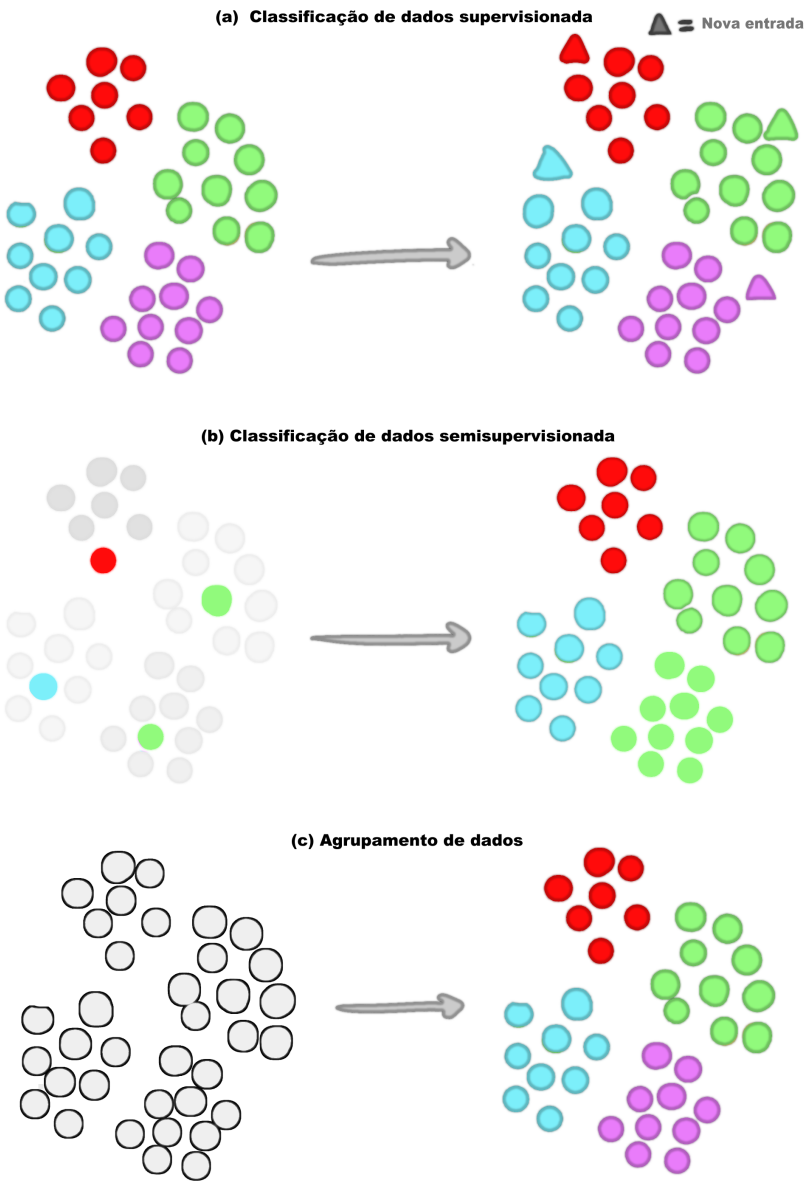


Figura 6 – Ilustração da diferença básica entre os principais paradigmas do aprendizado de máquina. Adaptado de (CARNEIRO, 2017)

exemplo na Figura 7. É amplamente utilizado em áreas como jogos, robótica e aplicações envolvendo o aprendizado de agentes autônomos.

O Aprendizado Profundo é uma vertente do Aprendizado de Máquina baseada em Redes Neurais Artificiais (RNAs) com múltiplas camadas de processamento (camadas "profundas") entre a entrada e a saída. Inspiradas na estrutura e função do cérebro humano, as RNAs são modelos computacionais compostos por neurônios interconectados que processam informações. A "profundidade" dessas redes permite que elas aprendam representações de dados em vários níveis de abstração, automaticamente extraindo características complexas e hierárquicas diretamente dos dados brutos, sem a necessidade de engenharia de *features* manual (LECUN; BENGIO; HINTON, 2015).

Essa capacidade de aprendizado hierárquico e automático de características é o que

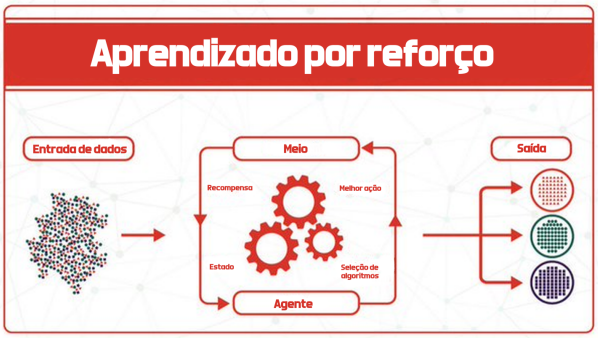


Figura 7 – Funcionamento do aprendizado por reforço. Fonte: Adaptado de Medium

confere ao Aprendizado Profundo sua notável performance em tarefas como reconhecimento de padrões, processamento de linguagem natural e visão computacional. As Redes Neurais Convolucionais (CNNs), por exemplo, são um tipo específico de arquitetura de rede profunda particularmente eficaz para dados com estrutura de grade, como imagens e sequências.

Apesar da notável capacidade do Aprendizado Profundo em extrair características hierárquicas e aprender representações complexas diretamente dos dados (LECUN; BENGIO; HINTON, 2015), modelos de aprendizado profundo tradicionais, como as CNNs, são adeptos na captura de padrões locais e dependências sequenciais, mas frequentemente carecem na representação de relações estruturais globais entre as amostras.

Nesse contexto, a teoria de Redes Complexas emerge como um paradigma complementar e igualmente poderoso. Redes complexas representam sistemas através de grafos com padrões de interconexão não triviais, permitindo a análise de características topológicas e funcionais que revelam propriedades intrínsecas e relacionamentos não lineares nos dados. A combinação dessas duas abordagens configura uma abordagem híbrida com potencial para aprimorar significativamente o desempenho em tarefas de classificação, especialmente em dados biomédicos complexos, como espectros ATR-FTIR. A próxima subseção aprofundará os conceitos de Redes Complexas e sua aplicabilidade.

Nas subseções a seguir, serão apresentados algoritmos de aprendizado de máquina e aprendizado profundo investigados nesta dissertação, sendo eles Rede neural convolucional (CNN), Máquina de Vetores (SVM) e Análise Discriminante Linear (LDA), os quais foram utilizados juntamente com a rede complexa para trabalhar no algoritmo híbrido.

2.2.1 Redes Neurais Convolucionais - CNN

Uma rede neural convolucional (CNN) é um dos algoritmos de aprendizado profundo que revolucionou o processamento de dados em diversas áreas, notadamente na visão computacional. Diferentemente de redes neurais simples, que operam com vetores de atributos como entrada, as CNNs são intrinsecamente projetadas para tratar dados com estruturas espaciais ou sequenciais, como imagens (atributos multidimensionais de altura, largura

e profundidade/canais de cores) (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Sua eficácia reside na capacidade de identificar e diferenciar aspectos complexos dos dados de entrada, realizando mapeamento e transformação espacial.

O poder das CNNs advém de suas características arquiteturais principais:

- **Camadas Convolucionais:** Através de filtros (*kernels*) que deslizam sobre os dados de entrada, as CNNs aprendem automaticamente características hierárquicas, desde padrões de baixo nível (bordas, texturas) até representações de alto nível. Esse processo de "peso compartilhado" (*weight sharing*) permite que um mesmo filtro detecte um padrão específico em diferentes posições dos dados, tornando a rede eficiente e robusta a pequenas variações e deslocamentos (LECUN; BENGIO et al., 1995; GOODFELLOW et al., 2016).

- **Camadas de Pooling:** Estas camadas reduzem a dimensionalidade dos dados (*downsampling*), o que ajuda a tornar o modelo mais robusto a pequenas distorções e a controlar o *overfitting*, além de diminuir a complexidade computacional.

- **Conectividade Local:** Cada neurônio em uma camada convolucional se conecta apenas a uma pequena região da camada anterior, o que imita a forma como o cérebro humano processa informações visuais, focando em características locais.

As CNNs não são restritas apenas ao processamento de imagens. Elas podem ser adaptadas e aplicadas com grande sucesso a dados de uma dimensão (1D), como séries temporais e, pertinentemente para esta dissertação, espectros de 1D (JIANG et al., 2021; KIRANYAZ; INCE; GABBOUJ, 2015). Para trabalhar com espectros de uma dimensão, os dados podem ser tratados como uma série temporal, onde cada ponto representa uma medida de intensidade em uma determinada posição do *wavenumber*. Neste caso, a arquitetura da CNN é modificada para lidar com entradas unidimensionais, utilizando convoluções 1D. Essa adaptação permite que a CNN extraia automaticamente padrões locais e sequenciais relevantes ao longo do espectro, como picos e bandas específicas, sem a necessidade de engenharia de *features* manual, o que é particularmente vantajoso para a análise de dados ATR-FTIR, já contribuindo também estudos com COVID-19 (KROHLING; KROHLING, 2023). A capacidade de aprender representações diretamente dos dados brutos e de ser robusta a variações intrínsecas dos espectros, como ruído ou pequenas variações de linha de base, torna as CNNs 1D uma ferramenta poderosa para a classificação espectral.

2.2.2 Máquinas de Vetores de Suporte - SVM

As Máquinas de Vetores de Suporte (SVMs) são algoritmos de aprendizado supervisionado amplamente utilizados para tarefas de classificação e regressão, destacando-se pela sua robustez e eficácia, especialmente em conjuntos de dados com alta dimensionalidade

ou quando o número de amostras é limitado (CORTES; VAPNIK, 1995). O princípio fundamental de uma SVM para classificação binária é encontrar um hiperplano ótimo que maximize a margem de separação entre as classes em um espaço de características. Essa margem é definida pela distância entre o hiperplano de separação e os pontos de dados mais próximos de cada classe, conhecidos como "vetores de suporte" (BURGES, 1998).

Para a classificação espectral, as SVMs podem ser eficazes na identificação de padrões discriminativos, lidando bem com a alta dimensionalidade dos espectros e a complexidade das relações entre as variáveis.

2.2.3 Análise Discriminante Linear - LDA

A Análise Discriminante Linear (LDA) é um método estatístico supervisionado utilizado tanto para a redução de dimensionalidade quanto para a classificação, com o objetivo principal de encontrar combinações lineares de variáveis que melhor discriminam entre duas ou mais classes de um conjunto de dados (XANTHOPOULOS et al., 2013). Ao contrário de técnicas de redução de dimensionalidade não supervisionadas como a Análise de Componentes Principais (PCA), que busca a variância máxima nos dados, a LDA otimiza a separação entre as classes, maximizando a razão entre a variância interclasse e a variância intraclasse (MARTINEZ; KAK, 2001).

O resultado da LDA são eixos discriminantes que, ao projetarem os dados, criam uma separação máxima entre os centróides das classes e uma variância mínima dentro de cada classe, facilitando a classificação (MARTINEZ; KAK, 2001). Após a projeção em um espaço de menor dimensão, um classificador simples (como um limite direto) pode ser utilizado para atribuir amostras a uma classe específica. A LDA é particularmente útil quando as distribuições de classe seguem uma premissa de normalidade e possuem covariâncias semelhantes. No contexto de dados espectrais, a LDA pode ser empregada para identificar as regiões do espectro que contribuem mais para a diferenciação entre as classes, tornando-a uma ferramenta valiosa para a classificação e análise de dados ATR-FTIR.

2.2.4 Interpretabilidade dos Modelos Com o Método SHAP

SHAP (SHapley Additive exPlanations) (LUNDBERG; LEE, 2017) é uma técnica de interpretação baseada na teoria dos jogos que atribui, de forma justa, a contribuição de cada variável para uma predição de modelo. Utilizando os valores de Shapley, o SHAP calcula o impacto marginal de cada atributo ao ser adicionado a diferentes combinações de variáveis, e depois faz a média dessas contribuições, garantindo uma explicação robusta e aditiva. Os valores SHAP indicam não apenas a importância de cada variável, mas também a direção do seu impacto (positivo ou negativo), permitindo interpretações

locais (por instância) e globais (por todo o conjunto). Em aplicações biomédicas com espectros ATR-FTIR, esta técnica permitirá quantificar a contribuição de cada atributo (tanto espectral quanto topológico) nas decisões dos classificadores, destacando as *features* (intensidade de um determinado número de onda cm^{-1}) mais importantes para a diferenciação entre as classes. Podemos ver um exemplo do SHAP na Figura 8, onde, o elemento que mais impacta para a diferenciação dos espectros é mostrado em primeiro lugar com o maior valor de SHAP. A análise do SHAP contribuirá para uma maior transparência dos modelos e para a validação qualitativa da aplicabilidade da abordagem híbrida em contextos clínicos reais.

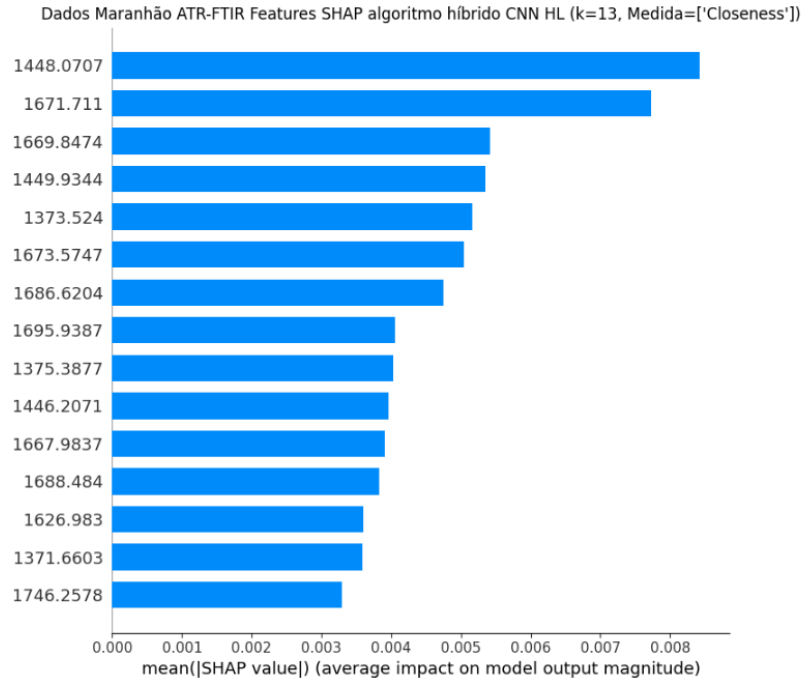


Figura 8 – Exemplo de valores SHAP evidenciando as regiões do espectro ATR-FTIR com maior contribuição para o modelo de classificação. Fonte: Autor

2.2.5 Métricas para Avaliação de Desempenho Preditivo

Uma matriz de confusão é uma tabela usada para definir o desempenho de um algoritmo de classificação, (SINGH et al., 2021) especialmente em problemas supervisionados. Ela resume os resultados das previsões do modelo, comparando os rótulos reais com os rótulos preditos. No caso de uma classificação binária, a matriz é mostrada na Tabela 1 abaixo:

Tabela 1 – Matriz de Confusão para Classificação Binária

	Classe Predita	
Classe Real	Positiva	Negativa
Positiva	VP (Verdadeiro Positivo)	FN (Falso Negativo)
Negativa	FP (Falso Positivo)	VN (Verdadeiro Negativo)

A matriz de confusão é uma ferramenta essencial na avaliação de modelos de classificação, especialmente em cenários binários. Ela organiza os resultados da predição em quatro categorias principais:

- **VP (Verdadeiro Positivo)**: Casos em que a amostra realmente pertence à classe positiva e o modelo previu corretamente.
- **VN (Verdadeiro Negativo)**: Casos em que a amostra realmente pertence à classe negativa e foi corretamente classificada como tal.
- **FP (Falso Positivo)**: Casos em que a amostra pertence à classe negativa, mas foi incorretamente classificada como positiva.
- **FN (Falso Negativo)**: Casos em que a amostra pertence à classe positiva, mas o modelo não a reconheceu, classificando-a como negativa.

A partir dessa matriz, é possível derivar métricas de desempenho importantes como acurácia, precisão, sensibilidade, especificidade e F1-score, que fornecem uma visão mais abrangente sobre o comportamento do classificador.

Para avaliar a eficácia da abordagem híbrida proposta, o desempenho do modelo é avaliado utilizando métricas de classificação padrão, incluindo Acurácia (AC), (Sensibilidade (SE)), Especificidade (ES), Precisão (PR) e F1-score. Essas métricas fornecem uma visão abrangente da capacidade do classificador em identificar corretamente amostras positivas (COVID-19) e negativas (controle), o que é de particular importância em diagnósticos médicos onde o custo de falsos negativos ou falsos positivos pode ser significativo.

- **Acurácia** mede a proporção geral de amostras corretamente classificadas e é definida como:

$$\text{Acurácia} = \frac{VP + VN}{VP + TN + FP + FN} \tag{1}$$

- **Sensibilidade** quantifica a capacidade do modelo de detectar corretamente casos positivos:

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{2}$$

- **Especificidade**mede o quão bem o modelo identifica casos negativos:

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{3}$$

- **Precisão** quantifica a proporção de casos positivos preditos que são, de fato, positivos. É uma medida da qualidade das predições positivas do modelo:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{4}$$

- **F1-score:** Esta métrica representa a média harmônica da precisão (*Precision*) e da sensibilidade (*Recall*). É particularmente útil em problemas de classificação com classes desbalanceadas, pois penaliza modelos que ignoram uma das classes. O F1-score é calculado como:

$$\text{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{5}$$

No contexto de aplicações médicas, essas métricas desempenham um papel crucial para garantir a confiabilidade e segurança dos sistemas diagnósticos. Alta sensibilidade é essencial para minimizar falsos negativos, que poderiam resultar em diagnósticos perdidos e maior transmissão da doença. Alta especificidade ajuda a prevenir falsos positivos, evitando tratamentos desnecessários, ansiedade ou alocação inadequada de recursos. Enquanto a acurácia oferece uma medida global, ela pode ser enganosa em cenários de classes desbalanceadas, que é comum em conjuntos de dados clínicos, destacando a necessidade da sensibilidade, especificidade, precisão e, especialmente, do F1-score como indicadores complementares da eficácia diagnóstica. Utilizaremos a acurácia, precisão, sensibilidade, especificidade e F1-Score nas tabelas de resultados, deixando a F1-Score como a principal medida por conta dela utilizar sensibilidade e precisão.

2.3 Aprendizado de Máquina em Redes Complexas

As redes complexas são baseadas em grafos, e foram popularizadas há décadas atrás e vem ganhando força a cada ano com sua capacidade de classificação. As redes complexas têm como base a complexidade da vida, e se chamam complexas pelo fato de serem nem completamente regulares nem completamente aleatórias (CARNEIRO, 2017). Dado que redes complexas podem descrever a relação entre eventos, mais pesquisas estão usando redes complexas para modelar problemas. Por exemplo, podemos usar uma rede para modelar compostos em pesquisa química, na qual nós e arestas representam respectivamente moléculas e ligações químicas entre moléculas.

Os estudos sobre as redes complexas têm crescido muito nas últimas décadas e têm sido objeto de pesquisa em diversas disciplinas, incluindo física, matemática, ciência da computação, biologia e sociologia (STROGATZ, 2001). Essas redes podem ser definidas como estruturas compostas por elementos interconectados, onde os elementos podem ser nós (ou vértices) e as conexões entre eles são chamadas de arestas. Um dos pontos mais interessantes sobre as redes complexas é que elas não seguem critérios determinísticos nem totalmente aleatórios na sua formação . Em vez disso, há um viés de formação que leva em conta diversos fatores, como por exemplo as redes livres de escala na qual há preferência por conexões com outros elementos que já possuem muitas conexões (CARNEIRO, 2017).

Diferentemente das abordagens tradicionais, que se baseiam em proximidade física ou separabilidade geométrica, a análise via redes complexas oferece uma vantagem funda-

mental ao explorar a topologia subjacente dos dados. Conforme ilustrado pela Figura 9, métodos como k-NN e SVM classificam incorretamente as instâncias de teste (triângulos) como pertencentes à classe dos quadrados vermelhos, pois estes são os vizinhos fisicamente mais próximos. Em contraste, a abordagem topológica, apresentada na Figura 10, primeiro mapeia o conjunto de dados em uma rede (Figura 10 (a)), onde cada ponto de dado se torna um nó e as conexões entre eles revelam o padrão estrutural do conjunto.

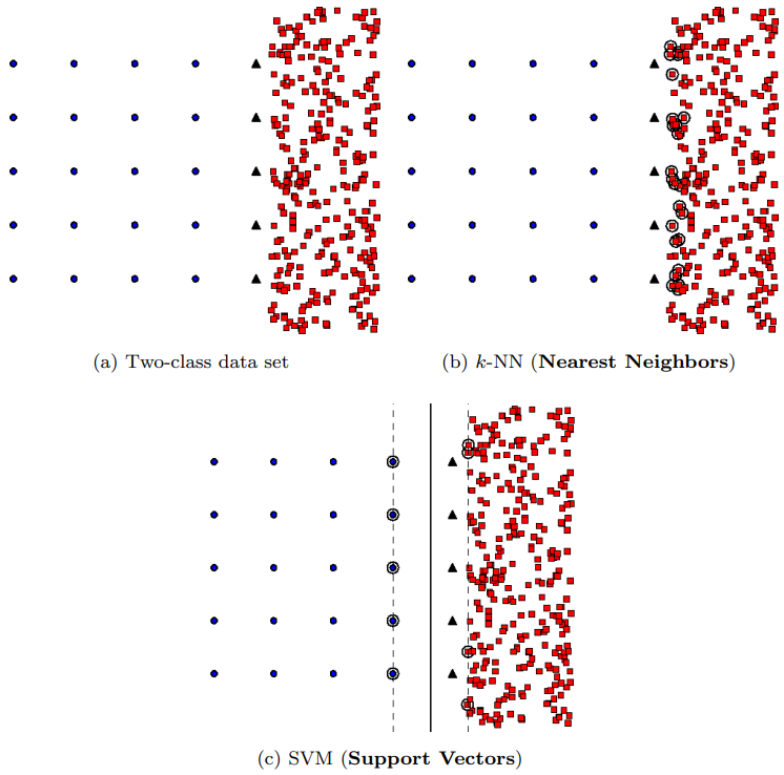
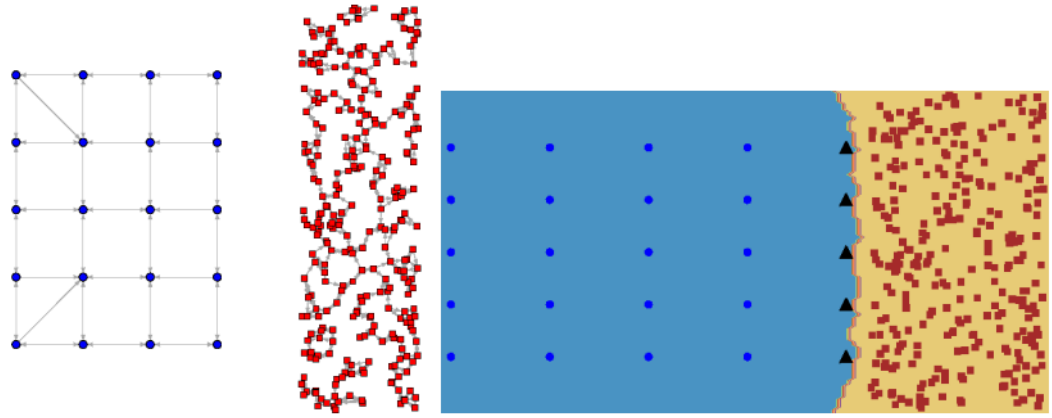


Figura 9 – Base de dados para ilustrar os métodos de classificação kNN, SVM Fonte: (CARNEIRO et al., 2019) Elsevier Licensa: 6112510255119

Ao analisar a rede, a técnica não avalia apenas a distância, mas também a função e a importância de cada nó dentro da estrutura global, olhando sua topologia em geral. Isso permite que o método identifique que os triângulos, apesar de sua proximidade com a classe vermelha, são uma continuação semântica e estrutural do padrão de grade formado pela classe dos círculos azuis. Como resultado (Figura 10b), a classificação baseada em redes complexas consegue capturar essa relação de padrão e agrupar corretamente as instâncias de teste com sua verdadeira classe, demonstrando uma capacidade superior de identificar padrões semânticos que são invisíveis para os métodos de classificação convencionais (CARNEIRO et al., 2019).

Tendo demonstrado visualmente o potencial da abordagem topológica, como podemos, de forma algorítmica, capturar a percepção de que os “triângulos” são uma continuação do padrão de “círculos”? Para que um classificador possa, de fato, “aprender” a topologia e tomar decisões a partir dela, é necessário converter as propriedades estruturais em um formato numérico. É neste ponto que as medidas de redes complexas se tornam



(a) Network formed from (b) Importance-based Classification

Figura 10 – Como a classificação baseada em importância da rede complexa funciona. Fonte: (CARNEIRO et al., 2019) Elsevier Licensa: 6112510255119

essenciais. Em vez de usar as coordenadas originais dos dados, passamos a descrever cada ponto por suas características topológicas: sua centralidade, sua conexão com vizinhos, seu papel como ponte entre grupos, entre outras. A seção a seguir detalhará como essas medidas extraem a inteligência topológica dos dados, permitindo que o classificador tome decisões baseadas na relevância estrutural de cada amostra, e não apenas em sua posição geométrica.

2.3.1 Métodos de Construção da Rede

No contexto de aprendizado de máquina, um dos métodos mais comuns é o grafo kNN *(N Neasrest Neighbors)*. A ideia consiste basicamente em conectar cada item de dado aos seus k vizinhos mais próximos.

Formalmente, seja kNN($\mathbf{x}_i$) os k vizinhos mais próximos de $\mathbf{x}_i$, podemos obter a matriz de adjacência do grafo $\mathbf{A}$ como:

$$A_{ij} = \begin{cases} 1 & \text{se } \mathbf{x_i} \in \text{kNN}(\mathbf{x}_j) \text{ e } y_i = y_j, \\ 0 & \text{senão.} \end{cases}$$

onde, y_i é a classe da amostra $\mathbf{x}_i$, se ela for igual a classe do vizinho $\mathbf{x}_j$, e.g $(y_i = y_j)$ ela será conectada.

Para otimizar a representação topológica, são investigadas variações da criação de redes por kNN, que são bastante utilizadas por tornarem o grafo resultante não direcionado. Isso inclui o **k-Nearest Neighbors Simétrico (SkNN)** e o **k-Nearest Neighbors Mútuo (mKNN)**, que promovem diferentes tipos de conectividade mútua e aprimoram a difusão da informação através do grafo (CARNEIRO et al., 2019; RESENDE; CAR-

NEIRO, 2020). O SkNN gera uma matriz de adjacência simétrica $\mathbf{A}'$ a partir da matriz de adjacência original $\mathbf{A}$ (obtida pelo kNN) da seguinte forma:

$$A'_{ij} = \max(A_{ij}, A^T_{ij}). \tag{6}$$

Outro método derivado é a rede kNN Mútuo (MkNN), onde a matriz de adjacência simétrica $\mathbf{A}'$ é obtida por:

$$A'_{ij} = \min(A_{ij}, A^T_{ij}). \tag{7}$$

Outro método bem conhecido é o vizinhança de raio ϵ (eN) (CARNEIRO; ZHAO, 2018) que basicamente conecta um item de teste a todos os vizinhos que estão até um raio máximo ϵ, obtendo uma melhor representação de regiões mais densas. Formalmente, seja D a matriz de distãncia onde D_{ij} é a distância entre $\mathbf{x}_i$ e $\mathbf{x}_j$, $\mathbf{A}$ pode ser obtida por:

$$A_{ij} = \begin{cases} 1 & \text{se } D_{ij} \leq \epsilon \text{ e } y_i = y_j, \\ 0 & \text{senão.} \end{cases}$$

A classificação de dados em redes complexas é uma tarefa fundamental do aprendizado de máquina que busca predizer rótulos para entidades dentro de uma estrutura de grafo. Essas entidades podem ser nós (*node classification*), arestas (*link prediction*) ou o grafo inteiro (*graph classification*). O campo tem progredido notavelmente, impulsionado pela necessidade de analisar dados relacionais em domínios como redes sociais, biologia, sistemas de recomendação e processamento de linguagem natural.

2.3.2 Medidas de Redes Complexas

- Grau Médio: Corresponde ao número médio de ligações que cada vértice faz no grafo. Pode ser obtido por:

$$GM = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{8}$$

- Comprimento Médio do Caminho: número médio de etapas ao longo dos caminhos mais curtos para todos os pares possíveis de nós de rede (RESENDE; CARNEIRO, 2021).

$$CMC = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j) \tag{9}$$

- Assortatividade (Assortativity): Esta medida determina o quanto os vértices tendem a se conectar de maneira assortativa. A medida pode assumir valores entre $[-1, 1]$, de modo que valores positivos indicam que pares de vértices diretamente conectados estão mais suscetíveis a se comportar da mesma forma, enquanto valores

negativos indicam maior probabilidade dos vértices conectados terem comportamentos distintos (CARNEIRO, 2017). Seja L o número de arestas na rede e i_u, k_u o grau dos vértices i e k que compõem uma aresta u, a assortatividade da rede pode ser calculada por:

$$r = \frac{L^{-1}\sum_u i_u k_u - [L^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}{L^{-1}\sum_u \frac{1}{2}(i_u^2 + k_u^2) - [L^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}$$

- Coeficiente de Agrupamento : De forma geral, esta medida quantifica o quanto os vértices tendem a agrupar-se. O coeficiente de agrupamento de um vértice mede o quão próximo ele está de formar um clique. Esta medida pode ser obtida por:

$$CC_i = \frac{|e_{us}|}{k_i(k_i - 1)} \tag{10}$$

onde $|e_{us}|$ representa quantos cliques de três vértices são formados a partir do vértice i, o que significa o número de conexões compartilhadas por vizinhos diretos de i . Para formar um clique, i precisa ser conectado a u, s e u também precisa estar conectado a s. k_i é o grau do vértice i (WATTS; STROGATZ, 1998).

O coeficiente de agrupamento da rede pode ser obtido pela equação:

$$CC = \frac{1}{N}\sum_{i=1}^{N} CC_i$$

- Centralidade de Proximidade: Esta medida avalia o quão "próximo" um vértice está estruturalmente de todos os demais vértices na rede, refletindo a eficiência potencial com que ele pode disseminar informação, Freeman a definiu em 1977 (FREEMAN et al., 2002). É definida como o inverso da soma das distâncias geodésicas entre o vértice i e todos os outros vértices da rede:

$$C_i = \frac{n-1}{\sum_{j \neq i} d(v_i, v_j)},$$

onde n é o número total de vértices da rede e $d(v_i, v_j)$ representa a distância do caminho mínimo entre i e j. Valores mais altos de C_i indicam nós que podem alcançar outros vértices com menor número de etapas, caracterizando nós centrais e eficientes em termos de comunicação ou fluxo de informação.

- Intermedialidade ou Grau de Intermediação: A centralidade de intermediação, conforme definida por (FREEMAN et al., 2002; CARNEIRO, 2017), mede a frequência com que um nó aparece nos caminhos mais curtos entre outros nós na rede. Um nó com alta centralidade de intermediação atua como uma ponte, sendo crucial para o

fluxo na rede. Ela mede a centralidade de um vértice i com base em sua presença nos caminhos geodésicos entre outros pares de vértices (u, v).

Primeiro, a fração de caminhos mais curtos entre u e v que passam por i é dada por η^i_{uv}:

$$\eta^i_{uv} = \frac{ng^i_{uv}}{ng_{uv}}$$

Onde ng_{uv} é o número total de caminhos mais curtos entre u e v, e ng^i_{uv} é o número desses caminhos que incluem i.

O grau de intermediação total, b_i, é a soma dessas frações para todos os pares de vértices (u, v) distintos de i:

$$b_i = \sum_{u,v \in V - \{i\}} \eta^i_{uv}$$

2.3.3 Aprendizado de Representação em Grafos (Network Embedding)

Para superar as limitações da engenharia manual de características, surgiram as técnicas de *network embedding*. O objetivo é aprender representações vetoriais de baixa dimensão (*embeddings*) para os nós da rede, de forma que as relações e a estrutura do grafo original sejam preservadas nesse novo espaço vetorial (CUI et al., 2018; GOYAL; FERRARA, 2018). Uma vez que os *embeddings* são gerados, eles podem ser utilizados como entrada para qualquer algoritmo de aprendizado de máquina.

Métodos pioneiros como o **DeepWalk** (PEROZZI; AL-RFOU; SKIENA, 2014) utilizam caminhadas aleatórias (*random walks*) para gerar sequências de nós, que são então processadas por algoritmos como o Word2Vec para aprender os *embeddings*. O **Node2vec** (GROVER; LESKOVEC, 2016) aprimorou essa ideia ao introduzir uma estratégia de busca enviesada, permitindo um balanço entre a exploração de vizinhanças locais (homofilia) e globais (equivalência estrutural). Outras abordagens, como o **LINE** (TANG et al., 2015), concentram-se em preservar as proximidades de primeira e segunda ordem na rede. Esses métodos automatizam a extração de características e demonstraram grande eficácia em tarefas como classificação de nós e predição de arestas (CUI et al., 2018; GOYAL; FERRARA, 2018).

2.3.4 Classificação de alto nível e classificação híbrida

Complementando as abordagens anteriores, o estudo de (SILVA; ZHAO, 2012) propõe um modelo híbrido de classificação supervisionada que integra o aprendizado de baixo nível (baseado em atributos físicos) com o aprendizado de alto nível (baseado em padrões semânticos e topológicos), oferecendo uma solução inovadora para problemas de reconhecimento de padrões complexos. Esse método combina a robustez de classificadores

tradicionais, como SVM e k-NN, com uma abordagem fundamentada em redes complexas, na qual os dados são representados por grafos cujas propriedades estruturais, como assortatividade, coeficiente de agrupamento e grau médio, capturam padrões intrínsecos específicos de cada classe.

Durante a fase de treinamento, cada classe é atribuída a um componente isolado da rede por meio de uma estratégia adaptativa de conexão que combina as abordagens de ε-raio e k-vizinhos mais próximos. No momento da classificação, a pertinência de uma instância de teste é determinada não apenas por sua similaridade física com os elementos das classes (termo de baixo nível $T_i^{(j)}$), mas também pela sua conformidade com os padrões topológicos característicos das redes representativas de cada classe (termo de alto nível $C_i^{(j)}$), expressos por meio da seguinte combinação convexa:

$$M_i^{(j)} = (1 - \lambda)T_i^{(j)} + \lambda C_i^{(j)}, \tag{11}$$

em que o parâmetro $\lambda \in [0, 1]$ controla a contribuição relativa de cada termo, sendo especialmente relevante em cenários com classes sobrepostas ou padrões não lineares. Os resultados experimentais demonstram que o modelo proposto supera classificadores tradicionais em tarefas como o reconhecimento de dígitos manuscritos (MNIST), nas quais a identificação de variações e distorções depende da análise semântica das relações entre os dados.

O trabalho de (SILVA; ZHAO, 2012) avalia a conformidade de uma nova instância de teste em relação ao padrão de formação topológica de cada classe. Matematicamente, a pertinência de alto nível $C_i^{(j)}$ de uma instância x_i à classe j é definida pela agregação das variações observadas em diferentes medidas de redes complexas, conforme a Equação 12:

$$C_i^{(j)} = \frac{\sum_{u=1}^{K} \alpha(u) \left[1 - f_i^{(j)}(u)\right]}{\sum_{g \in \mathcal{L}} \sum_{u=1}^{K} \alpha(u) \left[1 - f_i^{(g)}(u)\right]} \tag{12}$$

onde $\alpha(u)$ representa o peso de influência de cada medida topológica u (como assortatividade ou coeficiente de aglomeração), satisfaça a restrição.

$$\sum_{u=1}^{K} \alpha(u) = 1 \tag{13}$$

A função $f_i^{(j)}(u)$, detalhada na Equação 14, mensura o impacto estrutural causado pela inserção da instância x_i na rede da classe j:

$$f_i^{(j)}(u) = \Delta G_i^{(j)}(u) p^{(j)} \tag{14}$$

Nesta formulação, $\Delta G_i^{(j)}(u)$ corresponde à variação normalizada da medida topológica u calculada antes e depois da inserção da instância de teste. O termo $p^{(j)}$, definido na Equação 15, atua como um fator de ponderação que considera a proporção de vértices

pertencentes à classe j em relação ao total de vértices V da rede, mitigando vieses em bases de dados desbalanceadas:

$$p^{(j)} = \frac{1}{V} \sum_{v=1}^{V} I_{\{y_v=j\}} \tag{15}$$

onde I é a função indicadora que retorna 1 se o vértice v pertence à classe j, e 0 caso contrário.

A capacidade do modelo de capturar padrões globais reside na escolha das medidas topológicas que compõem $G(u)$. Uma das medidas fundamentais utilizadas é a **Assortatividade** (r), que avalia a tendência dos vértices de se conectarem a outros com graus similares, explicada anteriormente. A assortatividade da classe j, denotada por $r^{(j)}$, é calculada através do coeficiente de correlação de Pearson dos graus dos vértices conectados pelas arestas E_j, conforme a Equação 16:

$$r^{(j)} = \frac{L^{-1} \sum_{u \in \mathcal{U}_j} j_u k_u - [L^{-1} \sum_{u \in \mathcal{U}_j} \frac{1}{2}(j_u + k_u)]^2}{L^{-1} \sum_{u \in \mathcal{U}_j} \frac{1}{2}(j_u^2 + k_u^2) - [L^{-1} \sum_{u \in \mathcal{U}_j} \frac{1}{2}(j_u + k_u)]^2} \tag{16}$$

onde j_u e k_u são os graus dos vértices nas extremidades da aresta e, e L é o número total de arestas na componente. A variação desta medida $(\Delta Gi^{(j)}(1))$ após a inserção da instância de teste é dada pela diferença normalizada entre a assortatividade original $r^{(j)}$ e a nova assortatividade $r'^{(j)}$:

$$\Delta G_i^{(j)}(1) = \frac{|r'^{(j)} - r^{(j)}|}{\sum_{v \in \mathcal{L}} |r'^{(u)} - r^{(u)}|} \tag{17}$$

Complementarmente, utiliza-se o **Coeficiente de Agrupamento** (*Clustering Coefficient*) para capturar a coesão local da estrutura. O coeficiente local $CC_i^{(j)}$ de um vértice i na classe j quantifica a conectividade entre seus vizinhos diretos (Equação 18):

$$CC_i^{(j)} = \frac{|e_{uk}|}{k_i^{(j)}(k_i^{(j)} - 1)} \tag{18}$$

onde $|e_{uk}|$ é o número de arestas existentes entre os vizinhos do vértice i e $k_i^{(j)}$ é o seu grau. Para caracterizar a classe como um todo, calcula-se a média dos coeficientes locais de todos os n_j vértices da componente (Equação 19):

$$CC^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} CC_i^{(j)} \tag{19}$$

Após realizadas as medidas, é necessário obter a variação dos coeficientes, que é dada através do cálculo de $(\Delta Gi^{(j)}(2))$, sendo assim possível utilizá-la na 14 para conseguir obter a variação de cada medida e finalizar o cálculo de $C_i^{(j)}$ dada a classificação utilizando várias medidas de rede e aplicá-las na classificação híbrida do algoritmo.

$$\Delta G_i^{(j)}(2) = \frac{|CC'^{(j)} - CC^{(j)}|}{\sum_{u \in \mathcal{L}} |CC'^{(u)} - CC^{(v)}|} \tag{20}$$

A principal contribuição do trabalho de (SILVA; ZHAO, 2012) reside na generalidade do arcabouço proposto, onde, em um termo de alto nível, combinando suas medidas, opera de forma independente do classificador de baixo nível utilizado, e na demonstração de que a incorporação de informações estruturais pode melhorar significativamente o desempenho em configurações de classes complexas. Esta pesquisa estabelece um marco para técnicas de classificação que vão além de medidas convencionais de similaridade, abrindo caminho para aplicações em áreas como visão computacional, bioinformática, análise de redes sociais e este trabalho que utiliza redes complexas para trabalhar com dados ATR-FTIR na área da saúde.

2.4 Trabalhos Relacionados

Os crescentes desafios de saúde globais, exemplificados pela recente pandemia de COVID-19, têm sublinhado a necessidade urgente de ferramentas diagnósticas rápidas, precisas e não invasivas (ALYASSERI et al., 2022). A espectroscopia vibracional, em particular a espectroscopia por refletância total atenuada e transformada de Fourier (ATR-FTIR), tem emergido como uma técnica promissora para o diagnóstico biomédico devido à sua capacidade de capturar informações moleculares de amostras biológicas (SANTOS; MORAIS; LIMA, 2020). Quando acompanhada de métodos computacionais avançados, como algoritmos de *machine learning* (ML) e *deep learning* (DL), a ATR-FTIR detém um potencial significativo para a detecção e prognóstico de doenças (TEKLEMARIAM et al., 2024). Esta seção revisa a literatura relevante sobre a aplicação da espectroscopia ATR-FTIR na saúde, com ênfase particular na sua integração com a inteligência artificial para o diagnóstico de doenças, especialmente no contexto da COVID-19 utilizando amostras salivares. Adicionalmente, explora-se o potencial sinérgico da combinação de arquiteturas de *deep learning*, como as Redes Neurais Convolucionais (CNNs), com arcabouços analíticos como redes complexas, para aprimorar a precisão e a interpretabilidade diagnóstica (FILHO et al., 2024).

Trabalhos como os de (ALAJAJI et al., 2025; SITNIKOVA et al., 2020; SILVA et al., 2020; OLIVEIRA et al., 2023; BUTLER et al., 2019; HONÓRIO-SILVA et al., 2024; GARCIA-JUNIOR et al., 2025) fornecem uma visão perspicaz dos avanços na espectroscopia ATR-FTIR combinada com métodos de *machine learning* para o diagnóstico e prognóstico de várias condições de saúde. Esses estudos destacam a capacidade do ATR-FTIR em identificar mudanças bioquímicas sutis em biofluídos e tecidos, que, quando analisadas por algoritmos sofisticados de *machine learning*, podem resultar em altas acurácias diagnósticas. Outro exemplo, o estudo de (DELRUE; BRUYNE; SPEECKAERT, 2025), oferece uma boa revisão de biofluídos aplicados à ATR-FTIR para a descoberta de diversas doenças oncológicas. Embora os artigos mencionados acima se concentrem principalmente em câncer, HIV e doenças virais infecciosas, seus trabalhos estabelecem os princípios fun-

damentais e demonstram a ampla aplicabilidade dessa abordagem integrada, ressaltando o potencial para sua expansão em crises de saúde emergentes (RALBOVSKY; LEDNEV, 2020; MALAMUD, 2011; BUNACIU; HOANG; ABOUL-ENEIN, 2015).

Os esforços globais para o enfrentamento da pandemia de COVID-19, a identificação de perfis espectrais específicos por ATR-FTIR despontou como uma abordagem sustentável promissora para o diagnóstico biomédico. A caracteristíca de teste sem uso de reagentes permitiria uma detecção da COVID-19, mesmo em situações com ausência de reagentes para execução dos testes padrão-ouro por RT-PCR. A patente internacional WO2021237330A1 descreve de forma pioneira a utilização da espectroscopia ATR-FTIR para detectar modos vibracionais com intensidades características em amostras biológicas de pacientes positivos para SARS-CoV-2 e pacientes com síndromes gripais (FILHO et al., 2021). A associação com inteligência artificial a essa abordagem permite a identificação de assinaturas espectrais únicas da infecção, o que pode viabilizar a construção de plataformas diagnósticas rápidas, não invasivas e de baixo custo. O sistema proposto foi avaliado inicialmente por validação cruzada, alcançando 94,9% de acurácia, 96,5% de especificidade e 93,4% de sensibilidade. Em um segundo experimento de validação com um conjunto cego de amostras — não utilizado na etapa de treinamento — obteve-se 85% de acurácia média, 67,3% de especificidade e 98,5% de sensibilidade, demonstrando alta capacidade de generalização na detecção da COVID-19 por meio de espectros obtidos via ATR-FTIR (FILHO et al., 2021).

Avançando na aplicação do *Deep learning*, (ZHANG et al., 2024) apresentou uma contribuição significativa com seu trabalho. Este estudo demonstrou o poder das Redes Neurais Convolucionais, utilizando os mínimos quadrados penalizados interativamente reponderados adaptativamente (airPLS) aplicados em uma CNN (especificamente, um modelo PLS-1D-CNN aprimorado com atenção baseada em canais) para a triagem precisa e rápida de SARS-CoV-2 usando espectros ATR-FTIR. Eles alcançaram acurácias de classificação notáveis, destacando as capacidades superiores de extração de características das CNNs a partir de dados espectrais complexos, mesmo com conjuntos de dados limitados. Sua sofisticada arquitetura de *Deep Learning* fornece um precedente convincente para a integração de CNNs em sistemas de diagnóstico de alta precisão baseados em ATR-FTIR, inspirando diretamente o componente CNN do nosso algoritmo híbrido.

Um estudo altamente relevante (SANTOS et al., 2023; JUNIOR et al., 2025) que aborda o desafio da detecção de COVID-19 a partir de espectros ATR-FTIR de saliva não invasivos utilizando uma técnica proposta de CNN-BiLSTM e também CNN-1D no trabalho envolvendo espectroscopia de Raman. Sua arquitetura CNN-BiLSTM representa uma combinação híbrida de *deep learning* que mescla a robustez das Redes Neurais Convolucionais (CNNs) na extração de características de dados complexos com a capacidade das redes de Memória de Longa Curto Prazo Bidirecional (BiLSTM) de lidar com dependências sequenciais.

A técnica proposta pelo (JUNIOR et al., 2025), foi rigorosamente comparada com uma CNN autônoma e outros modelos convencionais de *machine learning*, incluindo Random Forest (RF), XGBoost (XGB) e Support Vector Machine (SVM). O modelo CNN-BiLSTM consistentemente alcançou resultados superiores em várias métricas de desempenho, notavelmente demonstrando uma acurácia média de 0.80. Além disso, o modelo exibiu um F1-score de 0.80, demonstrando um desempenho robusto e consistente na detecção de COVID-19 a partir de espectros ATR-FTIR de saliva não invasivos.

A arquitetura CNN-BiLSTM, adotada para a classificação espectral, consiste em camadas convolucionais unidimensionais (CNN 1D) seguidas por uma camada *Bidirectional Long Short-Term Memory* (BiLSTM). O BiLSTM é uma variação das redes LSTM tradicionais, capaz de processar informações em duas direções (para frente e para trás), o que permite capturar relações contextuais mais amplas dentro de sequências. No contexto dos espectros ATR-FTIR, a utilização de BiLSTM é particularmente relevante, pois os espectros podem ser interpretados como séries de dados que variam ao longo dos números de onda, apresentando dependências estruturais não triviais.

A arquitetura específica é composta pelos seguintes componentes (JUNIOR et al., 2025):

- Duas camadas convolucionais (`Conv1D`) com função de ativação ReLU e regularização L2;
- Camadas de `MaxPooling1D` entre as convoluções, para redução de dimensionalidade;
- Uma camada `BiLSTM` com retorno de sequências ativado, permitindo o aproveitamento de dependências bidirecionais;
- Uma camada `Flatten`, responsável por transformar os dados em vetor unidimensional;
- Duas camadas densas totalmente conectadas, com aplicação de `Dropout` para regularização;
- Uma camada final densa com ativação `sigmoid`, apropriada para classificação binária (positivo ou negativo para COVID-19).

Os hiperparâmetros desta rede foram otimizados automaticamente utilizando a biblioteca Optuna, com até 200 execuções por simulação para identificar as melhores combinações de parâmetros. O modelo foi treinado com validação cruzada k-fold com $k = 10$, utilizando 90% dos dados para treino (com 20% para validação interna) e 10% para teste. A métrica de otimização foi o F1-Score, considerando o balanceamento do conjunto de dados.

Quando comparada com os outros modelos de *machine learning*, a inclusão da camada BiLSTM demonstrou melhorias significativas em todas as métricas de desempenho, ressaltando a importância da informação temporal na modelagem de tais dados e resultando em maior estabilidade e confiabilidade.

Com base neste estudo fundamental, nosso trabalho visa aprimorar ainda mais esses resultados, integrando Redes Complexas com uma abordagem de classificação híbrida semelhante. As redes complexas têm se destacado como uma abordagem promissora para a classificação de alto nível devido à sua capacidade de modelar as relações estruturais e dinâmicas subjacentes aos dados. Conforme (FERNANDES; SABINO-SILVA; CARNEIRO, 2025) ressalta, o desempenho dessas redes está intimamente relacionado à sua arquitetura, o que justifica a utilização de métodos de otimização para aprimorar sua construção. No estudo apresentado, a autora propõe o GANet, um método que emprega algoritmos genéticos para otimizar redes complexas, utilizando medidas de importância de vértices, tais como PageRank, grau, intermediação, proximidade e comprimento do caminho mais curto, com o intuito de identificar relações relevantes para a tarefa de classificação. Os experimentos realizados com bases de dados reais evidenciam que o GANet supera métodos tradicionais, como o k-vizinhos mais próximos, em termos de acurácia, demonstrando a eficácia da abordagem para cenários de alta complexidade e dimensionalidade. Dessa forma, as redes complexas otimizadas apresentam-se como uma ferramenta robusta e eficiente para aplicações que demandam reconhecimento e classificação precisos, como no caso da análise ATR-FTIR.

Um outro método inovador proposto por (FILHO; SABINO-SILVA; CARNEIRO, 2025), onde foi investigado a detecção do Transtorno do Espectro Autista (TEA) utilizando amostras de saliva coletadas com aparelho ATR-FTIR, a novidade deste estudo foi, além da aplicação de classificação de alto nível utilizando RCs, foi aplicado um novo método de criação de grafos utilizando o grafo de visibilidade (VisG2). O VisG2 é um método de construção de metagrafos projetado para mapear relações sequenciais tanto dentro de um único espectro quanto entre múltiplos espectros. Especificamente, o VisG2 consiste em dois componentes principais: a representação de cada espectro como um grafo de visibilidade e um procedimento de geração de metagrafos baseado na similaridade entre esses grafos. Este é um método inovador para este tipo de problema e trouxe bons resultados comparados a métodos mais conhecidos para criação de grafo. Projetos futuros podem integrar o ATLA com seu método híbrido a novos métodos de criação da rede para fins de melhorar o desempenho dos algoritmos de predição.

De forma diretamente alinhada à proposta desta dissertação, o trabalho de (FILHO et al., 2024), membro do nosso grupo de pesquisa AINET, apresenta uma aplicação robusta da classificação de alto nível para a detecção de câncer oral a partir de espectroscopia ATR-FTIR. A pesquisa valida a hipótese de que a representação de dados espectrais como redes complexas permite a extração de características topológicas com alto poder

discriminativo. Ao converter os espectros em grafos e analisar suas propriedades estruturais, os autores demonstram a capacidade do método para identificar padrões sutis associados à condição patológica, que não seriam facilmente capturados por algoritmos de classificação convencionais. Este estudo é particularmente relevante, pois não apenas reforça a viabilidade da fusão entre espectroscopia vibracional e a teoria de redes complexas, mas também serve como um precedente metodológico importante, aplicando com sucesso esta abordagem inovadora em um desafio clínico concreto e de grande relevância que estamos tratando nesta dissertação.

Capítulo 3

ATLA: ATR-FTIR Topological Learning Analysis

A presente dissertação propõe o desenvolvimento e a validação do método de classificação espectral híbrida de alto nível ATLA, que integra sinergicamente representações topológicas derivadas da teoria de redes complexas com algoritmos de classificação supervisionada de aprendizado profundo e modelos tradicionais de *machine learning*. O objetivo principal é aprimorar significativamente o desempenho na tarefa de classificação de espectros obtidos por espectroscopia ATR-FTIR (SANTOS; MORAIS; LIMA, 2020; BUNACIU; HOANG; ABOUL-ENEIN, 2015), especialmente em cenários com alta similaridade espectral entre classes (ALYASSERI et al., 2022). A abordagem busca capitalizar a capacidade de abstração estrutural e captura de relações globais das redes complexas (CARNEIRO; ZHAO, 2018) com o poder discriminativo de classificadores como Redes Neurais Convolucionais Unidimensionais (CNN 1D) e BiLSTM (que se destacam na extração de características locais e dependências sequenciais), bem como modelos convencionais como Máquinas de Vetores de Suporte (SVM) e Análise Discriminante Linear (LDA).

Como principal contribuição, este método híbrido é aplicado e validado extensivamente em múltiplas bases de dados espectrais ATR-FTIR, representando diversos e desafiadores contextos clínicos e infecciosos. Isso inclui COVID-19, Hepatite Delta (HDV) e doença de Chagas, permitindo uma avaliação abrangente da robustez e generalização do modelo proposto em cenários biomédicos variados. A expectativa é reunir evidências de que a integração de atributos topológicos melhora substancialmente a acurácia, sensibilidade e especificidade dos diagnósticos em comparação com abordagens puramente baseadas em características espectrais tradicionais.

Além disso, a proposta contempla a aplicação de técnicas de pré-processamento espectral, como correção de baseline e normalização, visando à otimização dos sinais espectrais para sua posterior conversão em redes complexas e utilização eficaz nos classificadores. Este trabalho não só contribuiu para o avanço das técnicas de análise espectral, mas tam-

bém implementou uma abordagem interpretável utilizando SHAP, permitindo a análise do impacto de cada atributo espectral e topológico nas decisões dos classificadores. Espera-se que essa interpretabilidade ofereça *insights* clínicos valiosos sobre as regiões espectrais mais relevantes para a diferenciação entre classes, aumentando a transparência e a aplicabilidade prática das metodologias computacionais voltadas ao diagnóstico assistido por espectroscopia.

3.1 Visão Geral

A presente seção detalha o metódo de pesquisa empregado para o desenvolvimento e a validação do método de classificação espectral híbrida, denominado ATLA. Esta técnica visa superar as limitações de modelos convencionais na diferenciação de casos espectralmente similares, integrando o poder de extração de características de modelos de aprendizado profundo com a capacidade de representação estrutural de redes complexas. O método, conforme ilustrado na Figura 11, é dividido em cinco etapas principais: 1) Coleta dos dados; 2) Preparação e pré-processamento dos espectros; 3) Treinamento dos algoritmos de classificação base e de classificação baseado em redes complexas; 4) Predições e aplicação do algoritmo híbrido; e 5) Avaliação de desempenho e interpretabilidade.

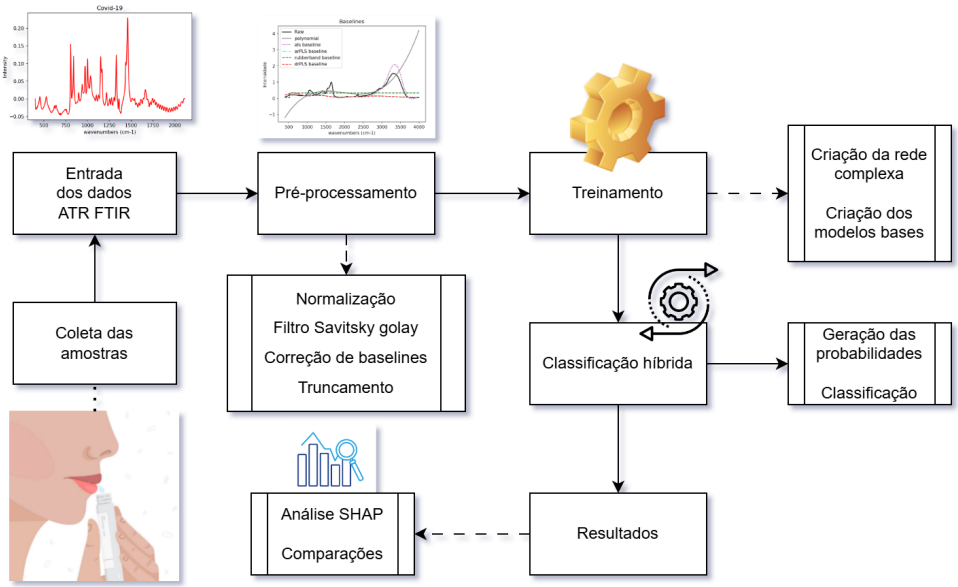


Figura 11 – Diagrama do funcionamento geral do método ATLA Fonte: Autor

3.2 Coleta das Amostras

As amostras foram coletadas utilizando a técnica de espectroscopia ATR-FTIR, um método amplamente reconhecido por ser não invasivo e eficiente. Nossa base de dados foi composta por amostras de saliva de pacientes, o que ressalta a aplicabilidade da técnica

em análises clínicas. Além disso, o ATR-FTIR se destaca em contextos que demandam diagnósticos rápidos e precisos, pois permite a coleta de dados espectrais de forma ágil e vantajosa (KAZARIAN; CHAN, 2013). Entre os principais benefícios dessa metodologia está a capacidade de obter espectros altamente precisos e reproduzíveis, com preparação mínima ou mesmo sem necessidade de preparo das amostras, reduzindo significativamente o tempo e o custo dos procedimentos laboratoriais. Ademais, o ATR-FTIR possibilita a análise rápida e não destrutiva de materiais em diferentes estados físicos — sólidos, líquidos ou pastosos — ampliando sua aplicabilidade em diversas áreas. No contexto da classificação, a técnica fornece dados espectrais ricos e detalhados, que, quando combinados a métodos computacionais avançados, permitem a identificação e diferenciação de substâncias com elevada sensibilidade e especificidade (NASEER; ALI; QAZI, 2021).

A Figura 2 mostra como a amostra é coletada, após obter os espectros ATR-FTIR, serão armazenados e transformados em dados numa tabela para conseguirmos utilizá-los para treinamento dos modelos. A primeira base de dados estudada foi a de COVID-19, esta, seguiu os preceitos éticos da Declaração de Helsinki e obteve aprovação do Comitê de Ética, sob o protocolo de número 4.602.081 (SOUSA, 2022). Todos os participantes forneceram consentimento livre e esclarecido por escrito antes de qualquer procedimento. As amostras foram coletadas simultaneamente de 200 indivíduos recrutados em centros de saúde descentralizados no Brasil, entre março e setembro de 2021. O conjunto de dados foi composto por 100 pacientes com diagnóstico de COVID-19, confirmado por RT-PCR a partir de *swabs* de nasofaringe, e 100 indivíduos do grupo controle (não-COVID-19). Para a análise espectral, foram coletadas amostras de saliva não estimulada por 3 minutos em tubos estéreis, sendo os participantes instruídos a não comer, beber (exceto água) ou fumar por pelo menos uma hora antes da coleta. Subsequentemente, as amostras salivares foram armazenadas a -20°C até o momento da análise (SOUSA, 2022).

Além da base de dados de COVID-19, também foi testado o método na base de dados da doença de Chagas e base de dados de HDV, ambas foram concebidas pelo grupo da Universidade Federal de Uberlândia coordenada pelo Prof. Robinson Sabino Silva e aluna Mariana Araújo Costa, as bases vieram do LACEN-AC - Laboratório Central de Saúde Pública do ACRE, a base da doença de Chagas consiste em 30 amostras do grupo positivo e 50 do grupo controle, que não contém a doença, dada esta situação, a base não está equilibrada, logo é um desafio a mais para nosso método. A base de Hepatite Delta (HDV), que é uma infecção viral hepática causada pelo vírus da hepatite D, um vírus RNA defeituoso que só consegue infectar e replicar-se na presença do vírus da Hepatite B (HBV). Essa coinfecção ou superinfecção pode levar a formas mais graves e progressivas da doença hepática. Após esta coleta, precisamos executar a preparação e o pré-processamento destes dados para conseguir aplicá-los nos modelos de aprendizado de máquina.

3.3 Preparação e Pré-processamento dos Espectros

A preparação dos espectros ATR-FTIR é uma etapa crucial para otimizar a qualidade dos dados e maximizar o desempenho dos modelos de classificação. Desafios como variações indesejadas na linha de base, a presença de ruídos e a alta dimensionalidade dos espectros são abordados por uma sequência de técnicas de pré-processamento, conforme detalhado na Seção 2.1. As etapas aplicadas são:

- **Filtro Savitzky-Golay:** Inicialmente, o filtro Savitzky-Golay (SAVITZKY; GOLAY, 1964) é aplicado para a remoção de ruídos e suavização dos espectros, crucial para preservar a morfologia dos picos de absorção e a informação bioquímica contida nos dados. Esta técnica utiliza o ajuste de um polinômio de baixo grau a uma janela deslizante de pontos para obter um sinal suavizado, minimizando a distorção do sinal original. A configuração específica do filtro utilizada envolveu a definição de um polinômio de "4º grau" e um tamanho de janela de tamanho 25.

- **Correção de *Baseline*:** A correção da linha de base é realizada para remover interferências e desvios de intensidade causados por fatores externos à amostra (PENG et al., 2011). Esta etapa garante a reprodutibilidade e a comparabilidade dos espectros. Entre os métodos investigados e aplicados estão o Polinomial (*poly*), *Asymmetric least squares* (*als*), *Asymmetrically reweighted penalized least squares* (*arPLS*), *Doubly reweighted partial least squares* (*drPLS*) e *Rubberband*.

- **Truncamento do Espectro:** As regiões do espectro correspondentes a cada amostra são truncadas para $900\,\mathrm{cm}^{-1}$ a $1800\,\mathrm{cm}^{-1}$ e $2800\,\mathrm{cm}^{-1}$ a $3050\,\mathrm{cm}^{-1}$. O objetivo é evitar regiões de elevado ruído ou interferências (como as de água) (SOUSA, 2022), que poderiam dificultar a capacidade de generalização dos modelos. A relevância dessas regiões para a classificação será investigada e confirmada por meio da análise de interpretabilidade utilizando SHAP.

- **Normalização pela Amida I:** Como última etapa de pré-processamento, os espectros são normalizados pela região da Amida I ($1630\,\mathrm{cm}^{-1}$ a $1660\,\mathrm{cm}^{-1}$). Essa região é um marcador interno confiável para a caracterização de proteínas (SOUSA, 2022), sendo utilizada para compensar variações de intensidade não relacionadas à composição bioquímica intrínseca da amostra.

3.4 Classificador de Alto Nível

O Classificador de Alto Nível é o componente da abordagem híbrida do ATLA responsável por gerar probabilidades de classificação baseadas na conformidade topológica de uma amostra com as redes complexas de referência predefinidas para cada classe. Seu funcionamento abrange duas fases principais: treinamento e teste.

3.4.1 Fase de Treinamento e Construção das Redes de Referência

Na fase de treinamento, o classifcador de alto nível constrói um conjunto de redes complexas de referência, uma para cada classe presente nos dados de treinamento. Este processo é central para a extração de atributos topológicos que complementarão as características espectrais. Utilizando o método de construção de grafos k-vizinhos mais próximos (kNN), explicado na sessão 2.3.1, cada amostra de treinamento é representada como um nó. As arestas entre os nós são estabelecidas com base nas similaridades de seus vetores de características, formando grafos específicos para cada classe. Como mostrado (CARNEIRO, 2017) estes métodos de construção de redes são bastante utilizados pela literatura e bem otimizados para ATR-FTIR também (FILHO et al., 2024).

Para otimizar a representação topológica, foram investigadas variações da criação de redes por kNN, incluindo o k-Nearest Neighbors Simétrico (SkNN) e o k-Nearest Neighbors Mútuo (mKNN)(CARNEIRO, 2017), que promovem diferentes tipos de conectividade mútua e aprimoram a difusão da informação através do grafo. Após a construção destas redes, um conjunto de medidas topológicas é extraído para caracterizar as propriedades estruturais dos grafos resultantes, fornecendo atributos descritivos adicionais para cada nó. As medidas computadas incluem: Grau Médio, Comprimento Médio do Caminho, Assortatividade, Coeficiente de Agrupamento, Centralidade de Proximidade (*Closeness Centrality*) e Centralidade de Intermediação (*Betweenness Centrality*) (CARNEIRO, 2017; FILHO et al., 2024; GARCIA-JUNIOR et al., 2025; HONÓRIO-SILVA et al., 2024). Estas medidas de referência são armazenadas para cada rede de classe, representando o perfil topológico "esperado" para uma amostra pertencente àquela classe. Os parâmetros de inicialização do classificador incluem o número de vizinhos (k), a métrica de distância (Euclidiana ou Cosseno) e a medida topológica a ser utilizada para a classificação.

3.4.2 Fase de Classificação

Para classificar uma nova amostra de teste, o classificador de alto nível determina sua probabilidade de pertencer a cada classe com base em sua conformidade topológica:

1. **Cálculo de Distâncias e Vizinhança:** Para cada amostra de teste, são calculadas as distâncias em relação a todas as amostras do conjunto de treinamento, essa distância foi calculada por meio distância euclidiana e do cosseno. Em seguida, são identificados os k vizinhos mais próximos da amostra de teste no conjunto de treinamento.

2. **Geração de conexões temporárias e Avaliação Topológica:** Para cada classe existente (e.g., COVID-19 positivo e controle), uma conexão temporária é criada

a partir do grafo de referência daquela classe. A amostra de teste é adicionada a este grafo temporário e conectada aos seus vizinhos mais próximos que pertencem a essa classe. Em seguida, a medida topológica selecionada (e.g., Assortatividade) é recalculada para este grafo temporário.

3. **Cálculo do Score de Conformidade:** A diferença absoluta entre a medida topológica do grafo temporário (com a amostra de teste adicionada) e a medida topológica de referência da classe é calculada. Uma diferença menor indica uma maior conformidade topológica da amostra de teste com aquela classe. Como mostrado por (CARNEIRO, 2017), a conformidade de padrão de uma nova instância de teste y em relação a um componente de rede α é calculada pela técnica de classificação de alto nível, denotada por $H_y^{(\alpha)}$. Este cálculo representa a probabilidade de y pertencer ao componente α e é formalmente definido pela seguinte equação, que pondera as variações de múltiplas medidas de redes complexas (CARNEIRO, 2017):

$$H_y^{(\alpha)} = \frac{\sum_{u=1}^{Z} \delta_y(u)[1 - f_y^{(\alpha)}(u)]}{\sum_{g \in G} \sum_{u=1}^{Z} \delta_y(u)[1 - f_y^{(g)}(u)]}$$

onde o termo $f_y^{(\alpha)}(u)$ representa a variação normalizada da u-ésima medida de rede, considerando a proporção de instâncias $p^{(\alpha)}$ que pertencem ao componente α:

$$f_y^{(\alpha)}(u) = \frac{\Delta G_y^{(\alpha)}(u) p^{(\alpha)}}{\sum_{\alpha} \Delta G_y^{(\alpha)}(u) p^{(\alpha)}}$$

Para evitar que uma única medida de rede domine a decisão, um coeficiente de influência $\delta_y(u)$ é definido de maneira automática para balancear a contribuição de cada medida, subtraindo de 1 a diferença entre a máxima e a mínima variação observada para aquela medida em todos os componentes e, em seguida, normalizando:

$$\delta_y(u) = \frac{1 - \left(\max_{\alpha \in G} \Delta G_y^{(\alpha)}(u) - \min_{\alpha \in G} \Delta G_y^{(\alpha)}(u)\right)}{\sum_{u=1}^{Z} \left[1 - \left(\max_{\alpha \in G} \Delta G_y^{(\alpha)}(u) - \min_{\alpha \in G} \Delta G_y^{(\alpha)}(u)\right)\right]}$$

Dessa forma, a conformidade de padrão é estabelecida como uma avaliação ponderada e balanceada das perturbações que a inserção do item de teste y causa na topologia de cada componente da rede de referência (CARNEIRO, 2017).

4. **Normalização e Geração de Probabilidades:** Os scores de conformidade calculados para todas as classes são normalizados. Em seguida, são ponderados pela proporção de amostras de cada classe no conjunto de treinamento. O resultado final é uma matriz de probabilidades, onde cada entrada representa a probabilidade da amostra de teste pertencer a cada classe, baseada em sua conformidade topológica.

3.5 Classificador Híbrido

O algoritmo de classificação híbrido ATLA combina as probabilidades geradas pela rede complexa com as probabilidades obtidas por outros algoritmos de classificação base, como a Rede Neural Convolucional (CNN) ou outros algoritmos tradicionais (SVM, LDA). O processo de combinação ocorre com a obtenção de probabilidades dos classificadores base, onde, a amostra de teste é primeiramente classificada pelo classificador base, como a CNN neste caso, gerando uma matriz de probabilidades. Juntamente, o classificador da rede complexa, como explicado na seção 3.4 gera a matriz de probabilidades de alto nível. Em seguida, a matriz de probabilidades a matriz de probabilidades do classificador base são combinadas linearmente, conforme o Algoritmo 3 da Metodologia. Um parâmetro de balanceamento $\lambda \in [0, 1]$ é utilizado para ponderar a contribuição de cada classificador. Após isso, é feita a predição final, isto é, a classe final da amostra de teste é determinada pelo índice da classe com a maior probabilidade resultante da combinação.

O ATLA, visa combinar as forças dos classificadores de aprendizado profundo/tradicionais com a análise de redes complexas. A estratégia de combinação será implementada por meio de uma fusão tardia, isto é, a combinação das saídas de probabilidade dos classificadores base (e.g., CNN-BiLSTM, SVM, LDA) com a probabilidade topológica derivada da classificação baseada em redes complexas. A combinação é realizada através de uma soma ponderada, utilizando um parâmetro de balanceamento $\lambda \in [0, 1]$. O valor de λ será otimizado via busca em grade no conjunto de validação, permitindo um equilíbrio dinâmico entre as informações de características locais e estruturais globais. Especificamente, quanto menor o valor de λ (próximo a 0), maior o peso atribuído às probabilidades dos algoritmos de base; inversamente, quanto mais próximo de 1.0, maior a influência das probabilidades geradas pelo algoritmo de alto nível (Rede Complexa).

Com isso, conseguimos utilizar o ATLA para classificar os espectros ATR-FTIR do com alta performance utilizando o aprendizado das redes complexas de alto nível utilizando a topologia e também a vantagem de estar utilizando os algoritmos bases como CNN-1D, SVM e LDA já conhecidos do estado da arte, assim melhorando o resultado final obtido e aumentando a confiabilidade na predição para área de saúde e detecção de doenças, conseguindo auxiliar médicos em diagnósticos e também em diagnósticos mais rápidos dependendo da situação.

Algoritmo 3 Estratégia de Fusão do Classificador Híbrido

Input: `prob_alto_nivel`: matriz de probabilidades do classificador de alto nível
`prob_baixo_nivel`: matriz de probabilidades do classificador de baixo nível
`lambda_blend`: parâmetro de balanceamento $\lambda \in [0, 1]$
Output: `resultado`: vetor de índices de classe predita (n_amostras × 1)

Função `fusao_classificador_hibrido`(*prob_alto_nivel, prob_baixo_nivel, lambda_blend*)**:**
  n_linhas, n_colunas ← dimensões de `prob_alto_nivel`
  prob_hibrida ← matriz de zeros de dimensão (n_linhas × n_colunas)
  for $i \leftarrow 0$ **to** $n\_linhas - 1$ **do**
    for $c \leftarrow 0$ **to** $n\_colunas - 1$ **do**
      `prob_hibrida`[i][c] ← (1 − lambda_blend) × prob_baixo_nivel[i][c] + lambda_blend × prob_alto_nivel[i][c]
    end
  end
  `resultado` ← índices dos valores máximos em cada linha de `prob_hibrida`
  return *resultado*
fim

3.6 Interpretabilidade dos Modelos

Para validar a performance do classificador proposto em comparação com outros algoritmos da literatura. As métricas consideradas serão: Acurácia, Precisão, Sensibilidade, Especificidade e F1-Score, explicadas anteriormente na Seção 2.2.5. Essas métricas são cruciais em diagnósticos médicos, permitindo identificar a capacidade do algoritmo de classificar corretamente casos positivos e negativos, minimizando falsos negativos e falsos positivos, e garantindo a confiabilidade e eficiência do modelo proposto, sendo importantes para comparação entre cada técnica e entender os pontos fortes de cada abordagem. Analisando as matrizes de confusão podemos ter uma boa visão de como o modelo está se comportando e os casos em que ele está mais errado, assim conseguindo ter mais noção da base e de como algum parâmetro possa ser ajustado.

Adicionalmente, para promover a interpretabilidade dos modelos e fornecer informações sobre as regiões espectrais mais relevantes, será realizada uma **análise de interpretabilidade utilizando o método** SHAP, explicado na Seção 2.2.4, nesta etapa de avaliação pelo método SHAP, pedimos a especialistas da área de espectroscopia para analisar os melhores componentes que o SHAP retorna. Com isso, os resultados de explicação obtidos pelo ATLA, onde, estes especialistas analisam os modos vibracionais que mais contribuíram para as predições e eventualmente associam componentes biológicos específicos que podem contribuir na discriminação de grupos alvo e controle. Ademais, essa etapa também pode contribuir no processo de seleção de atributos, dirigindo o refinamento dos modelos de aprendizado para regiões potenciais dos espectros, biologicamente associadas

com potencial discriminador para condições alvo e controle dos dados investigados.

Capítulo 4

Resultados experimentais

Este capítulo detalha os experimentos conduzidos e apresenta a análise dos resultados obtidos, visando avaliar a proposta do ATLA para a classificação espectral ATR-FTIR em múltiplas bases de dados biomédicas. Serão demonstradas a eficácia, robustez e capacidade de generalização do algoritmo híbrido em cenários clínicos diversos e desafiadores.

4.1 Ambiente Experimental

Para execução do pipeline a seguir foi montado um pipeline, conforme mostrado na Figura 11, para seguirmos e manter como um padrão para manter conciso todos os experimentos. O sistema foi executado em um computador pessoal Desktop com os seguintes componentes processador: AMD Ryzen 5 7600 6-Core, com velocidade 3.80 GHz, memória RAM instalada de 32,0 GB, sistema operacional Windows 11 Pro, códigos executados na linguagem de programação Python versão 3.10, utilizando Jupyter Notebook, com a placa de video NVIDIA RTX 3060Ti.

Com a configuração de hardware disponível, o treinamento do ATLA demandou um tempo computacional considerável. Esta demanda é inerente à complexidade do algoritmo e ao extenso processo de otimização de hiperparâmetros, que envolve a exploração de diversos valores para o número de vizinhos $k \in [1, 15]$ na criação das RCs e para o parâmetro de balanceamento $\lambda \in [0, 1]$. O parâmetro λ é crucial na estratégia de combinação do algoritmo híbrido, ponderando as probabilidades dos classificadores base (SVM, LDA e CNN) e de alto nível baseado em RCs.

A rede foi criada utilizando o método kNN, definido a partir de testes empíricos realizados tanto com a distância euclidiana quanto com a distância de cosseno para o cálculo da similaridade entre as amostras. As medidas topológicas extraídas das RCs para compor as probabilidades de alto nível incluíram Grau Médio, Comprimento Médio do Caminho, Assortatividade, Coeficiente de Agrupamento, Centralidade de Proximidade e Centralidade de Intermediação.

Para o treinamento dos modelos, foi realizado a normalização pela *Amida I* e aplicação do filtro *Savitzky-golay*. Para cada técnica, foi simulada a estratégia de validação cruzada *k-Fold*, configurada com 10 grupos e 3 repetições, garantindo uma avaliação robusta e estatisticamente significativa do desempenho dos classificadores. Os resultados gerados pelas medidas de redes do classifiador de alto nível e dos algoritmos de aprendizado de máquina convencionais (SVM e LDA) e algoritmos de aprendizado profundo (CNN-BiLSTM), foram treinados independentemente e depois as probabilidades foram combinadas com o ATLA para melhorar o resultado em geral. Na parte de experimentação foram testados Random Forest e também SVM não linear, porém decidimos continuar apenas com LDA e SVM por testes empíricos, onde o LDA foi utilizado o da biblioteca *sklearn* SVM foi testados diversos parâmetros C e números de interação. Ambos foram os melhores algoritmos bases que conseguimos em grande parte dos testes, onde seus resultados foram superiores.

4.2 ATLA Aplicado na Detecção de COVID-19

Esta é a investigação principal da dissertação, realizada com objetivo de melhorar o resultado obtido pela CNN-BiLSTM publicada em (JUNIOR et al., 2025) através da nossa técnica ATLA. Os parâmetros da CNN foram mostrados na sessão 2.4.

4.2.1 Descrição da Base de Dados COVID-19

A base de dados utilizada neste estudo é um conjunto meticulosamente curado de amostras salivares destinado à classificação de COVID-19. A obtenção desta base seguiu os preceitos éticos da Declaração de Helsinki e obteve aprovação do Comitê de Ética em Pesquisa da UNA, sob o protocolo de número 4.602.081 (SOUSA, 2022). Todos os participantes forneceram consentimento livre e esclarecido por escrito antes de qualquer procedimento. A média dos espectros na base de dados de COVID-19 podem ser vistos na Figura 1. As amostras foram coletadas de 200 indivíduos recrutados em centros de saúde descentralizados no Brasil, entre março e setembro de 2021. O conjunto de dados foi composto por 100 pacientes com diagnóstico de COVID-19, confirmado por RT-PCR a partir de *swabs* de nasofaringe, e 100 indivíduos do grupo controle, os quais embora sintomáticos, testaram negativo para a COVID-19, estabelecendo uma proporção de 1:1 entre as classes. Para a análise espectral, foram coletadas amostras de saliva não estimulada por 3 minutos em tubos estéreis, sendo os participantes instruídos a não comer, beber (exceto água) ou fumar por pelo menos uma hora antes da coleta. Subsequentemente, as amostras salivares foram armazenadas a -20°C até o momento da análise (SOUSA, 2022). A validação por RT-PCR fornece rótulos de verdade absoluta, essenciais para o aprendizado supervisionado. Os espectros de absorbância foram coletados na faixa do

infravermelho médio (de $4000\,\mathrm{cm}^{-1}$ a $400\,\mathrm{cm}^{-1}$ com resolução de $4\,\mathrm{cm}^{-1}$), gerando vetores espectrais de alta dimensionalidade, onde cada banda de infravermelho(*wavenumber*) codifica informações bioquímicas específicas sobre a composição molecular da amostra. Os espectros ATR-FTIR das amostras de saliva, tanto positivas para COVID-19 quanto do grupo controle, exibem ruído de *baseline* significativo e variabilidade interamostral, características inerentes à complexa composição bioquímica da saliva como foi explicado na fundamentação. Podemos ver um exemplo de rede formada a partir destes espectros na Figura 12, onde temos ambas classes representadas - grupo alvo e controle - conseguimos ver que estão bem distintas uma da outra na representação.

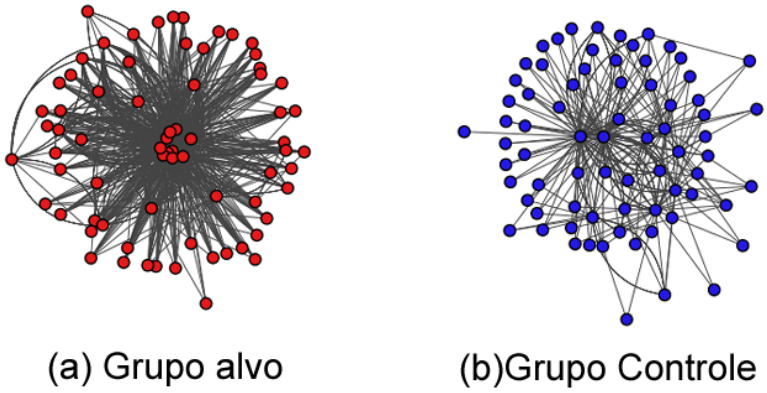


Figura 12 – Plotagem da rede complexa da base de dados de COVID-19 (a) Grupo alvo, (b) Grupo controle. Fonte: Autor

4.2.2 Resultados de Modelos Base para Detecção de COVID-19

Os resultados médios obtidos pelo modelo CNN-BiLSTM ao longo das 10 simulações demonstram um forte desempenho de linha de base na detecção de COVID-19, conforme resumido na Tabela 2. A arquitetura CNN-BiLSTM superou todos os modelos comparativos avaliados individualmente (incluindo uma CNN simples com os mesmos hiperparâmetros, porém sem a camada BiLSTM). A inclusão da camada BiLSTM aprimorou significativamente a capacidade do modelo em capturar padrões temporais nos espectros, conferindo maior robustez, estabilidade e precisão à CNN-BiLSTM. O melhor modelo testado na base de COVID-19 foi o da CNN-BiLSTM, obtendo 83% de Acurácia, Precisão de 87%, Sensibilidade 76%, Especificidade 89% e F1-Score de 81%. Sendo o modelo com melhor performance, apenas perdendo para SVM que conseguiu um melhor resultado na sensibilidade de 88%, isso nos mostra que ele também é um modelo que possui um bom desempenho com esse tipo de dado, onde utilizamos no ATLA resultados obtidos pelo SVM. O modelo também apresentou os menores desvios padrão entre as execuções, indicando alta consistência em seus resultados.

Tabela 2 – Desempenho médio dos modelos na detecção de COVID-19 (JUNIOR et al., 2025).

Modelo	**Acurácia**	**Precisão**	**Sensibilidade**	**Especificidade**	**F1-Score**
SVM	0.54 ± 0.0937	0.52 ± 0.0642	**0.88± 0.1932**	0.20 ± 0.1333	0.65 ± 1932
CNN	0.76 ± 0.0568	0.74 ± 0.0696	0.83 ± 0.1252	0.69 ± 0.1197	0.77 ± 0.1252
CNN-BiLSTM	**0.80±0.0926**	**0.80±0.1254**	**0.82±0.1033**	**0.77±0.1567**	**0.80±0.0860**

4.2.3 Resultados de Classificação de Alto Nível de COVID-19 com ATLA

Esta seção apresenta os resultados obtidos da aplicação da abordagem híbrida proposta, detalhando o desempenho dos classificadores nas tarefas de detecção de COVID-19 em espectros ATR-FTIR. As análises focam na comparação entre modelos baseados em RCs e arquiteturas de aprendizado profundo (CNN-BiLSTM), bem como classificadores tradicionais como SVM e LDA, para avaliar a eficácia do ATLA.

Os resultados, apresentados nas Tabelas 4 e 3, mostram os resultados da estratégia de classificação híbrida que integra uma arquitetura CNN-BiLSTM com modelos baseados em RCs. Modelos tradicionais de aprendizado profundo, como CNNs e BiLSTMs, são proficientes na extração de características espaciais locais e na captura de dependências temporais, respectivamente. No entanto, frequentemente, não conseguem representar as relações estruturais globais entre as amostras. Para abordar essa limitação, incluímos medidas topológicas de redes complexas, que aprimora a capacidade discriminativa do modelo base ao incorporar informações relacionais no processo de classificação. As siglas utilizadas para as medidas de rede serão: Grau médio, *Average Degree* (ADG), Menor caminho médio - *Average shortest path* (ASP), Assortatividade, *Assortativity* (ASS), Intermedialidade, *Betwenness* (BET), CLO, Coeficiente de Agrupamento - Cluster Coefficient (CCF).

Tabela 3 – Comparação dos resultados ATLA aplicado em algoritmos tradicionais (SVM e LDA).

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA(kNN BET + SVM)	0.8	0.69± 0.043	0.62± 0.0358	0.47±0.0272	**0.94±0.0608**	**0.74±0.046**
ATLA(kNN BET + LDA)	0.9	**0.72±0.0832**	**0.73±0.0700**	**0.79±0.114**	0.65 ± 0.1107	0.69 ± 0.0776
ATLA (kNN ASP)	0.7	0.69 ± 0.0931	0.65 ± 0.0756	0.63 ± 0.1146	0.76 ± 0.1215	0.70 ± 0.0891
SVM Base	0	0.60±0.0143	0.60 ± 0.0122	0.60 ± 0.0272	0.60 ± 0.0121	0.60 ± 0.0079
LDA Base	0	0.47±0.0115	0.45 ± 0.0098	0.37 ± 0.1231	0.59 ± 0.0244	0.51 ± 0.0154

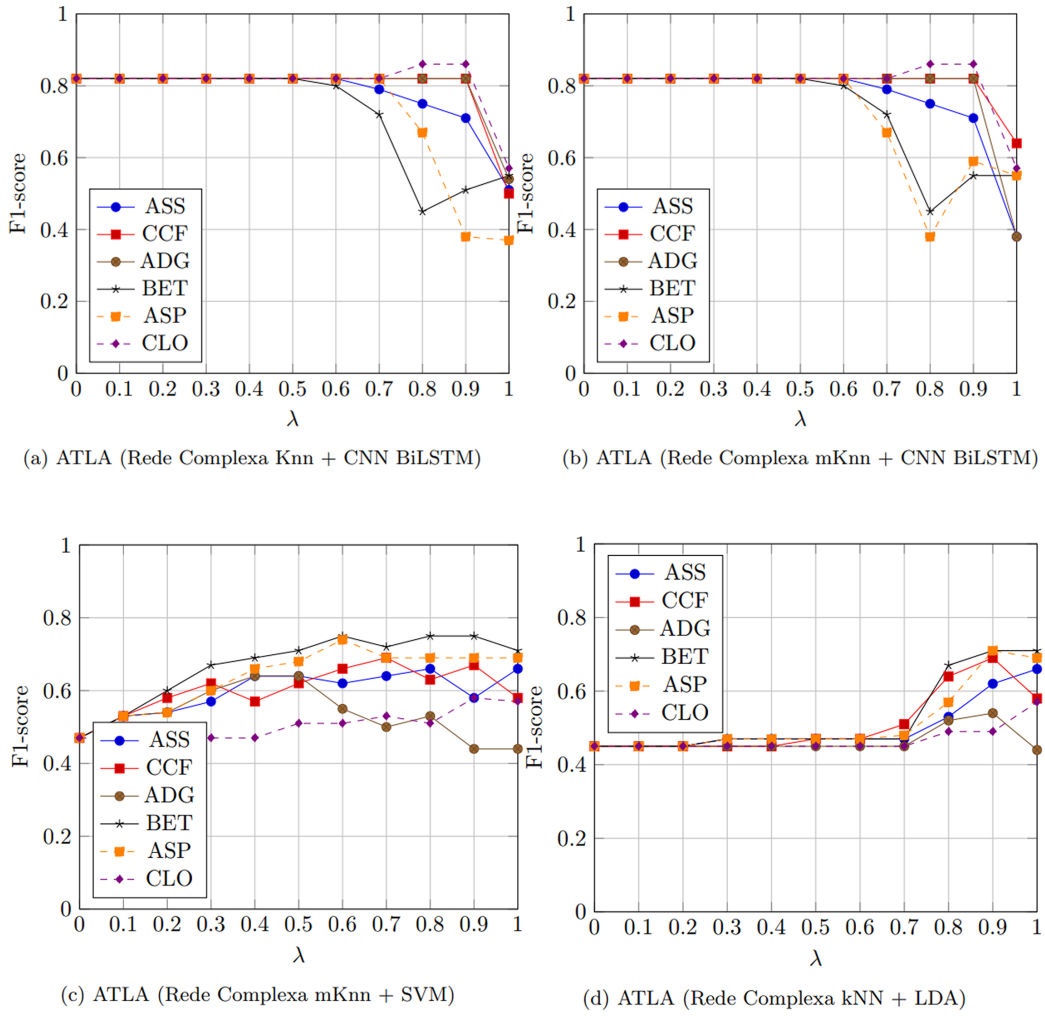
Tabela 4 – Comparação dos resultados ATLA utilizando métodos diferentes para criação da rede (kNN, mKnn, sKNN) em termos de Acurácia (AC), Precisão (PR), Especificidade (ES), Sensibilidade (SE) e F1-Score

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA(kNN CLO + CNN)	0.8	**0.86 ± 0.0859**	**0.88± 0.0779**	**0.89±0.1315**	**0.82±0.2076**	**0.85±0.0299**
ATLA(MkNN CCF + CNN)	0.9	0.83 ± 0.0098	0.87 ± 0.0034	0.89 ± 0.0012	0.76 ± 0.0207	0.82 ± 0.0135
ATLA(sKNN CCF + CNN)	0.7	0.83 ± 0.0309	0.87 ± 0.1588	0.89 ± 0.0951	0.76 ± 0.1647	0.81 ± 0.207
CNN BiLSTM Base	0	0.80±0.0926	0.82 ± 0.1033	0.77 ± 0.1567	0.82 ± 0.1033	0.80 ± 0.0860

Conforme evidenciado na Tabela 3, o modelo CNN-BiLSTM alcançou uma acurácia e um F1-score de 0.83 e 0.82, respectivamente. Quando integrado a vários métodos de construção de grafos baseados em kNN, como kNN simples, kNN mútuo (MkNN), kNN simétrico (SkNN), o desempenho da CNN demonstrou melhorias notáveis. As configurações kNN Graph + CNN atingiu o desempenho mais elevado, com F1-scores de 0.86 e 0.85, respectivamente, mostrando que a combinação feita pelo ATLA realmente melhora os resultados. Essas melhorias evidenciam que a incorporação de informações estruturais via medidas topológicas baseadas em grafos, como a centralidade de proximidade (*closeness centrality*) e o coeficiente de agrupamento, pode contribuir positivamente para o resultado da classificação.

Os gráficos exibidos na Figura 13 mostram os resultados preditivos obtidos em relação à variação do parâmetro λ (lambda) para várias medidas de RC, tais como caminho mais curto médio (ASP), centralidade de proximidade (CLO), centralidade de intermediação (BET) e coeficiente de agrupamento (CCF). Tais resultados mostram que os modelos híbridos mantêm um desempenho superior em uma ampla gama de valores de λ, particularmente para a Centralidade de Proximidade nos casos utilizando a CNN BiLSTM. Isso sugere que o modelo híbrido pode possuir algum nível de robustez a mudanças na esparsidade do grafo e pode capturar características estruturais nos dados, mesmo sob densidades de grafo variadas.

A Tabela 3 dos resultados, as Figuras 14 que mostra a evolução do desempenho em relação a quantos vizinhos "k" criamos o grafo, também nas Figuras 13(c) e (d), fornecem *insights* comparativos adicionais ao avaliar estratégias híbridas que incluem classificadores tradicionais como SVM e LDA. Quando utilizados isoladamente, os resultados dos algoritmos SVM e LDA apresentaram desempenho inferior (F1-scores de 0.60 e 0.51, respectivamente). No entanto, o ATLA aprimorou significativamente sua capacidade de classificação, por exemplo, com as configurações ATLA (SVM + RC) e ATLA (LDA + RC) esses modelos alcançaram F1-scores de 0.74 e 0.69. As medidas de rede da rede complexa base (quando $\lambda = 1$ consegue alcançar resultados melhores do que o SVM e o LDA base, porém quando combinadas consegue um melhor desempenho. Isso reforça a noção de que tanto modelos de aprendizado profundo quanto modelos tradicionais podem se beneficiar da integração de informações relacionais baseadas em grafos, corroborando com nossa hipótese.



(a) ATLA (Rede Complexa Knn + CNN BiLSTM)

(b) ATLA (Rede Complexa mKnn + CNN BiLSTM)

(c) ATLA (Rede Complexa mKnn + SVM)

(d) ATLA (Rede Complexa kNN + LDA)

Figura 13 – Gráficos de comparação do melhor resultado de cada modelo por cada medida de rede para base de COVID-19.

Finalmente, o gráfico da Figura 13 que apresenta os F1-scores para diferentes valores de k para o modelo RC kNN+CNN (onde $\lambda = 0.8$) destaca a importância da seleção de parâmetros de grafo apropriados. O melhor desempenho foi alcançado em torno de $k = 13$, enfatizando ainda mais o valor do ajuste fino de parâmetros estruturais em arcabouços híbridos. Além disso, a média dos valores de F1-Score é próximo de 0.8, mostrando a robustez do método ATLA. Estes padrões reconhecidos, podem nos ajudar a encontrar o valor específico e otimizado de K-vizinhos mais próximos para conseguir otimizar o algoritmo para conseguir o resultado mais rápido.

4.2.4 Análise de Interpretabilidade SHAP para COVID-19

Para oferecer transparência às decisões do modelo híbrido ATLA e identificar os atributos mais relevantes para a classificação de COVID-19, foi realizada uma análise de interpretabilidade utilizando o método SHAP para o melhor resultado obtido na base de

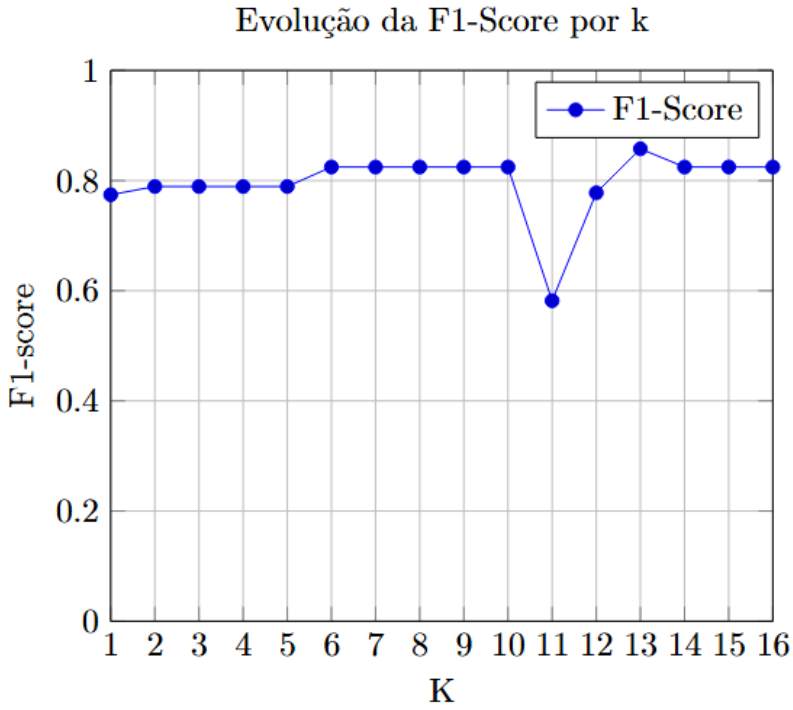


Figura 14 – K-vizinhos mais próximo do melhor resultado do ATLA para COVID-19 utilizando CNN + RC kNN (Medida: Closeness)

COVID-19. A Figura 15 ilustra os valores SHAP médios (impacto médio absoluto na magnitude da saída do modelo) para os *wavenumbers* mais importantes, considerando a configuração híbrida que utilizou a centralidade de proximidade (*Closeness*) como medida de rede complexa e $k = 13$ vizinhos.

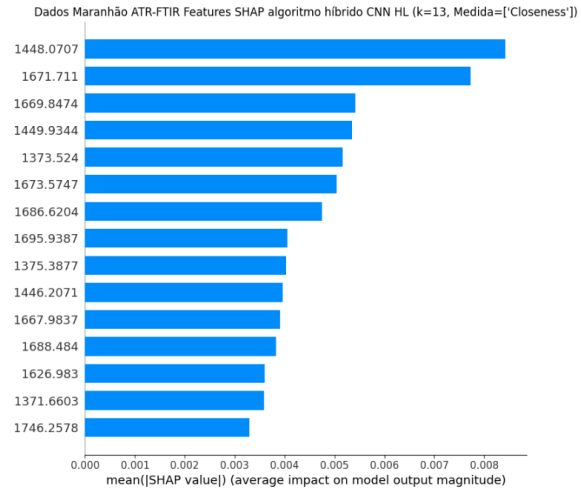


Figura 15 – SHAP do melhor resultado do ATLA (RC kNN CLO + CNN)

A Figura 15 revela que os *wavenumbers* com maior impacto na classificação da COVID-19 concentram-se predominantemente em regiões específicas do espectro. Destacam-se valores como 1446.0707 cm^{-1}, 1671.711 cm^{-1}, 1669.8474 cm^{-1} e 1449.9344 cm^{-1} como os de maior relevância. A alta contribuição de *wavenumbers* próximos à região da Amida I (1630−1660 cm^{-1}), como 1671.711 cm^{-1}, 1669.8474 cm^{-1}, 1673.5747 cm^{-1} e 1656.6204 cm^{-1},

corrobora a importância da normalização por esta banda e da informação de proteínas para a diferenciação de amostras. Outras regiões, como as próximas a $1440\,\mathrm{cm}^{-1}$ e $1370\,\mathrm{cm}^{-1}$ (associadas a deformações de C-H em lipídios e proteínas), também mostram impacto significativo.

Esta análise de interpretabilidade não apenas valida a relevância das faixas espectrais selecionadas no pré-processamento, mas também fornecendo informações valiosas sobre os biomarcadores potenciais associados à infecção por COVID-19. Ao destacar quais *wavenumbers* são cruciais para as decisões do ATLA, o SHAP contribui para a maior transparência do modelo e para o direcionamento de futuras pesquisas clínicas na busca por correlações biológicas específicas.

Em resumo, os resultados experimentais demonstram que métodos de classificação híbridos, que alavancam tanto representações de redes neurais profundas quanto estruturas de redes complexas, melhoram significativamente a capacidade do modelo de distinguir entre classes. Isso é especialmente verdadeiro em casos com variações interclasses sutis, onde modelos tradicionais podem apresentar dificuldades.

4.3 ATLA Aplicado na Detecção da Doença de Chagas

Os experimentos nesta base seguiram a mesma metodologia geral de pré-processamento, criação de redes, treinamento e avaliação descrita detalhadamente na Seção 4.1. Contudo, observou-se uma particularidade relevante no pré-processamento: o melhor desempenho para esta base foi obtido utilizando o espectro em sua forma *Base*, isto é, sem correção de baseline e aplicando-se apenas a normalização pela Amida I e o método de suavização *Savitzky-Golay*. Isto sugere que, para certas características desta base, a correção de *baseline* pode não ser necessária ou até mesmo introduzir artefatos que dificultam o aprendizado dos padrões pelo algoritmo, provando a importância de testar diferentes estratégias de pré-processamento.

4.3.1 Descrição da Base de Dados de Chagas

A base de dados da doença de chagas veio do LACEN-AC - Laboratório Central de Saúde Pública do ACRE, esta base de dados é composta por espectros ATR-FTIR, sendo 30 amostras do grupo que testou positivo a esta doença (grupo alvo), e 50 amostras do grupo controle.

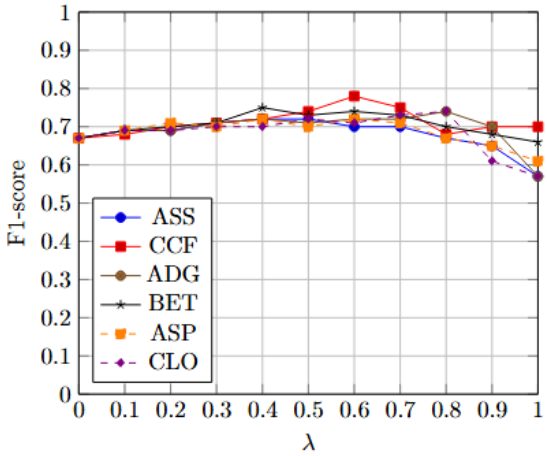


Figura 16 – Melhor resultado obtido da base de doença de Chagas, Target X Control - (ATLA - Cluster coefficient com SVM)

4.3.2 Resultados de Classificação de Alto nível de Chagas com ATLA

A Tabela 5 sumariza o desempenho dos modelos de classificação (LDA, SVM, Rede Complexa individual e o híbrido SVM+RC) na detecção da doença de Chagas. O algoritmo híbrido (SVM+RC) alcançou os melhores resultados, com um F1-Score de **0.78** e acurácia de **0.81**. Este desempenho superior demonstra a eficácia da abordagem de combinação na tarefa de classificação desta base.

Tabela 5 – Comparação dos resultados ATLA utilizando algoritmos tradicionais na base de dados da doença de Chagas

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA (kNN CCF + SVM)	0.6	**0.81±0.0369**	**0.72± 0.0499**	**0.79±0.1493**	0.85±0.1828	**0.78±0.1083**
ATLA (kNN BET + LDA)	0.8	0.72 ± 0.0412	0.59 ± 0.0911	0.62 ± 0.1128	**0.87 ± 0.1107**	0.70 ± 0.0329
ATLA (kNN BET)	1	0.70 ± 0.0841	0.59 ± 0.0809	0.67 ± 0.2430	0.74 ± 0.1870	0.66 ± 0.1329
SVM	0	0.74±0.0035	0.65 ± 0.0017	0.76 ± 0.0070	0.70 ± 0.0196	0.67 ± 0.0099
LDA	0	0.71±0.0006	0.62 ± 0.0009	0.73 ± 0.0013	0.68 ± 0.0016	0.65 ± 0.0008

A Figura 16 ilustra o comportamento do F1-score em função do parâmetro λ para as diferentes medidas de rede complexa. Como podemos observar, o desempenho do modelo híbrido é aprimorado conforme o valor de λ aumenta até um ponto ótimo, com o melhor resultado (F1-score de 0.78) sendo observado em $\lambda = 0.6$. Neste caso específico, a melhor medida de rede para a Rede Complexa foi o Coeficiente de Cluster, que, ao ser combinada com o SVM, otimizou a classificação. Isso reforça que a otimização de λ é crucial para maximizar os benefícios da abordagem híbrida, e que testar diferentes medidas topológicas e condições de pré-processamento é essencial para cada base de dados.

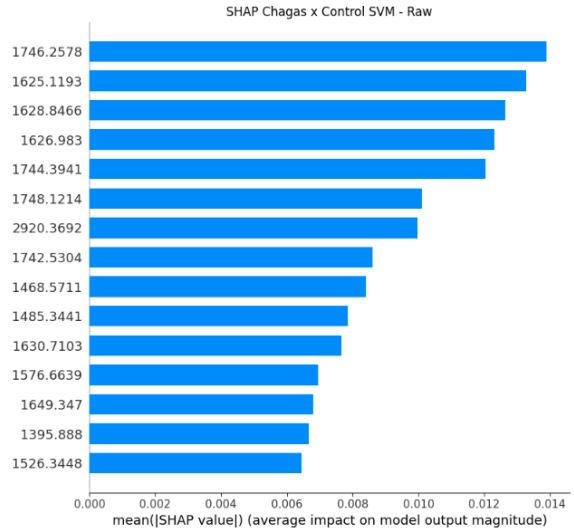


Figura 17 – SHAP - Análise da base de chagas do melhor resultado obtido

4.3.3 Análise de Interpretabilidade SHAP para Chagas

A análise de interpretabilidade SHAP para o melhor modelo híbrido na base de Chagas (SVM+RC) foi realizada para identificar os wavenumbers que mais contribuíram para as decisões de classificação. A Figura 17 revela que os *wavenumbers* com maior impacto na classificação da doença de Chagas estão concentrados em regiões como $1746.2578\,\mathrm{cm}^{-1}$, sendo o modo vibracional com potencial para discriminar a doença de Chagas de amostras controles. seguido por $1625.1193\,\mathrm{cm}^{-1}$, $1628.8466\,\mathrm{cm}^{-1}$ e $1626.983\,\mathrm{cm}^{-1}$. Estes valores são notavelmente próximos à região da Amida I ($1630 - 1660\,\mathrm{cm}^{-1}$), o que valida a relevância da normalização por esta banda e corrobora a importância da informação de proteínas para o diagnóstico. Outros *wavenumbers* relevantes incluem $2920.3692\,\mathrm{cm}^{-1}$ (associado a vibrações de estiramento de C-H) e $1468.5711\,\mathrm{cm}^{-1}$. A predominância dessas regiões ressalta a capacidade do algoritmo híbrido de identificar biomarcadores espectrais cruciais para a diferenciação de casos de Chagas, fornecendo descobertas importantes que podem guiar futuras pesquisas clínicas.

4.4 ATLA Aplicado na Detecção de Hepatite B e D

Nesta base de dados, foi obtido amostras espectrais ATR-FTIR de Hepatite Delta, esta é uma infecção viral hepática causada pelo vírus da hepatite D (HDV), um vírus RNA defeituoso que só consegue infectar e replicar-se na presença do vírus da Hepatite B (HBV). Essa coinfecção ou superinfecção pode levar a formas mais graves e progressivas da doença hepática.

4.4.1 Descrição da Base de dados de HDV

Para investigar a capacidade do algoritmo ATLA na diferenciação de estados relacionados a essa condição, foi utilizada uma base de dados específica de espectros ATR-FTIR com quatro grupos distintos de amostras, denominados targets: Target A (grupo controle sem vacina BCG, 87 amostras), Target B (grupo controle com vacina BCG, 94 amostras), Target C (indivíduos infectados apenas com Hepatite B, 93 amostras) e Target D (indivíduos coinfectados com Hepatite B e D, 85 amostras). No contexto deste estudo experimental, foram realizadas as seguintes comparações de classificação binária, visando isolar e identificar padrões espectrais específicos relacionados à infecção por HDV e outras condições hepáticas: Target A x D, Target B x D, Target C x D e Target A x C.

4.4.2 Resultados de classificação de alto nível de Hepatite B e D com ATLA

A metodologia de pré-processamento espectral, criação de redes complexas, treinamento de classificadores e aplicação do algoritmo híbrido, conforme descrito na Seção 3.1, foi rigorosamente aplicada a todas as comparações nesta base de dados.

A Tabela 6 apresenta o desempenho dos modelos na classificação entre o grupo controle sem vacina BCG (Target A) e o grupo coinfectado com Hepatite B e D (Target D).

Tabela 6 – Resultado base de dados de HDV - Target A x Target D Baseline: Als

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA (RC kNN CCF + SVM)	0.6	**0.83 ± 0.0669**	**0.83 ± 0.0994**	**0.83 ± 0.0657**	0.73 ±0.0637	**0.83 ± 0.0584**
ATLA (RC kNN CCF + LDA)	0.5	0.80 ±0.0556	0.81 ±0.0608	0.83 ±0.0946	0.76 ±0.0372	0.79 ±0.1300
ATLA (kNN ASS)	1.0	0.62 ±0.0267	0.61 ±0.0212	0.60 ±0.1351	0.64 ±0.1155	0.62 ±0.0347
SVM Base	0.0	0.81 ±0.2052	0.81 ±0.2038	0.82 ±0.2045	**0.80 ± 0.2060**	0.80 ±0.2049
LDA Base	0.0	0.79 ±0.1894	0.81 ±0.1950	0.83 ±0.2024	0.76 ±0.1762	0.78 ±0.1851

Os resultados da Tabela 6 indicam que o ATLA (SVM+RC) obteve o melhor desempenho, com um F1-score de 0.83, superando significativamente os classificadores de linha de base. A medida de rede Coeficiente de cluster foi a que mais contribuiu para o ATLA nesta comparação. O gráfico de sensibilidade ao parâmetro λ para esta classificação, gráfico da Figura 18(a) mostra que o desempenho do F1-score se manteve robusto para uma ampla gama de valores de λ, com um pico de performance em $\lambda = 0.5$.

A Tabela 7 sumariza o desempenho dos modelos na classificação entre o grupo controle com vacina BCG (‘Target B‘) e o grupo coinfectado (‘Target D‘).

Tabela 7 – Resultado base de dados de HDV - Target B x Target D Baseline: ArPLS

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA (RC kNN CCF + SVM)	0.3	**0.81 ± 0.0953**	**0.80 ± 0.0993**	**0.80 ± 0.2697**	**0.81 ±0.0991**	**0.81 ± 0.0330**
ATLA (RC kNN CCF + LDA)	0.9	0.75 ±0.1286	0.77 ±0.1258	0.79 ±0.3347	0.71 ±0.0999	0.74 ±0.0511
ATLA (kNN ASS)	1.0	0.67 ±0.0349	0.66 ±0.0260	0.64 ±0.2430	0.70 ±0.1989	0.68 ±0.0524
SVM Base	0.0	0.78 ±0.2125	0.77 ±0.2133	0.76 ±0.2192	0.81 ±0.2050	0.79 ±0.2091
LDA Base	0.0	0.73 ±0.2019	0.75 ±0.2033	0.77 ±0.2115	0.70 ±0.1914	0.72 ±0.1972

Conforme a Tabela 7, o ATLA (SVM+RC) também demonstrou um desempenho superior nesta comparação, alcançando um F1-score de 0.81. A medida "Coeficiente de cluster" novamente se mostrou a mais eficaz. O gráfico de sensibilidade ao λ (segundo gráfico da Figura 18) indica um comportamento similar de robustez, com o pico de F1-score em $\lambda = 0.5$.

A Tabela 8 apresenta os resultados para a classificação entre indivíduos infectados apenas com Hepatite B ('Target C') e os coinfectados ('Target D').

Tabela 8 – Resultado base de dados de HDV - Target C x Target D Baseline: Rubberband

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA (RC kNN ASP + SVM)	0.6	**0.76 ± 0.0139**	**0.76 ± 0.0168**	**0.78 ± 0.0857**	**0.75 ± 0.0892**	**0.75± 0.0376**
ATLA (RC kNN BET + LDA)	0.8	0.67 ±0.0150	0.66 ±0.0207	0.70 ±0.0833	0.63 ±0.0724	0.65 ±0.0251
ATLA (kNN ASS	1.0	0.69 ±0.0288	0.70 ±0.0297	0.75 ±0.0679	0.64 ±0.1307	0.66 ±0.0769
SVM Base	0.0	0.70 ±0.1712	0.70 ±0.1625	0.74 ±0.1588	0.65 ±0.1847	0.67 ±0.1729
LDA Base	0.0	0.66 ±0.1648	0.65 ±0.1622	0.70 ±0.1752	0.62 ±0.1534	0.64 ±0.1577

Nesta classificação mostrada na Tabela 8, o ATLA (SVM+RC) também obteve o melhor F1-score (0.75), embora a melhoria em relação aos modelos de linha de base seja um pouco menor, o que pode refletir a maior sutileza na diferenciação entre HBV e HBV+HDV. A medida "Coeficiente de cluster" manteve-se como a principal contribuinte para a Rede Complexa. O gráfico de sensibilidade ao λ (terceiro gráfico da Figura 18) mostra um comportamento mais instável para altos valores de λ, sugerindo que para esta comparação a contribuição dos classificadores de baixo nível é mais relevante.

Tabela 9 – Resultado base de dados de HDV - Target A x Target C Baseline: Base Normalizada

Algoritmo	λ	Acurácia	Precisão	Especificidade	Sensibilidade	F1-Score
ATLA (RC kNN CCF + SVM)	0.9	0.74 ±0.0027	0.76 ±0.0091	0.75 ±0.0190	0.74 ±0.0175	0.75 ± 0.0046
ATLA (RC kNN ASS + LDA)	0.7	**0.77 ± 0.0256**	**0.79 ± 0.0415**	**0.78 ± 0.1026**	**0.76 ± 0.0790**	**0.77 ± 0.0308**
ATLA (kNN ADG)	1.0	0.51 ±0.0265	0.51 ±0.0255	0.97 ±0.1289	0.74 ±0.1550	0.67 ±0.0923
SVM Base	0.0	0.73 ±0.1967	0.74 ±0.1964	0.71 ±0.1899	0.75 ±0.2029	0.74 ±0.1996
LDA Base	0.0	0.69 ±0.1819	0.71 ±0.1871	0.67 ±0.1858	0.71 ±0.1782	0.70 ±0.1826

Os resultados da Tabela 9 indicam que o ATLA (SVM+RC) manteve a melhor performance (F1-score de 0.77), destacando sua capacidade de diferenciar o grupo controle de indivíduos com Hepatite B. O gráfico de sensibilidade ao λ para esta classificação (Figura 18) reitera a importância de encontrar o λ ótimo para cada cenário de classificação. Em geral os desvios padrões estão bem sólidos com não tanta diferença assim, os métodos bases geralmente possuem um desvio maior sendo diminuídos e melhorados quando aplicado nosso método ATLA.

4.4.3 Análise de Interpretabilidade SHAP para Hepatite B e D

A análise de interpretabilidade utilizando SHAP foi aplicada aos melhores modelos híbridos para cada uma das comparações da base de HDV, visando identificar os *wa-*

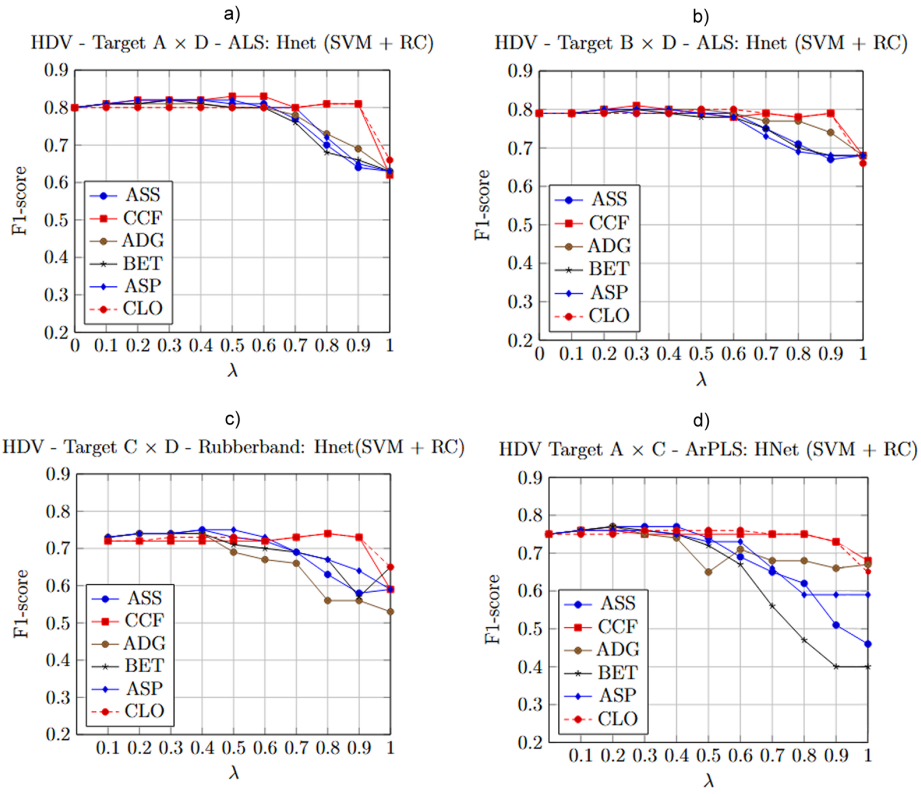


Figura 18 – Desempenho de λ dos melhores resultados do ATLA na base de dados HDV

venumbers e atributos topológicos mais relevantes para as decisões de classificação. As Figuras 19, 20 apresentam os valores SHAP médios para os *wavenumbers* mais importantes em cada cenário de classificação.

A análise dos gráficos SHAP revela a complexidade da diferenciação em HDV, com *wavenumbers* importantes variando significativamente entre as comparações, o que sugere que diferentes aspectos bioquímicos são relevantes para distintas condições de infecção e pessoas vacinadas. No geral, regiões próximas à Amida I ($1630 - 1660\,\mathrm{cm}^{-1}$) e às bandas de estiramento de acima de $2800\,\mathrm{cm}^{-1}$ frequentemente emergem como atributos cruciais, o que corrobora a relevância do pré-processamento focado nessas faixas. Esta análise proporciona maior transparência sobre as decisões do ATLA e aponta para potenciais biomarcadores espectrais específicos para cada diferenciação no contexto da HDV.

4.5 Discussão Geral dos Resultados

Os resultados obtidos em múltiplas bases de dados ATR-FTIR biomédicas, incluindo COVID-19, Chagas e Hepatite B e D demonstram consistentemente que o ATLA, o algoritmo de classificação espectral híbrida de alto nível proposto, é capaz de contribuir na melhoria preditiva de algoritmos de linha de base (SVM, LDA e CNN). Tais melhorias, observadas em métricas como acurácia, precisão, sensibilidade, especificidade e, em especial, F1-score, trazem evidências para validar a hipótese de que a integração sinérgica de

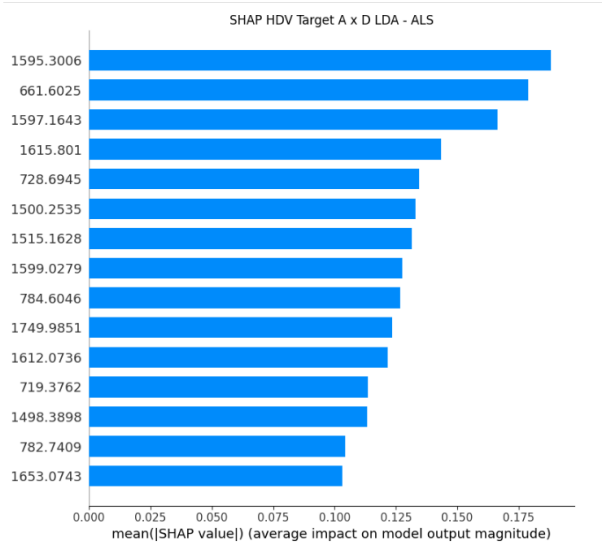


Figura 19 – SHAP do melhor resultado obtido pelo ATLA nos dados de Target A (Grupo controle sem vacina BCG) x Target D (Indivíduos coinfectados com Hepatite B e D) , pertecentes da base de dados de HDV

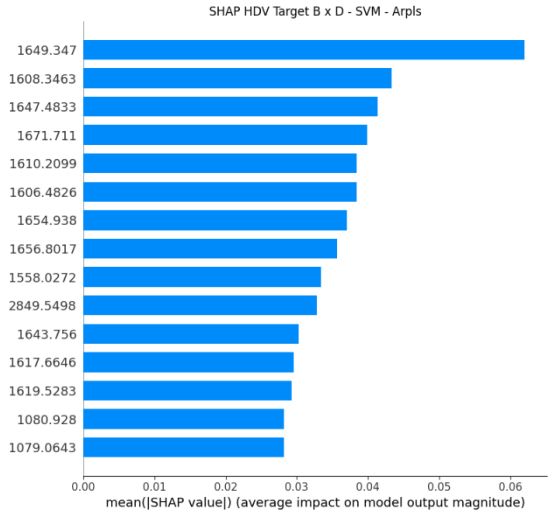


Figura 20 – SHAP do melhor resultado obtido pelo ATLA nos dados de Target B (Grupo controle com vacina BCG) x Target D (Indivíduos coinfectados com Hepatite B e D), pertecentes da base de dados de HDV

atributos topológicos de redes complexas com classificadores de aprendizado profundo e tradicional é capaz de aprimorar a classificação de dados espectrais em cenários desafiadores. O desempenho do ATLA e sua notável capacidade de generalização para diferentes condições clínicas e espectros de alta similaridade confirmam a eficácia de sua arquitetura. Entre as principais vantagens do ATLA, destacam-se:

- **Capturar Relações Complexas:** A integração de Redes Complexas revelou-se eficaz na incorporação de informações relacionais e estruturais globais dos dados, que não são plenamente exploradas por algoritmos de aprendizado profundo isolados. As redes complexas nos permite a representação de dados em rede e extração

de informações topológicas e estruturais da rede, assim, conseguindo nos dar informações das medidas de rede, papel fundamental para a classificação das mesmas. Este aspecto foi crucial para diferenciar classes com sutis variações espectrais.

- **Otimização do Pré-processamento:** A investigação de um protocolo robusto de pré-processamento (incluindo filtro Savitzky-Golay, correção de *baseline* por múltiplos métodos e normalização pela Amida I) foi fundamental para otimizar a qualidade dos espectros, resultando em dados mais limpos e padrões mais evidentes para os classificadores. A variabilidade na performance com diferentes métodos de *baseline* e a observação de que para certas bases (como Chagas) o espectro *Raw* foi mais eficaz, ressalta a importância da adaptabilidade do pré-processamento.

- **Interpretabilidade e *Insights* Clínicos:** A aplicação da análise SHAP forneceu *insights* valiosos sobre os *wavenumbers* e atributos topológicos mais relevantes para a classificação em cada doença. Isso não só aumenta a transparência do modelo híbrido, que é intrinsecamente complexo, mas também oferece direções para a identificação de biomarcadores espectrais potenciais, validando a relevância clínica de regiões como a Amida I e outras faixas de truncamento.

- **Consistência e Otimização de Parâmetros:** A exploração sistemática de parâmetros como o número de vizinhos (k) para a criação das redes complexas e o parâmetro de combinação λ demonstrou a capacidade de ajustar o ATLA para obter o desempenho ótimo em diferentes bases, reforçando sua versatilidade.

Apesar do sucesso, algumas limitações inerentes ao escopo desta dissertação e aos dados disponíveis podem ser apontadas:

- **Tamanho e Diversidade dos Dados:** Embora o ATLA tenha sido validado em múltiplas bases, a quantidade de amostras em cada *dataset* ainda pode ser considerada limitada em um contexto clínico de larga escala. Estudos futuros com bases de dados maiores e mais heterogêneas seriam benéficos para fortalecer ainda mais a generalização.

- **Custo Computacional:** A otimização extensiva de hiperparâmetros (como no caso do CNN-BiLSTM com Optuna) e a criação de redes complexas para cada iteração do treinamento podem implicar um custo computacional elevado, o que deve ser considerado para futuras aplicações em tempo real ou em ambientes com recursos limitados.

- **Variabilidade de uso local:** A padronização de dados ATR-FTIR entre diferentes laboratórios e equipamentos continua sendo um desafio, o que pode impactar a transferibilidade direta dos modelos sem reajustes ou calibrações.

Em suma, as evidências apresentadas, decorrentes dos experimentos rigorosos e da análise detalhada, corroboram a hipótese central desta dissertação: a integração sinérgica de atributos topológicos de RCs com classificadores de aprendizado de máquina (tradicional e profundo) resulta em uma abordagem híbrida que melhora significativamente o desempenho da classificação de dados espectrais ATR-FTIR em múltiplas bases de dados biomédicas desafiadoras. O ATLA oferece uma solução promissora para o diagnóstico assistido por IA, pavimentando o caminho para futuras investigações e aplicações clínicas.

Capítulo 5

Conclusão

A presente pesquisa foi motivada pela crescente demanda por métodos diagnósticos rápidos, precisos e não invasivos para diversas patologias, destacando a espectroscopia ATR-FTIR como uma tecnologia promissora, apesar dos desafios inerentes à complexidade e similaridade espectral dos dados. Diante desse cenário, esta dissertação apresenta o desenvolvimento e avaliação do ATLA, um método inovador de classificação espectral híbrida de alto nível utilizando a topologia estrutural dos espectros para aperfeiçoar o seu desempenho preditivo de classificadores base.

O ATLA foi concebido para integrar sinergicamente a capacidade de aprendizado profundo (utilizando a arquitetura CNN-BiLSTM para extração de características discriminativas) com a análise de Redes Complexas (para modelar relações estruturais globais e de alto nível), visando superar as limitações de modelos tradicionais que frequentemente falham em capturar informações topológicas cruciais em dados complexos. A metodologia empregou um protocolo robusto de pré-processamento espectral (incluindo filtro Savitzky-Golay, correção de *baseline* e normalização pela Amida I), validação cruzada k-Fold e sua interpretabilidade foi aprimorada pela análise via SHAP. O ATLA foi aplicado e extensivamente avaliado em múltiplas bases de dados ATR-FTIR biomédicas, abrangendo condições desafiadoras como COVID-19, Hepatite Delta e doença de Chagas.

5.1 Validação da Hipótese e Principais Contribuições

Conclui-se, portanto, que esta dissertação cumpriu com sucesso seu objetivo principal, validando a hipótese de que a integração sinérgica entre a análise topológica de redes complexas e o aprendizado supervisionado aprimora a classificação de dados espectrais ATR-FTIR. O desenvolvimento e a validação do modelo ATLA não apenas resultaram em um desempenho classificatório superior em múltiplos cenários biomédicos desafiadores, mas também estabeleceram uma nova abordagem robusta e generalizável. As principais contribuições científicas e técnicas desta dissertação concentram-se no desenvolvimento e validação de um modelo híbrido inovador denominado ATLA (ATR-FTIR Topological

Learning Analysis), projetado para classificação espectral de alto nível a partir de dados ATR-FTIR. Este modelo integra representações topológicas derivadas de medidas de Redes Complexas com algoritmos supervisionados (como SVM e LDA) e de aprendizado profundo (CNN-BiLSTM), promovendo uma sinergia entre abordagens que, isoladamente, apresentam limitações na captação de padrões espectrais globais e locais. Além disso,é apresentado a formulação de um protocolo otimizado de pré-processamento espectral, incorporando técnicas como filtro de Savitzky-Golay, diversos métodos de correção de baseline (poly, ALS, ArPLS, DrPLS, Rubberband) e normalização pela banda da Amida I, o que resultou em maior qualidade e clareza dos dados espectrais. Desta forma, a dissertação não apenas atinge seus objetivos propostos, mas também abre novos horizontes para a aplicação multidisciplinar da inteligência artificial na espectroscopia diagnóstica.

5.2 Contribuições em Produção Bibliográfica

As seguintes patentes e publicações, resultantes direta ou indiretamente do trabalho desenvolvido nesta dissertação, foram depositadas ou publicadas, evidenciando o caráter inovador e o potencial de aplicação prática da pesquisa:

❑ **Patentes:**

- **DETECÇÃO DE HEPATITE DELTA BASEADA EM MODOS VIBRACIONAIS INFRAVERMELHOS ACOPLADOS EM ALGORITMO DE INTELIGÊNCIA ARTIFICIAL.** Autores: CAIXETA, D. C.; SILVA, R. S.; SOUZA, R. C.; MARTINS, M. M.; COSTA, M. A.; DIAS, J. C. V. E.; CARNEIRO, M. G.; GUEVARA-VEGA, M.; SANTOS, F. A. A.; GOULART FILHO, L. R.; MUNDIM FILHO, A. C. Tipo: Patente: Privilégio de Inovação. Número de Registro: BR10202302285. Instituição de Registro: INPI - Instituto Nacional da Propriedade Industrial. Data de Depósito: 31/10/2023. País: Brasil.
- **DETECÇÃO SALIVAR DA INFECÇÃO POR CHIKUNGUNYA VÍRUS BASEADA EM MODOS VIBRACIONAIS INFRAVERMELHOS ACOPLADOS COM ALGORITMO DE APRENDIZADO DE MÁQUINA.** Autores: MUNDIM FILHO, A. C.; CARNEIRO, M. G.; DIAS, J. C. V. E.; GUEVARA-VEGA, M.; JARDIM, A. C. G.; MARTINS, M. M.; CUNHA, T. M.; SILVA, R. S.; FERREIRA, G. M.; ROSA, R. B. Tipo: Patente: Privilégio de Inovação. Número de Registro: BR1020240248490. Instituição de Registro: INPI - Instituto Nacional da Propriedade Industrial. Data de Depósito: 28/11/2024. País: Brasil.

❑ **Artigos em Periódicos:**

- **Hybrid Deep Learning and Complex Networks for Enhanced Non-Invasive Covid-19 Diagnosis using Salivary ATR-FTIR Spectroscopy** Autores: MUNDIM FILHO, Anagê C;; SABINO-SILVA, Robinson; CARNEIRO, Murillo G. Ano: 2025. Status: em redação para submissão
- **Salivary detection of Chikungunya virus infection using a portable and sustainable biophotonic platform coupled with artificial intelligence algorithms.** Autores: Guevara-Vega, M., Rosa, R.B., Caixeta, D.C. et al. Ano: 2024. DOI: 10.1038/s41598-024-71889-z URL: <https://doi.org/10.1038/s41598-024-71889-z>. Status: Publicado na Nature

❑ **Artigos em Conferência:**

- **OCANSpectra: um sistema de detecção de câncer oral a partir da espectroscopia salivar ATR-FTIR.** Autores: MUNDIM FILHO, Anagê C.; FERNANDES, Janayna M.; SABINO-SILVA, Robinson; CARNEIRO, Murillo G. Ano: 2023. Conferência: ENIAC (Encontro Nacional de Inteligência Artificial e Computacional). Nota: Publicado nos Anais do evento.
- **Convolutional Neural Networks for the Molecular Detection of COVID-19.** Autores: SANTOS, Anisio P. Jr.; MUNDIM FILHO, Anage C.; SABINO-SILVA, Robinson; CARNEIRO, Murillo G. Ano: 2023. Em: Tuba, M.; Akter, S.; Akter, S.; Akter, S. (Eds.). *Deep Learning for COVID-19: Diagnosis and Drug Discovery*. Springer Cham. DOI: 10.1007/978-3-031-45389-2 4. URL: <https://doi.org/10.1007/978-3-031-45389-2_4>.

5.3 Limitações e Trabalhos Futuros

Esta dissertação demonstrou o potencial transformador de uma abordagem híbrida que une o poder das redes neurais profundas à riqueza estrutural das redes complexas para o diagnóstico espectral. O ATLA não apenas oferece uma ferramenta de classificação de alto desempenho para cenários biomédicos complexos, mas também enfatiza a importância da interpretabilidade em IA, pavimentando o caminho para o desenvolvimento de soluções mais transparentes, confiáveis e impactantes na saúde. Os resultados obtidos representam um avanço significativo na área, abrindo novas perspectivas para o diagnóstico não invasivo e o monitoramento de diversas doenças. Apesar dos avanços significativos alcançados com o desenvolvimento do ATLA, algumas limitações ainda se fazem presentes. Em primeiro lugar, embora o modelo tenha sido validado em múltiplas bases de dados relevantes, o número de amostras por conjunto ainda é relativamente reduzido frente ao contexto clínico real, o que pode limitar sua generalização em larga escala. Além disso, o processo de otimização intensiva de hiperparâmetros, especialmente no caso

do CNN-BiLSTM com o uso do Optuna, aliado à construção recorrente de redes complexas durante o treinamento, acarreta um custo computacional elevado. Tal limitação pode comprometer a viabilidade do uso do ATLA em aplicações em tempo real ou em ambientes com recursos restritos. Soma-se a isso a variabilidade inter-laboratorial inerente aos dados espectrais ATR-FTIR, que representa um desafio adicional à reprodutibilidade e à transferibilidade direta dos modelos entre diferentes dispositivos ou centros de coleta.

Diante dessas limitações, diversas perspectivas de trabalhos futuros se mostram promissoras. Uma das principais direções é a ampliação da avaliação do ATLA em bases de dados mais robustas, incluindo amostras de diferentes patologias, populações e etnias, a fim de fortalecer sua aplicabilidade clínica. Também é relevante investigar abordagens que reduzam o custo computacional do modelo, como técnicas de poda de redes neurais, otimizações nos algoritmos de construção das redes complexas e métodos de seleção de atributos mais agressivos. A extensão do modelo para outros biofluídos, como sangue e urina, bem como para diferentes técnicas espectroscópicas, como espectroscopia Raman e NIR, pode ampliar ainda mais seu escopo. Do ponto de vista metodológico, o desenvolvimento de redes complexas adaptativas ou autoajustáveis pode dispensar a definição manual de parâmetros como o número de vizinhos e a métrica de distância. Além disso, reforça-se a importância de aprofundar a interpretabilidade dos modelos por meio de ferramentas como SHAP, em colaboração com especialistas clínicos, visando identificar biomarcadores espectrais relevantes. Uma outra limitação, foi a falta de realização de uma investigação sistemática de heurísticas e modelos de construção e otimização de redes, a qual representa uma oportunidade para trabalhos futuros. Por fim, destaca-se a possibilidade de integrar o ATLA em sistemas clínicos reais, viabilizando sua aplicação prática como ferramenta de apoio ao diagnóstico médico.

Referências

ALAJAJI, S. A. et al. A scoping review of infrared spectroscopy and machine learning methods for head and neck precancer and cancer diagnosis and prognosis. **Cancers**, MDPI, v. 17, n. 5, p. 796, 2025.

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2020.

ALYASSERI, Z. A. A. et al. Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. **Expert Systems**, v. 39, n. 3, p. e12759, 2022.

BAEK, S.-J. et al. Baseline correction using asymmetrically reweighted penalized least squares smoothing. **Analyst**, Royal Society of Chemistry, v. 140, n. 1, p. 250–257, 2015.

BAKER, M. J. et al. Using Fourier transform FTIR spectroscopy to analyze biological materials. **Nature protocols**, Nature Publishing Group, v. 9, n. 8, p. 1771–1791, 2014.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4-5, p. 175–308, 2006.

BOELENS, H. F.; EILERS, P. H.; HANKEMEIER, T. Sign constraints improve the detection of differences between complex spectral data sets: Lc- ir as an example. **Analytical chemistry**, ACS Publications, v. 77, n. 24, p. 7998–8007, 2005.

BOZKURT, O.; SEVERCAN, M.; SEVERCAN, F. Diabetes induces compositional, structural and functional alterations on rat skeletal soleus muscle revealed by ftir spectroscopy: a comparative study with edl muscle. **Analyst**, Royal Society of Chemistry, v. 135, n. 12, p. 3110–3119, 2010.

BUNACIU, A. A.; HOANG, V. D.; ABOUL-ENEIN, H. Y. Applications of FTIR spectrophotometry in cancer diagnostics. **Critical reviews in analytical chemistry**, Taylor & Francis, v. 45, n. 2, p. 156–165, 2015.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, Springer, v. 2, n. 2, p. 121–167, 1998.

BUTLER, H. J. et al. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. **Nature communications**, Nature Publishing Group, v. 10, n. 1, p. 1–9, 2019.

BYLER, D. M.; SUSI, H. Examination of the secondary structure of proteins by deconvolved FTIR spectra. **Biopolymers: Original Research on Biomolecules**, Wiley Online Library, v. 25, n. 3, p. 469–487, 1986.

CAIXETA, D. C. et al. Salivary ATR-FTIR spectroscopy coupled with support vector machine classification for screening of type 2 diabetes mellitus. **Diagnostics**, MDPI, v. 13, n. 8, p. 1396, 2023. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/13/8/1396>.

CARNEIRO, M.; ZHAO, L. Analysis of graph construction methods in supervised data classification. In: IEEE. **2018 7th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.], 2018. p. 390–395.

CARNEIRO, M. G. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. Tese (Doutorado) — Universidade de São Paulo, 2017.

CARNEIRO, M. G. et al. Particle swarm optimization for network-based data classification. **Neural Networks**, Elsevier, v. 110, p. 243–255, 2019.

CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on the importance concept of complex networks. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, n. 8, p. 3361–3373, 2018.

CIOTTI, M. et al. The COVID-19 pandemic. **Critical Reviews in Clinical Laboratory Sciences**, Taylor & Francis, v. 57, n. 6, p. 365–388, 2020. ISSN 1040-8363.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, p. 273–297, 1995.

CUI, P. et al. A survey on network embedding. **IEEE transactions on knowledge and data engineering**, IEEE, v. 31, n. 5, p. 833–852, 2018.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: **Machine learning techniques for multimedia**. [S.l.]: Springer, 2008. p. 21–49.

DELRUE, C.; BRUYNE, S. D.; SPEECKAERT, M. M. The promise of infrared spectroscopy in liquid biopsies for solid cancer detection. **Diagnostics**, MDPI, v. 15, n. 3, p. 368, 2025.

FERNANDES, J. M.; SABINO-SILVA, R.; CARNEIRO, M. G. **Network-Based Detection of Autism Spectrum Disorder Using Sustainable and Non-invasive Salivary Biomarkers**. 2025. Disponível em: <https://arxiv.org/abs/2509.16126>.

Luiz Ricardo Goulart Filho, Thulio Marquez Cunha, Mário Machado Martins, Léia Cardoso de Sousa, Paula de Souza Santos, Thays Crosara Abrahão Cunha, Emília Rezende Vaz, Tatiane Martins de Lima Crosara Bastos, Arlene Bispo dos Santos Nossol, Anderson Rodrigues dos Santos, Luciana Machado Bastos, Robinson Sabino Silva e Murillo Guimarães Carneiro. **Perfil espectral para diagnóstico de COVID-19, uso do mesmo, método, sistema e plataforma de diagnóstico de COVID-19**. 2021. WO2021237330A1. Disponível em: <https://patents.google.com/patent/WO2021237330A1/en>.

FILHO, R. B. L. et al. High-Level Network-based Detection of Oral Cancer from ATR-FTIR Spectroscopy. In: **2024 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2024. p. 1–8.

FILHO, R. B. L.; SABINO-SILVA, R.; CARNEIRO, M. G. High-level classification based on meta-graph of visibility graphs for autism detection. In: **2025 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2025. p. 1–8.

FREEMAN, L. C. et al. Centrality in social networks: Conceptual clarification. **Social network: critical concepts in sociology. Londres: Routledge**, v. 1, n. 3, p. 238–263, 2002.

GARCIA-JUNIOR, M. A. et al. Artificial-Intelligence Bio-Inspired Peptide for Salivary Detection of SARS-CoV-2 in Electrochemical Biosensor Integrated with Machine Learning Algorithms. **Biosensors**, v. 15, n. 2, 2025. ISSN 2079-6374. Disponível em: <https://www.mdpi.com/2079-6374/15/2/75>.

GIAMOUGIANNIS, P. et al. 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**, Springer, v. 413, n. 20, p. 5095–5107, 2021.

GOODFELLOW, I. et al. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1.

GOYAL, P.; FERRARA, E. Graph embedding techniques, applications, and performance: A survey. **Knowledge-Based Systems**, Elsevier, v. 151, p. 78–94, 2018.

GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. In: **Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 855–864.

HONÓRIO-SILVA, G. et al. Development of a novel sustainable, portable, fast, and non-invasive platform based on ATR-FTIR technology coupled with machine learning algorithms for Helicobacter pylori detection in human saliva. **Talanta Open**, Elsevier, v. 10, p. 100383, 2024.

JIANG, S. et al. Using ATR-FTIR spectra and convolutional neural networks for characterizing mixed plastic waste. **Computers & Chemical Engineering**, Elsevier, v. 155, p. 107547, 2021.

JUNIOR, A. P. S. et al. **CNN-BiLSTM for sustainable and non-invasive COVID-19 detection via salivary ATR-FTIR spectroscopy**. 2025. Disponível em: <https://arxiv.org/abs/2509.12241>.

KAZARIAN, S. G.; CHAN, K. A. ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. **Analyst**, Royal Society of Chemistry, v. 138, n. 7, p. 1940–1951, 2013.

KIRANYAZ, S.; INCE, T.; GABBOUJ, M. Real-time patient-specific ecg classification by 1-d convolutional neural networks. **IEEE transactions on biomedical engineering**, IEEE, v. 63, n. 3, p. 664–675, 2015.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KROHLING, B. A.; KROHLING, R. A. 1D Convolutional neural networks and machine learning algorithms for spectral data classification with a case study for COVID-19. **arXiv preprint arXiv:2301.10746**, 2023.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y.; BENGIO, Y. et al. Convolutional networks for images, speech, and time series. **The handbook of brain theory and neural networks**, Citeseer, v. 3361, n. 10, p. 1995, 1995.

LOSQ, C. L. **Rampy: a Python library for processing spectroscopic (IR, Raman, XAS...) data.** Zenodo, 2018. Disponível em: <https://doi.org/10.5281/zenodo.4715040>.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MADEJOVÁ, J. Ftir techniques in clay mineral studies. **Vibrational spectroscopy**, Elsevier, v. 31, n. 1, p. 1–10, 2003.

MALAMUD, D. Saliva as a diagnostic fluid. **Dental Clinics**, Elsevier, v. 55, n. 1, p. 159–178, 2011.

MANSUR, H. S. et al. Ftir spectroscopy characterization of poly (vinyl alcohol) hydrogel with different hydrolysis degree and chemically crosslinked with glutaraldehyde. **Materials Science and Engineering: C**, Elsevier, v. 28, n. 4, p. 539–548, 2008.

MARTINEZ, A. M.; KAK, A. C. PCA versus LDA. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 23, n. 2, p. 228–233, 2001.

MARTINEZ-CUAZITL, A. et al. ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients. **Scientific Reports**, Nature Publishing Group, v. 11, n. 1, p. 1–14, 2021.

MENEZES, M. E.; LIMA, L. M.; MARTINELLO, F. Diagnóstico laboratorial do SARS-CoV-2 por transcrição reversa seguida de reação em cadeia da polimerase em tempo real (RT-PCR). **A Tempestade do Coronavírus. Rev RBAC**, v. 52, n. 2, p. 122–30, 2020.

MITCHELL, T. M.; MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-hill New York, 1997. v. 1.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, Manole Ltda, v. 1, n. 1, p. 32, 2003.

MORAIS, C. L. et al. Standardization of complex biologically derived spectrochemical datasets. **Nature protocols**, Nature Publishing Group, v. 14, n. 5, p. 1546–1577, 2019.

MOVASAGHI, Z.; REHMAN, S.; REHMAN, D. I. ur. Fourier transform infrared (ftir) spectroscopy of biological tissues. **Applied Spectroscopy Reviews**, Taylor & Francis, v. 43, n. 2, p. 134–179, 2008.

NASEER, K.; ALI, S.; QAZI, J. ATR-FTIR spectroscopy as the future of diagnostics: a systematic review of the approach using bio-fluids. **Applied Spectroscopy Reviews**, Taylor & Francis, v. 56, n. 2, p. 85–97, 2021.

NOGUEIRA, M. S. et al. FTIR spectroscopy as a point of care diagnostic tool for diabetes and periodontitis: A saliva analysis approach. **Photodiagnosis and Photodynamic Therapy**, v. 40, p. 103036, 2022. ISSN 1572-1000. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1572100022003222>.

OLIVEIRA, S. W. et al. 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**, MDPI, v. 13, n. 8, p. 1443, 2023.

PENG, J. et al. Baseline correction combined partial least squares algorithm and its application in on-line fourier transform infrared quantitative analysis. **Analytica chimica acta**, Elsevier, v. 690, n. 2, p. 162–168, 2011.

PENG, X. et al. Early diagnosis and bioimaging of lung adenocarcinoma cells/organs based on spectroscopy machine learning. **Journal of Innovative Optical Health Sciences**, World Scientific, v. 15, n. 02, p. 2250011, 2022.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. In: **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2014. p. 701–710.

PÓSFAI, M.; BARABASI, A.-L. **Network Science**. [S.l.]: Citeseer, 2016.

PRESS, W. H.; TEUKOLSKY, S. A. Savitzky-golay smoothing filters. **Computers in Physics**, American Institute of Physics, v. 4, n. 6, p. 669–672, 1990.

RALBOVSKY, N. M.; LEDNEV, I. K. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. **Chemical Society Reviews**, Royal Society of Chemistry, v. 49, n. 20, p. 7428–7453, 2020.

RESENDE, V. H.; CARNEIRO, M. G. High-level classification for multi-label learning. In: IEEE. **2020 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2020. p. 1–8.

_____. Analysis of complex network measures for multi-label classification. **International Journal on Artificial Intelligence Tools**, World Scientific, v. 30, n. 04, p. 2150023, 2021.

ROHMAN, A. et al. The use of FTIR and Raman spectroscopy in combination with chemometrics for analysis of biomolecules in biomedical fluids: A review. **Biomedical Spectroscopy and Imaging**, SAGE Publications Sage UK: London, England, v. 8, n. 3-4, p. 55–71, 2019.

SALA, A. et al. Biofluid diagnostics by ftir spectroscopy: A platform technology for cancer detection. **Cancer letters**, Elsevier, v. 477, p. 122–130, 2020.

SANTOS, A. P. et al. Convolutional neural networks for the molecular detection of covid-19. In: NALDI, M. C.; BIANCHI, R. A. C. (Ed.). **Intelligent Systems**. Cham: Springer Nature Switzerland, 2023. p. 51–62. ISBN 978-3-031-45389-2.

SANTOS, M. C.; MORAIS, C. L.; LIMA, K. M. ATR-FTIR spectroscopy for virus identification: A powerful alternative. **Biomedical Spectroscopy and Imaging**, v. 9, n. 3-4, p. 103–118, 2020.

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.

SCHMITT, J.; FLEMMING, H.-C. FTIR-spectroscopy in microbial and material analysis. **International Biodeterioration & Biodegradation**, Elsevier, v. 41, n. 1, p. 1–11, 1998.

SILVA, L. G. et al. ATR-FTIR spectroscopy in blood plasma combined with multivariate analysis to detect HIV infection in pregnant women. **Scientific reports**, Nature Publishing Group, v. 10, n. 1, p. 1–7, 2020.

SILVA, T. C.; ZHAO, L. Network-based high level data classification. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 6, p. 954–970, 2012.

SINGH, P. et al. Chapter 5 - diagnosing of disease using machine learning. In: SINGH, K. K. et al. (Ed.). **Machine Learning and the Internet of Medical Things in Healthcare**. Academic Press, 2021. p. 89–111. ISBN 978-0-12-821229-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128212295000033>.

SITNIKOVA, V. E. et al. Breast cancer detection by atr-ftir spectroscopy of blood serum and multivariate data-analysis. **Talanta**, Elsevier, v. 214, p. 120857, 2020.

SOUSA, L. C. de. **Desenvolvimento de plataforma biofotônica de larga escala para aplicação na triagem diagnóstica da COVID-19 por meio da saliva**. Tese (Tese de Doutorado) — Universidade Federal de Uberlândia, Uberlândia, MG, 2022.

STROGATZ, S. H. Exploring complex networks. **nature**, Nature Publishing Group UK London, v. 410, n. 6825, p. 268–276, 2001.

SURYASA, I. W.; Rodríguez-Gámez, M.; KOLDORIS, T. The COVID-19 pandemic. **International Journal of Health Sciences**, v. 5, n. 2, p. 572194, 2021. ISSN 2550-6978. Article ID 572194. Disponível em: <https://doi.org/10.53730/ijhs.v5n2.572194>.

SUTTON, R. S.; BARTO, A. G. et al. **Reinforcement learning: An introduction**. [S.l.]: MIT press Cambridge, 1998. v. 1.

TANG, J. et al. Line: Large-scale information network embedding. In: **Proceedings of the 24th international conference on world wide web**. [S.l.: s.n.], 2015. p. 1067–1077.

TEKLEMARIAM, Z. et al. Deep learning for FTIR-based disease classification: A comprehensive review. **Expert Systems with Applications**, v. 238, p. 124771, 2024.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

XANTHOPOULOS, P. et al. Linear discriminant analysis. **Robust data mining**, Springer, p. 27–33, 2013.

XU, D. et al. Baseline correction method based on doubly reweighted penalized least squares. **Applied optics**, Optica Publishing Group, v. 58, n. 14, p. 3913–3920, 2019.

ZHANG, L. et al. Fast screening and primary diagnosis of covid-19 by atr–ft-ir. **Analytical chemistry**, ACS Publications, v. 93, n. 4, p. 2191–2199, 2021.

ZHANG, W. et al. On-site precise screening of sars-cov-2 systems using a channel-wise attention-based pls-1d-cnn model with limited infrared signatures. **arXiv preprint arXiv:2410.20132**, 2024.

ZHAO, L.; SILVA, T. C.; CARNEIRO, M. G. **Classificação de Dados de Alto Nível em Redes Complexas**. São Paulo, SP: [s.n.], 2021. 181–205 p.