Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

THIAGO FRANÇA DA SILVA

“Inovação no diagnóstico de brucelose e tuberculose bovina: métodos de fotodiagnóstico baseados em espectroscopia FTIR e machine learning”

Campo Grande – MS
01/2025

Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

"Inovação no Diagnóstico de Brucelose e Tuberculose Bovina: Métodos de Fotodiagnóstico Baseados em Espectroscopia FTIR e Machine Learning"

THIAGO FRANÇA DA SILVA

Orientador: *Prof. Dr. Cícero Rafael Cena da Silva*

Tese apresentada à Fundação Universidade Federal de Mato Grosso do Sul – UFMS, Instituto de Física - INFI, para obtenção do título de Doutor em Ciência dos Materiais.

Campo Grande – MS
01/2025

INSTITUTO DE FÍSICA- INFI
Cidade Universitária | Unidade 5 |
Fone 67 3345 7485
79070-900 | Campo Grande | MS

Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul
Coordenadoria de Pós-Graduação (CPG/PROPP)

Ata de Defesa de Tese
Programa de Pós-Graduação em Ciência dos Materiais
Doutorado

Aos dezessete dias do mês de janeiro do ano de dois mil e vinte e cinco, às treze horas e trinta minutos, na Videoconferência, da Fundação Universidade Federal de Mato Grosso do Sul, reuniu-se a Banca Examinadora composta pelos membros: Cicero Rafael Cena da Silva (UFMS), Carlos Alberto do Nascimento Ramos (UFMS), Diogo Duarte dos Reis (UFMS), Flábio Ribeiro de Araújo (Embrapa/CNPGC) e Giselle Maria Rachid Viana (UFPA), sob a presidência do primeiro, para julgar o trabalho do aluno: **THIAGO FRANÇA DA SILVA**, CPF ***.317.851-**, do Programa de Pós-Graduação em Ciência dos Materiais, Curso de Doutorado, da Fundação Universidade Federal de Mato Grosso do Sul, apresentado sob o título **"Inovação no Diagnóstico de Brucelose e Tuberculose Bovina: Métodos de Fotodiagnóstico Baseados em Espectroscopia FTIR e Machine Learning"** e orientação de Cicero Rafael Cena da Silva. O presidente da Banca Examinadora declarou abertos os trabalhos e agradeceu a presença de todos os Membros. A seguir, concedeu a palavra ao aluno que expôs sua Tese. Terminada a exposição, os senhores membros da Banca Examinadora iniciaram as arguições. Terminadas as arguições, o presidente da Banca Examinadora fez suas considerações. A seguir, a Banca Examinadora reuniu-se para avaliação, e após, emitiu parecer expresso conforme segue:

EXAMINADOR	ASSINATURA	AVALIAÇÃO
Dr. Cicero Rafael Cena da Silva (Interno)	Documento assinado digitalmente gov.br CICERO RAFAEL CENA DA SILVA Data: 17/01/2025 17:55:21-0300 Verifique em https://validar.iti.gov.br	Aprovado
Dr. Bruno Spolon Marangoni (Interno) (Suplente)	Documento assinado digitalmente gov.br DIOGO DUARTE DOS REIS Data: 17/01/2025 18:50:13-0300 Verifique em https://validar.iti.gov.br	
Dr. Carlos Alberto do Nascimento Ramos (Externo)	Documento assinado digitalmente gov.br CARLOS ALBERTO DO NASCIMENTO RAMOS Data: 17/01/2025 20:55:47-0300 Verifique em https://validar.iti.gov.br	Aprovado
Dr. Diogo Duarte dos Reis (Interno)		Aprovado
Dr. Flábio Ribeiro de Araújo (Externo)		
Dra. Giselle Maria Rachid Viana (Externo)	Documento assinado digitalmente gov.br GISELLE MARIA RACHID VIANA Data: 20/01/2025 19:21:34-0300 Verifique em https://validar.iti.gov.br	Aprovado
Dra. Natalia Oliveira Alves (Externo) (Suplente)	Documento assinado digitalmente gov.br NATALIA OLIVEIRA ALVES Data: 17/01/2025 19:46:52-0300 Verifique em https://validar.iti.gov.br	Aprovado

RESULTADO FINAL:

[X] Aprovação [ ] Aprovação com revisão [ ] Reprovação

OBSERVAÇÕES:

Nada mais havendo a ser tratado, o Presidente declarou a sessão encerrada e agradeceu a todos pela presença.

Assinaturas:

Documento assinado digitalmente gov.br CICERO RAFAEL CENA DA SILVA Data: 21/01/2025 08:06:46-0300 Verifique em https://validar.iti.gov.br

Presidente da Banca Examinadora

Documento assinado digitalmente gov.br THIAGO FRANCA DA SILVA Data: 21/01/2025 09:18:04-0300 Verifique em https://validar.iti.gov.br

Aluno

AGRADECIMENTOS

O Presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo: 140506/2021-7, Título: CH MAI/DAI 2020 - Chamada CNPq Nº 12/2020 - Programa de Mestrado e Doutorado Acadêmico para Inovação - MAI/DAI.

RESUMO

Brucelose e tuberculose bovina são zoonoses com grande impacto na saúde pública, saúde animal e economia. Causadas por bactérias dos gêneros *Brucella* e *Mycobacterium*, respectivamente, são transmitidas aos humanos principalmente por produtos de origem animal contaminados ou pelo contato direto com animais infectados. Este estudo avaliou o uso da espectroscopia FTIR (Infravermelho por Transformada de Fourier) associada ao aprendizado de máquina como ferramenta de triagem para o diagnóstico das infecções, utilizando amostras de soro sanguíneo bovino. O estudo investigou a influência das condições de preparação das amostras no desempenho da classificação, dividindo-se em três etapas. Inicialmente, foram comparadas amostras de um grupo Controle e grupo Brucelose (animais comprovadamente infectados pela bactéria *Brucella abortus*) utilizando duas abordagens: soro seco em estufa, método consolidado na literatura, e soro líquido com *background* em água deionizada, uma opção inovadora, prática e rápida. Em seguida, com amostras de soro líquido, foram realizadas classificações binárias entre Controle x Tuberculose (animais comprovadamente infectados pela bactéria *Mycobacterium bovis*) e Controle x Brucelose, permitindo a validação e otimização do modelo, preparando-o para a classificação multiclasse. Por fim, a classificação multiclasse foi aplicada para distinguir simultaneamente os três grupos (Controle x Brucelose x Tuberculose). Os resultados mostraram que o soro líquido teve desempenho superior ao soro seco, com acurácia e sensibilidade de 100% no diagnóstico de brucelose, superando métodos convencionais. Embora a acurácia para tuberculose tenha sido de 83,3%, a abordagem multiclasse alcançou 90,5% de acurácia e sensibilidade de até 100%, destacando a eficácia do método na diferenciação entre animais controle e infectados. A análise revelou a contribuição conjunta dos modos vibracionais de moléculas de diferentes grupos, como lipídios, proteínas e carboidratos. A combinação de espectroscopia FTIR com aprendizado de máquina, utilizando soro sanguíneo líquido com *background* em água deionizada, mostrou-se um método inovador, rápido e eficiente, dispensando etapas complexas de preparação de amostras, com potencial para diagnóstico *in loco* e controle dessas zoonoses, reduzindo sua disseminação entre animais e humanos.

Palavras-chave: Brucelose bovina, Tuberculose bovina, Espectroscopia FTIR, Aprendizado de máquina, Diagnóstico *in loco*.

ABSTRACT

Bovine brucellosis and tuberculosis are zoonotic diseases with significant impacts on public health, animal health, and the economy. Caused by bacteria of the genera *Brucella* and *Mycobacterium*, respectively, these diseases are primarily transmitted to humans through contaminated animal-derived products or direct contact with infected animals. This study evaluated the use of Fourier Transform Infrared (FTIR) spectroscopy combined with machine learning as a screening tool for diagnosing these infections using bovine blood serum samples. The study investigated the influence of sample preparation conditions on classification performance, divided into three stages. Initially, samples from a Control group and a Brucellosis group (animals confirmed to be infected with *Brucella abortus*) were compared using two approaches: oven-dried serum, a well-established method in the literature, and liquid serum with a deionized water background, an innovative, practical, and rapid alternative. Subsequently, with liquid serum samples, binary classifications were conducted between Control vs. Tuberculosis (animals confirmed to be infected with *Mycobacterium bovis*) and Control vs. Brucellosis, enabling model validation and optimization for multiclass classification. Finally, multiclass classification was applied to simultaneously distinguish among the three groups (Control vs. Brucellosis vs. Tuberculosis). The results demonstrated that liquid serum outperformed dried serum, achieving 100% accuracy and sensitivity in diagnosing brucellosis, surpassing conventional methods. Although the accuracy for tuberculosis was 83.3%, the multiclass approach achieved 90.5% accuracy and up to 100% sensitivity, underscoring the method's effectiveness in differentiating between control and infected animals. The analysis revealed the joint contribution of vibrational modes from molecules belonging to different groups, such as lipids, proteins, and carbohydrates. The combination of FTIR spectroscopy with machine learning, using liquid blood serum with a deionized water background, proved to be an innovative, rapid, and efficient method, eliminating complex sample preparation steps. This approach holds potential for on-site diagnostics and the control of these zoonoses, reducing their spread among animals and humans.

Keywords: Bovine brucellosis, Bovine tuberculosis, FTIR spectroscopy, Machine learning, On-site diagnosis.

LISTA DE FIGURAS

LISTA DE ACRÔNIMOS E SIGLAS

AAT – Antígeno Acidificado Tamponado

AD – Água deionizada

ATR – Reflectância Total Atenuada (*Attenuated Total Reflectance*)

CFT – Teste de Fixação do Complemento (*Complement Fixation Test*)

DRIFTS – Refletância Difusa

ELISA – Ensaio Imunoenzimático (*Enzyme-Linked Immunosorbent Assay*)

FIR – Infravermelho Distante (*Far Infrared*)

FN – Falso Negativo

FP – Falso Positivo

FPA – Polarização de Fluorescência (*Fluorescence Polarization Assay*)

FTIR – Infravermelho por Transformada de Fourier (*Fourier Transform Infrared*)

iELISA - Ensaio Imunoenzimático Indireto (*Indirect Enzyme-Linked Immunosorbent Assay*)

IR – Infravermelho (*Infrared*)

LFDA – Laboratório Federal de Defesa Agropecuária

LOOCV – Validação Cruzada de Deixar-Um-Fora (*Leave-One-Out Cross-Validation*)

MIR – Infravermelho Médio (*Mid Infrared*)

NIR – Infravermelho Próximo (*Near Infrared*)

NIRS – Espectroscopia no Infravermelho Próximo (*Near Infrared Reflectance Spectroscopy*)

PAS – Espectroscopia Fotoacústica (*Photoacoustic Spectroscopy*)

PC's – Componentes Principais (*Principal Components*)

PCA – Análise de Componentes Principais (*Principal Component Analysis*)

PNCEBT – Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose

RBF – Função de Base Radial (*Radial Basis Function*)

RBT – Teste Rose Bengal (*Rose Bengal Test*)

SAL – Soroaglutinação Lenta

SG – Savitzky-Golay

SNR – Relação Sinal-Ruído (*Signal-to-Noise Ratio*)

SNV – Variação Normal Padrão (*Standard Normal Variate*)

SVM – Máquina de Vetores de Suporte (*Support Vector Machine*)

TAL – Teste do Anel em Leite

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

ZnSe – Seleneto de Zinco

SUMÁRIO

1. INTRODUÇÃO

A brucelose bovina é uma doença infectocontagiosa causada principalmente pela bactéria *Brucella abortus*, que afeta bovinos e bubalinos. Essa doença compromete a eficiência reprodutiva dos animais, levando a perdas econômicas significativas, associadas a problemas como abortos no final da gestação (1). Além disso, a brucelose bovina é uma zoonose, podendo ser transmitida ao homem através do consumo de leite e carne contaminados ou pelo contato direto com animais infectados (2).

O diagnóstico da brucelose bovina é realizado por meio da detecção de anticorpos específicos contra *Brucella abortus* no soro sanguíneo, utilizando uma combinação de testes de triagem e confirmatórios (3). No Brasil, os testes mais comuns são o Teste do Anel em Leite (TAL) e o teste de Antígeno Acidificado Tamponado (AAT), enquanto os testes confirmatórios incluem o teste de Fixação do Complemento (FC), entre outros (4).

A tuberculose bovina, causada principalmente por *Mycobacterium bovis*, é outra zoonose relevante, com impactos semelhantes em termos de saúde pública e na economia (5). A transmissão para os humanos ocorre principalmente através do consumo de produtos lácteos não pasteurizados (6) ou contato direto com animais infectados (7). O diagnóstico é comumente realizado por meio de testes intradérmicos com tuberculina, mas esses métodos apresentam limitações de sensibilidade e especificidade, dificultando o controle da doença nos rebanhos (8).

Uma alternativa promissora para o diagnóstico de brucelose e tuberculose bovina é a espectroscopia no Infravermelho por Transformada de Fourier (FTIR), combinada com aprendizado de máquina (9). Essa técnica permite uma análise rápida e eficiente do soro sanguíneo, sem a necessidade de preparação complexa de amostras (10). O procedimento consolidado na literatura para análise de soro sanguíneo envolve a deposição da amostra em substrato e secagem em estufa, o que elimina a interferência da água, mas pode introduzir viés devido a variações de temperatura ou à distribuição não homogênea dos componentes (11). O objetivo deste estudo é avaliar o potencial do uso de soro líquido com *background* em água deionizada (12), uma abordagem que visa melhorar a precisão e reduzir o tempo de diagnóstico, superando as limitações do método tradicional.

2. OBJETIVOS

2.1 Objetivo Geral

Avaliar o potencial da espectroscopia FTIR (infravermelho por transformada de Fourier) associada ao aprendizado de máquina como ferramenta de triagem no diagnóstico de brucelose e tuberculose bovina, utilizando amostras de soro sanguíneo. O estudo visa desenvolver um método diagnóstico rápido e eficaz, contribuindo para o controle da disseminação dessas zoonoses.

2.2 Objetivos Específicos

Comparar o desempenho da espectroscopia FTIR associada ao aprendizado de máquina na classificação de amostras do grupo controle e brucelose bovina, utilizando soro seco (obtido por desidratação em estufa) e soro líquido (com *background* em água deionizada);

Avaliar o desempenho da espectroscopia FTIR associada ao aprendizado de máquina na classificação de amostras de soro líquido com *background* em água deionizada para o diagnóstico de brucelose e tuberculose bovina, de forma isolada, utilizando classificadores binários;

Avaliar o desempenho da espectroscopia FTIR associada ao aprendizado de máquina na classificação de amostras de soro líquido com *background* em água deionizada para o diagnóstico de brucelose e tuberculose bovina, utilizando classificadores multiclasse. Demonstrar a aplicabilidade prática do método proposto, comparando os resultados obtidos com os métodos tradicionais considerados padrão-ouro.

3. REVISÃO BIBLIOGRÁFICA

3.1 Brucelose e Tuberculose Bovina

A brucelose bovina, causada principalmente pela bactéria *Brucella abortus*, é uma doença zoonótica que representa sérios riscos à saúde tanto do gado quanto dos humanos, ao mesmo tempo em que causa perdas econômicas consideráveis ao setor agrícola. A doença é caracterizada por falhas reprodutivas no gado, incluindo abortos, natimortos e infertilidade, o que pode impactar severamente a produtividade das operações de gado leiteiro e de corte (13,14,15).

As implicações econômicas da brucelose bovina são profundas, com estimativas de 2013 indicando que o impacto anual no Brasil atinge cerca de R$ 892 milhões, em grande parte devido à diminuição da produção de leite, aos custos com intervenções veterinárias e ao abate de animais infectados (13,15). Os impactos da brucelose bovina na saúde vão além da pecuária, pois a doença também é transmissível aos humanos, causando sintomas como febre, sudorese e dor musculoesquelética (16,17). Essa transmissão zoonótica ressalta a importância de desenvolvimento e implementação de medidas de controle eficazes e iniciativas de saúde pública para mitigar o risco de infecção entre animais e transmissão zoonótica (18).

A prevalência de brucelose em gado pode variar significativamente entre diferentes regiões, com alguns estudos relatando taxas de soroprevalência tão altas quanto 17,2% em certos estados brasileiros (19,20). Essa variabilidade exige estratégias de vigilância e controle direcionadas para gerenciar e reduzir efetivamente a incidência da doença.

O diagnóstico da brucelose bovina é essencial para controlar a doença e minimizar seus impactos econômicos e de saúde. Vários métodos de diagnóstico são empregados, incluindo testes sorológicos, Teste do Anel em Leite (TAL) e o teste de Antígeno Acidificado Tamponado (AAT), enquanto os testes confirmatórios incluem o teste de Fixação do Complemento (FC), entre outros (4,14,21,22).

Cada um desses métodos tem suas vantagens e desvantagens. Por exemplo, embora o ELISA seja conhecido por sua alta sensibilidade e especificidade, ele pode não ser prático para uso em condições de campo devido à necessidade de equipamento especializado (23,24). Por outro lado, o RBT é mais simples e pode ser conduzido em campo, mas pode produzir falsos positivos, particularmente em animais vacinados (25,26).

No Brasil, o cenário diagnóstico para brucelose bovina é instruído pelo Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose (PNCEBT), que determina a vacinação de bovinos fêmeas e a implementação de protocolos de testes para identificar

rebanhos infectados (20). Estudos conduzidos em vários estados brasileiros, como Maranhão e Acre, destacaram a importância da vigilância sorológica em abatedouros para monitorar a prevalência de brucelose e garantir a segurança alimentar (19,27).

O uso de testes imunocromatográficos rápidos também tem sido explorado como uma ferramenta potencial para diagnóstico no local, fornecendo resultados em minutos e facilitando a tomada de decisão imediata (26,14). No entanto, a eficácia desses testes rápidos pode variar, e sua confiabilidade em comparação aos métodos sorológicos tradicionais continua sendo um tópico de pesquisa em andamento (14,24).

As vantagens dos testes diagnósticos usados no Brasil incluem sua capacidade de identificar rapidamente animais infectados, o que é crucial para controlar surtos e prevenir novas transmissões (20,25). No entanto, as desvantagens incluem o potencial de falsos negativos e positivos, principalmente em populações vacinadas, o que pode complicar o manejo do rebanho e os esforços de controle de doenças (25,28).

Por exemplo, a cepa vacinal S19, comumente usada no Brasil, pode interferir nos testes sorológicos, levando a desafios no diagnóstico preciso da brucelose em rebanhos vacinados (25,28). Isso exige o desenvolvimento de ensaios diagnósticos mais específicos que possam diferenciar entre animais vacinados e infectados, melhorando assim a precisão dos programas de vigilância da brucelose (21,24).

Recentemente a acurácia dos testes sorológicos utilizados no diagnóstico da brucelose bovina foi investigado. Os resultados indicaram que, maior sensibilidade foi de 96,5% para Ensaio Imunoenzimático Indireto (iELISA) e maior especificidade de 99,7% para o Ensaio de Polarização da Fluorescência (FPA). Porém, a análise levanta preocupações sobre o impacto de testes superestimados em programas de controle e erradicação da doença (1).

A tuberculose bovina é uma doença zoonótica crônica e causa impactos na saúde pública e economia (5). Além dos bovinos, a doença também pode afetar bubalinos, causando lesões granulomatosas (29). A doença tem uma ecologia de infecção complexa e é difícil de controlar em diversos países (30). O agente etiológico da tuberculose bovina é o *Mycobacterium bovis*. No entanto, outros membros do complexo *Mycobacterium tuberculosis*, como *M. africanum*, *M. caprae*, *M. orygis* e *M. microti*, também podem causar a doença em bovinos (31). Entretanto, *M. bovis* apresenta maior prevalência e possui uma gama mais ampla de hospedeiros em comparação com as demais bactérias que causam tuberculose bovina (7).

A bactéria pode ser transmitida aos humanos através do consumo de leite e derivados não pasteurizados (6) e contato próximo com animais infectados (7). Os sintomas da doença em humanos infectados por *M. bovis* são semelhantes aos sintomas causados por *M.*

tuberculosis, embora a primeira tenha maior probabilidade de causar doença extrapulmonar (7). Mesmo com programas de controle da tuberculose bovina, a doença continua sendo um desafio para os setores veterinários e de saúde pública, principalmente em países em desenvolvimento. Os programas de controle da doença consistem principalmente em testes e abates, limitação da circulação de animais e inspeções *post-mortem* (32).

Os testes de rotina para o diagnóstico de tuberculose bovina são o Teste de Cervical Simples, Teste de Prega Caudal e Teste de Cervical Comparativo (este último, usado também como teste confirmatório). Os testes são realizados com inoculação intradérmica de tuberculina nos animais. Quando for obtido resultado positivo ou inconclusivo, o animal pode ser submetido ao Teste Cervical Comparativo em um intervalo de sessenta a noventa dias, destinados ao abate sanitário ou eutanásia. Animais que testarem positivo para a doença deverão ser marcados, isolados, afastados da produção leiteira e abatidos no máximo trinta dias após o diagnóstico (29). Os atuais testes de diagnóstico para a doença são baseados em respostas sorológicas ou mediadas por células, e embora sejam eficazes na identificação da maioria dos animais infectados, suas limitações podem levar a falhas no diagnóstico (8).

3.2 Infravermelho por Transformada de Fourier

A espectroscopia analisa a interação entre radiação eletromagnética e matéria, sendo governada pela dualidade onda-partícula e composta por vetores de campo elétrico e magnético perpendiculares (33,34). A interação ocorre por absorção, quando um fóton eleva o estado energético molecular; emissão, com liberação de energia em forma de luz; e espalhamento, que desvia fótons e revela propriedades estruturais e eletrônicas (33,35).

Na espectroscopia vibracional, as técnicas mais usadas são a absorção no infravermelho (IR), que identifica grupos funcionais por modos vibracionais, e o espalhamento Raman, que avalia mudanças de polarizabilidade durante vibrações moleculares (33,34). Complementares, o IR é sensível a vibrações dipolo-ativas, enquanto o Raman destaca alterações na polarizabilidade, oferecendo uma visão abrangente das vibrações moleculares (33,34).

As vibrações moleculares podem ser classificadas como estiramento, deformação angular e torção. O estiramento pode ocorrer de diferentes formas: simétrica, antissimétrica, degenerada, em fase, fora de fase, ou ainda como pulsação ou respiração de anel, representando modos únicos de alteração nos comprimentos de ligação entre átomos e contribuindo para as características vibracionais da molécula (36).

Por sua vez, a deformação angular pode se manifestar de diversas maneiras, como simétrica, "*wagging*", "*twisting*", "*rocking*", degenerada, no plano, fora do plano, deformação de anel ou torção, resultando em mudanças nos ângulos de ligação que afetam a geometria e simetria molecular (37,38). Os átomos das moléculas formam uma estrutura tridimensional, com distâncias de ligação química (internucleares) e ângulos de ligação bem definidos, criando uma simetria molecular. Para uma molécula contendo N átomos, o número de vibrações normais será 3N - 6, caso a estrutura seja não linear, ou 3N - 5, se for linear, destacando a importância da geometria molecular na determinação das características vibracionais (39).

A vibração normal que provoca uma variação no momento de dipolo será ativa no IR, resultando na observação de uma banda vibracional no espectro IR. Esse fenômeno ocorre porque o momento de dipolo muda durante a transição vibracional, permitindo a absorção da radiação IR (40).

Caso a vibração cause um dipolo induzido, ela será ativa no espalhamento Raman, e uma linha ou banda será observada no espectro Raman. Esse comportamento decorre das mudanças na polarizabilidade molecular, que governam a atividade Raman, com implicações distintas para moléculas com centro de simetria, onde vibrações ativas no IR não serão ativas

no Raman e vice-versa (41). No espectro vibracional, além das bandas fundamentais, podem ocorrer outras bandas, como as de combinações por soma ou diferença e bandas harmônicas, enriquecendo as informações estruturais e dinâmicas das moléculas (42). A espectroscopia IR pode ser subdividida em três regiões: 10 a 400 cm^{-1} (IR distante ou FIR), 400 a 4000 cm^{-1} (IR médio ou MIR) e 4000 a 12820 cm^{-1} (IR próximo ou NIR), sendo o MIR a região mais utilizada para análise de compostos orgânicos devido à riqueza de suas características espectrais (42).

A maioria dos estudos em IR utiliza a região do MIR, onde estão localizadas as frequências vibracionais fundamentais, que correspondem às transições entre o nível de energia vibracional fundamental e o primeiro nível vibracional excitado. O intervalo espectral de 400 a 1800 cm^{-1} é conhecido como a região de assinatura espectral, pois é nessa faixa que aparecem a maior parte das frequências vibracionais fundamentais (43,44).

Os aparelhos de IR podem ser do tipo espectrofotômetro dispersivo ou espectrômetro por transformada de Fourier (FTIR). O espectrofotômetro dispersivo utiliza um monocromador com rede de difração ou prisma para decompor a radiação. Contudo, este tipo de aparelho está em desuso devido à sua lentidão, alto custo e necessidade de mecânica de alta precisão (45,46). Por outro lado, o espectrômetro FTIR, que utiliza o interferômetro de Michelson como princípio de funcionamento, é mais rápido, preciso, reprodutível e acessível (47,48). O espectrômetro FTIR é composto por uma fonte de radiação, um interferômetro, um compartimento de amostra e um detector de radiação infravermelha. Os componentes do espectrômetro podem variar dependendo da região do IR analisada. A fonte de radiação infravermelha mais utilizada na região MIR é o Globar (CSi), que é resfriado a água (49,50). O interferômetro inclui um divisor de feixe, um espelho fixo e um espelho móvel. O divisor de feixe utilizado no MIR é formado por um par de janelas de KBr, separadas por uma camada de germânio, que permite dividir o feixe infravermelho em duas partes iguais, refletindo metade e transmitindo a outra (49).

O feixe de radiação infravermelha gerado pela fonte incide em um espelho côncavo, que converte o feixe em raios paralelos cilíndricos. Esses raios são então direcionados a um divisor circular, que divide o feixe em duas partes iguais. Uma metade do feixe é refletida em direção a um espelho plano fixo, enquanto a outra metade segue para um espelho plano móvel. A radiação dirigida ao espelho fixo é refletida e atinge novamente o divisor circular, com metade do feixe sendo refletida de volta para a fonte. O mesmo ocorre com a radiação direcionada ao espelho móvel, que reflete totalmente o feixe em direção ao divisor circular, sendo que metade é refletida perpendicularmente à direção de incidência, e a outra metade é transmitida ao divisor (51,52).

As componentes de radiação utilizadas são a metade proveniente do espelho fixo, transmitida pelo divisor circular, e a metade do espelho móvel, refletida pelo divisor. Essas duas componentes se recombinam no divisor, gerando interferências construtivas ou destrutivas. O feixe resultante passa pelo compartimento de amostra e é direcionado a um espelho côncavo, que focaliza a radiação para o detector. A fonte de IR emite radiação contínua ao longo de uma ampla faixa espectral. Ao incidir no divisor, cada comprimento de onda sofre interferência. O detector captura o somatório das interferências de cada radiação durante o deslocamento do espelho móvel, gerando o perfil da observação, conhecido como interferograma (53,54).

O interferograma é um gráfico que representa a resposta do detector em função da diferença de caminho óptico. Quando o espelho móvel realiza o deslocamento completo, um interferograma completo é gerado, e esse deslocamento total corresponde a uma varredura espectral (*scan*). O espectrômetro utiliza a operação de transformada de Fourier para obter um espectro IR natural (*raw*), que representa o perfil espectral de intensidade versus número de ondas, caracterizando a espectroscopia FTIR (55,56).

No espectrômetro FTIR, é possível realizar múltiplas varreduras e acumular os interferogramas, um processo denominado co-adição. A co-adição permite um aumento significativo na qualidade do espectro (57,58). Para medir a relação sinal-ruído (SNR) de um espectro FTIR, primeiro é necessário identificar uma região sem picos significativos para estimar o ruído, calculando o desvio padrão da intensidade do sinal nessa área. A identificação dessa região de ruído é essencial, pois fornece uma linha de base contra a qual o sinal pode ser medido. Após estabelecer o nível de ruído, escolhe-se um pico de absorção representativo, e o sinal é calculado como a razão entre a intensidade máxima do pico e a linha de base (59,60).

Outro parâmetro importante é a resolução, que se refere à capacidade de diferenciação entre bandas próximas no perfil espectral. A resolução está diretamente relacionada ao número de pontos de aquisição durante a varredura. Alta resolução implica na necessidade de aquisição em intervalos menores, enquanto baixa resolução permite incremento/passo em intervalos maiores. Na alta resolução, o valor do número de ondas é baixo, enquanto na baixa resolução, o valor é alto. Quanto maior a resolução, maior será o deslocamento do espelho móvel, tornando a aquisição mais lenta, o que acaba resultando em um espectro com uma SNR menor pela maior captação de ruído adicional (61,62).

A possibilidade de acoplar acessórios ao espectrômetro FTIR permite a obtenção de espectros de amostras em diferentes estados físicos. Os modos de obtenção mais utilizados incluem transmissão, Refletância Total Atenuada (ATR), Refletância Difusa (DRIFTS), fotoacústico (PAS), refletância especular e microscopia no infravermelho (63). O modo

transmissão é o mais utilizado, sendo aplicável a amostras sólidas, líquidas e gasosas. Sua principal vantagem é proporcionar um espectro com alta razão sinal/ruído, enquanto a desvantagem é a necessidade de uma espessura adequada da amostra, a fim de evitar a saturação ou a presença de bandas com intensidade muito fraca. A espessura da amostra que a radiação IR atravessa normalmente varia de 1 a 20 μm (64,65).

O acessório ATR é montado no compartimento de amostras e utiliza um cristal que deve ser transparente ao IR, além de possuir um alto índice de refração e um ângulo de incidência específico da radiação. O feixe incidente atinge a superfície do cristal, onde sofre reflexão interna total, propagando-se ao longo do cristal até alcançar a extremidade oposta, onde está o detector. Durante esse processo, o feixe atinge a superfície do cristal em contato físico com a amostra, e a radiação penetra uma pequena profundidade na superfície da amostra, interagindo com ela. A radiação IR que penetra na amostra e sofre atenuação é chamada de onda evanescente (66,67).

3.3 Pré-tratamento de Dados

Com o objetivo de destacar a aparência e extrair o máximo de informações do espectro, recorre-se à manipulação espectral. Contudo, é preciso ter cautela, pois o excesso de manipulação pode introduzir artefatos que não existem nas observações ou até mesmo destruir informações espectrais mais sutis (68).

A dispersão dos dados é comum em técnicas analíticas que utilizam luz, pois as partículas na amostra possuem dimensões muito próximas aos comprimentos de onda empregados. Os métodos de pré-processamento podem corrigir esses efeitos, e o SNV (Standard Normal Variate) é um dos métodos de pré-tratamento mais utilizados (69).

O SNV corrige o deslocamento constante da linha de base (*offset*) e normaliza os dados subtraindo a média de cada variável do espectro e, em seguida, dividindo esse valor pelo desvio padrão do espectro, ou seja, conforme a Equação 1:

$$x_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \qquad (1)$$

Onde xij é o valor da intensidade do espectro por exemplo transmitância, μi a média das intensidades do espectro e σi seu desvio padrão (69).

Já o filtro Savitzky-Golay (SG), introduzido por Savitzky e Golay em 1964, é uma técnica amplamente reconhecida para suavizar dados ruidosos, particularmente no contexto de análise espectral. Este método emprega ajuste de mínimos quadrados polinomiais para suavizar dados, preservando características importantes como altura e largura do pico, tornando-o particularmente útil em campos que vão da química analítica ao sensoriamento remoto e aplicações biomédicas (70,71,72).

O filtro SG opera ajustando subconjuntos sucessivos de pontos de dados adjacentes (janela) com um polinômio de baixo grau, o que permite a redução do ruído sem distorcer significativamente o sinal original (73,74).

Uma das principais vantagens do filtro SG é sua capacidade de manter a integridade da forma dos dados enquanto reduz efetivamente o ruído. Essa característica é crucial em aplicações onde a identificação precisa de características espectrais é necessária, como na análise de espectros Raman, onde picos sobrepostos podem obscurecer informações importantes (75,76).

Estudos demonstraram que a suavização SG pode melhorar a relação sinal-ruído na espectroscopia Raman, facilitando a identificação de compostos em misturas complexas

(75,70). A eficácia do filtro neste domínio levou à sua adoção em várias técnicas analíticas, incluindo espectroscopia NIRS, onde é usado para pré-processar dados espectrais para maior precisão na análise quantitativa (77).

Avanços recentes na aplicação do filtro SG também exploraram seu uso em contextos de aprendizado de máquina, onde ele serve como uma etapa de pré-processamento para melhorar o desempenho de modelos preditivos (43,78). Essa tendência destaca o crescente reconhecimento do filtro SG não apenas como uma ferramenta de suavização de dados, mas também como um componente crítico no pipeline de pré-processamento de dados no aprendizado de máquina.

A eficácia do filtro SG é ainda mais ressaltada por seu desempenho comparativo com outras técnicas de suavização. Pesquisas indicaram que a suavização SG frequentemente supera métodos tradicionais como médias móveis e filtros gaussianos, particularmente na preservação das características do sinal original enquanto reduz o ruído (43,76).

3.4 Análise de Componentes Principais

O PCA (Análise de Componentes Principais), método introduzido por Karl Pearson (1903) e formalmente descrito por Hotelling (1933), transforma as variáveis originais em novas variáveis chamadas Componentes Principais (PC's), que são não-correlacionadas entre si, mas preservam a informação estatística dos dados. Como as PC's são ordenadas de acordo com a variância, é possível reduzir a dimensionalidade dos dados, selecionando as componentes que capturam a maior parte da variação, sem perda significativa de informação (79).

O PCA é um método utilizado na análise exploratória de dados, permitindo a compreensão das relações entre as diferentes classes. Com caráter adaptativo, pode ser aplicado em dados numéricos de diversas áreas do conhecimento. O método utiliza uma matriz de dados X (n x p), onde n representa o número de amostras e p o número de variáveis numéricas para cada amostra. O objetivo do PCA é encontrar uma combinação linear das colunas da matriz X que maximize a variância. Por meio das combinações lineares de X, obtemos as PC's, cujos elementos são os autovalores, chamados de *scores*, e os autovetores, denominados *loadings* (79).

O PCA reduz a dimensionalidade dos dados ao capturar as direções de maior variância, permitindo identificar padrões (80). A variância explicada indica a proporção da variação total capturada por cada componente, sendo o primeiro componente geralmente responsável pela maior parcela dessa variância, seguido pelos demais componentes em ordem decrescente (81). Os *scores* representam as coordenadas dos dados no espaço das componentes principais, facilitando a visualização de agrupamentos (82). Já os *loadings* mostram as contribuições de cada variável para os componentes principais, revelando relações subjacentes entre variáveis (83).

O T^2 de Hotelling é uma generalização do teste t de Student, útil na identificação de *outliers* nos scores do PCA, avaliando se algum ponto no espaço reduzido está significativamente distante do centro, considerando a variância conjunta capturada pelas componentes principais (84). O T^2 utiliza a matriz de covariância das componentes principais e calcula a distância de cada ponto à média do espaço reduzido, identificando amostras atípicas (85). Valores elevados de T^2 indicam a presença de *outliers* e são comparados ao limite superior de controle calculado com base na distribuição F (86,87).

3.5 Aprendizado de Máquina

De forma geral, o principal objetivo da classificação multivariada é agrupar amostras com propriedades desconhecidas em classes previamente definidas, ou seja, classificá-las com base em modelos (matemáticos ou estatísticos) desenvolvidos a partir de amostras cujas propriedades são conhecidas ou avaliadas (88)

Os métodos de classificação multivariada podem ser divididos em paramétricos e não paramétricos. Os métodos paramétricos fazem suposições sobre a distribuição das variáveis, como é o caso do método LDA (Análise Discriminante Linear), que assume que as variáveis seguem uma distribuição normal com variâncias iguais para cada classe. Já os métodos não paramétricos, como o SVM (Máquinas de Vetores de Suporte), não fazem suposições sobre a distribuição dos dados e geram classificadores a partir da matriz de dados, utilizando a construção de um hiperplano que separa as classes com a maior margem possível, sem a necessidade de uma distribuição predefinida (89).

Um modelo de aprendizado de máquina é uma função que mapeia um conjunto de entradas (recursos) para saídas (rótulos). Os recursos ou *features* são as variáveis independentes que alimentam o modelo, enquanto os rótulos ou *labels* são as variáveis dependentes que o modelo tenta prever. (90,91).

Para avaliar o desempenho dos métodos de classificação, as amostras são divididas em dois conjuntos: um para treinamento e outro para teste (92). O processo de treinamento de um modelo envolve a otimização de seus parâmetros para minimizar a diferença entre as previsões do modelo e os rótulos reais (93). Após o treinamento, o modelo é validado utilizando um conjunto de dados separado, conhecido como conjunto de teste, para ajustar hiperparâmetros (94).

Os parâmetros de um modelo são valores aprendidos a partir dos dados durante o treinamento e são ajustados automaticamente pelo algoritmo de otimização (93). Já os hiperparâmetros, como os parâmetros de regularização, são definidos antes do treinamento, não sendo ajustados automaticamente, mas configurados manualmente ou por técnicas de busca, como a busca em grade (*Grid Search*) (95). Se o modelo se ajustar excessivamente ao conjunto de amostras de treinamento, poderá ter dificuldades para classificar novas amostras, ou seja, perderá sua capacidade de generalização, fenômeno conhecido como sobreajuste ou *overfitting* (88).

Overfitting e *underfitting* são dois problemas críticos que afetam a performance dos modelos de aprendizado de máquina. O *overfitting* ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, capturando ruídos e variações aleatórias, resultando em um desempenho ruim em dados de teste. Em contraste, o *underfitting* acontece quando um modelo é muito simples para capturar a complexidade dos dados, levando a um desempenho insatisfatório tanto em treinamento quanto em teste (96). A relação entre *overfitting* e *underfitting* é frequentemente descrita pelo *trade-off* entre viés e variância, onde modelos com alta variância tendem a *overfitting* e modelos com alto viés tendem a *underfitting* (97,98).

A validação cruzada é uma das estratégias mais utilizadas para realizar uma quantidade suficiente de testes com um número limitado de amostras, evitando o *overfitting* (99). Na validação cruzada, os dados são divididos de maneira aleatória em "k" subconjuntos (ou pastas) de tamanhos aproximadamente iguais, sendo esse método conhecido como *k-fold*. Uma alternativa eficiente ao *k-fold* é a validação cruzada LOOCV (*Leave-One-Out Cross Validation*), onde k (número de dobras) é igual ao número de amostras, e apenas uma dessas dobras é usada para validação, enquanto as demais são utilizadas para o treinamento. O LOOCV é amplamente utilizado em conjuntos de amostras pequenos (92).

Finalmente, o desempenho do modelo é avaliado em um conjunto de teste, que é utilizado para medir a generalização do modelo em dados não vistos (90,91). Dentre as abordagens mais comuns para a análise dos resultados da validação cruzada, destaca-se a matriz de confusão, na qual a classe estimada ou predita é comparada com a classe verdadeira ou real. Neste contexto, os termos verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN) são empregados para qualificar as predições. A acurácia, sensibilidade e especificidade são métricas comuns utilizadas para avaliar o desempenho de modelos de classificação. A acurácia é a proporção de previsões corretas em relação ao total de previsões, enquanto a sensibilidade mede a capacidade do modelo de identificar corretamente as classes positivas. A especificidade, por outro lado, avalia a capacidade do modelo de identificar corretamente as classes negativas (90,91).

3.6 Máquinas de Vetores de Suporte

O método SVM (Máquinas de Vetores de Suporte) constrói hiperplanos para definir uma fronteira ou limite de decisão que separa as observações de acordo com suas classes. Inicialmente concebido para problemas de classificação binária (duas classes) e separação linear, o SVM foi posteriormente adaptado para problemas multiclasse e não-lineares. A técnica baseia-se na construção de um novo espaço de dados (espaço de características) onde a classificação se torna mais simples, ou seja, a separação linear se torna viável (88).

Em dados reais, frequentemente ocorre sobreposição de classes devido a ruídos e *outliers*. Nesse contexto, é necessário estabelecer uma tolerância de erro em torno da margem de separação. Para isso, um termo de regularização é introduzido, uma constante que busca equilibrar a largura da margem de separação com os erros de classificação. O método, então, procura um equilíbrio entre a complexidade do modelo e o erro empírico, o que é conhecido como minimização do risco estrutural. Esse princípio visa otimizar a capacidade de generalização do modelo, controlando o trade-off entre a margem e os erros de classificação (88).

Para isso, o SVM utiliza funções chamadas kernel, que permitem o mapeamento dos dados para um espaço de características de maior dimensão. Esse processo é conhecido como o truque do *kernel*, pois ele evita a necessidade de construir explicitamente o mapeamento, realizando os cálculos diretamente no espaço original. Assim, apesar de trabalhar em um espaço de alta dimensão, o SVM não realiza cálculos explícitos nesse espaço, o que torna o processo computacionalmente eficiente. Dentre os principais tipos de funções *kernel*, temos: linear, polinomial, gaussiana ou função de base radial (RBF), e sigmoidal (88).

Uma propriedade importante do SVM é a esparsidade, que permite ao método convergir para uma única solução global, utilizando apenas os pontos de dados localizados próximos à fronteira de decisão. Esses pontos são denominados vetores de suporte, e são essenciais para a construção do hiperplano de separação. A esparsidade contribui para a eficiência do modelo, pois apenas os vetores de suporte influenciam diretamente a solução final (88).

O funcionamento do SVM é baseado na ideia de encontrar um hiperplano que separa as classes de dados com a maior margem possível. A margem é definida como a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte. O SVM busca maximizar essa margem, resultando em uma melhor generalização do modelo para dados não vistos (100).

A fronteira de decisão em uma SVM é definida pelo hiperplano que separa as classes. A posição e a orientação desse hiperplano são determinadas pelos vetores de suporte, que são os pontos de dados mais próximos da fronteira. Quanto maior a distância de um ponto à fronteira, maior a confiança na classificação desse ponto (101). A fronteira de decisão pode ser visualizada em um espaço bidimensional, onde as classes são separadas por uma linha (hiperplano). Em espaços de maior dimensão, a visualização se torna mais complexa, mas o conceito permanece o mesmo.

A certeza de uma predição em porcentagem pode ser avaliada através da margem de separação; quanto maior a distância de um ponto à fronteira de decisão, maior a confiança na classificação desse ponto. Em muitos casos, a saída da SVM pode ser convertida em uma probabilidade utilizando métodos como a calibração de Platt, que ajusta a saída da SVM para que ela represente uma probabilidade (102). A calibração de Platt envolve o ajuste de uma função logística aos resultados da SVM, permitindo que as saídas sejam interpretadas como probabilidades. Isso é especialmente útil em aplicações onde a interpretação da certeza é crítica, como em diagnósticos médicos (103).

Os hiperparâmetros desempenham um papel crucial no desempenho do modelo treinado, para o SVM os principais hiperparâmetros incluem a escolha da função kernel e a definição adequada do hiperparâmetro de regularização (C). O hiperparâmetro C controla o trade-off entre maximizar a margem e minimizar o erro de classificação. Um valor alto de C resulta em uma margem menor, permitindo que mais pontos de dados sejam classificados corretamente, mas pode levar ao overfitting. Por outro lado, um valor baixo de C pode resultar em uma margem maior, mas com maior risco de erros de classificação (104).

O SVM Gaussiano (RBF) é um dos kernels mais populares, que transforma os dados em um espaço de alta dimensão, permitindo a separação não linear. O RBF é especialmente eficaz em cenários com alta dimensionalidade e complexidade (105). Para este classificador o hiperparâmetro *gamma* controla a largura da função de base, afetando a forma da decisão (106). A escolha e a otimização desses hiperparâmetros são fundamentais para o sucesso do modelo. Métodos como validação cruzada e busca em grade (*Grid Search*) são frequentemente utilizados para encontrar a combinação ideal de hiperparâmetros (107). Além disso, técnicas como a otimização bayesiana têm sido aplicadas para melhorar a eficiência na busca por hiperparâmetros, permitindo uma exploração mais eficaz do espaço de parâmetros (108).

Para classificação de mais de duas classes (multiclasse) o SVM possui duas abordagens principais: *One-vs-One* e *One-vs-All*. No *One-vs-One*, são treinados k(k-1)/2

classificadores binários, onde cada um distingue entre um par de classes, e a classificação final é determinada por votação majoritária (109). Já no *One-vs-All*, cada classe é comparada contra todas as outras, simplificando o problema multiclasse em várias tarefas binárias, o que pode ser mais eficiente em grandes conjuntos de dados
(110).

Finalmente, a normalização dos dados é uma etapa essencial no pré-processamento antes de aplicar o SVM, especialmente quando os dados apresentam grande variância (como no caso dos *scores* de PCA). A normalização SNV elimina a influência de variações absolutas, permitindo que cada característica contribua igualmente para a distância calculada entre os pontos (111). A normalização não apenas melhora a convergência dos algoritmos de otimização, mas também ajuda a evitar que características com magnitudes diferentes dominem a função de decisão.

3.7 Diagnóstico de Brucelose e Tuberculose por FTIR e Aprendizado de Máquina

Os espectros de FTIR de soro sanguíneo refletem os movimentos vibracionais de grupos funcionais específicos, como proteínas, carboidratos e lipídios. Esses movimentos geram espectros infravermelhos característicos, descritos como “impressões digitais” moleculares (112). O soro sanguíneo reflete o estado fisiológico de um indivíduo e tem sido amplamente utilizado na identificação de biomarcadores de doenças (113).

O aumento na concentração de neopterina, por exemplo, está frequentemente relacionado a infecções bacterianas crônicas e reflete a ativação das respostas imunológicas mediadas por citocinas (114). As infecções podem afetar também os perfis lipídicos do soro. Estudos indicam que infecções virais, como a dengue, podem induzir dislipidemia, caracterizada por alterações nos níveis de lipoproteínas de baixa densidade (LDL) e triglicerídeos (115). As alterações bioquímicas no soro durante infecções não se restringem a proteínas e lipídios, abrangendo também mudanças em eletrólitos e subprodutos metabólicos. Por exemplo, infecções podem modificar os níveis de glicose e creatinina, indicando possível disfunção orgânica (116).

Adicionalmente, a resposta imune desencadeada pelas infecções pode causar modificações nos parâmetros hematológicos, como contagem de leucócitos e níveis de hemoglobina. Infecções como a malária, podem resultar em anemia e trombocitopenia, refletidas no perfil bioquímico do soro (117,118). No caso de infecções específicas, como a tuberculose, o perfil bioquímico do soro pode fornecer informações sobre a gravidade e a progressão da doença. Níveis elevados de biomarcadores, como a procalcitonina, têm sido associados a infecções bacterianas e podem atuar como indicadores da severidade da infecção (119,120).

Além disso, as infecções podem induzir estresse oxidativo, o que pode alterar ainda mais a composição do soro. O aumento de espécies reativas de oxigênio (ROS) durante as infecções pode modificar a conformação e a função das proteínas, levando a mudanças nos perfis proteicos do soro (121). A utilização da espectroscopia FTIR na análise das mudanças do soro durante infecções surge como uma abordagem promissora para o desenvolvimento de ferramentas diagnósticas rápidas.

O procedimento consolidado na literatura para análise de soro sanguíneo por FTIR consiste em depositar a amostra em um substrato de silício e secá-la em estufa. Esse procedimento elimina a interferência da água nos espectros, produzindo espectros mais nítidos, porém, a secagem pode introduzir viés, devido às variações de temperatura na estufa ou à

distribuição não homogênea dos componentes na gota seca durante a medição, exigindo maior controle do ambiente e a coletas em triplicata (11).

O soro sanguíneo centrifugado e sem anticoagulante é uma alternativa confiável, rápida e simples de operar devido à estabilidade e homogeneidade de seus componentes. Porém, a baixa concentração de analitos em relação a água e a sobreposição das bandas da água sobre as de outros componentes pode inviabilizar os resultados. Uma possível solução é o uso da técnica chamada "*Digital Drying"*, onde ocorre a subtração direta dos espectros de soro líquido pelo espectro da água. Um processo equivalente é a medida do soro líquido precedida de *background* no equipamento com água deionizada (12).

O uso de soro sanguíneo líquido no diagnóstico por FTIR e aprendizado de máquina tem mostrado resultados promissores, como no estudo da brucelose, que diferenciou ovelhas infectadas por *Brucella* de saudáveis (9). Da mesma forma, a espectroscopia Raman combinada com aprendizado de máquina tem se mostrado eficaz no diagnóstico de tuberculose humana (122). Essas técnicas têm a vantagem de não exigirem o uso de reagentes ou processos invasivos, oferecendo um método rápido para a detecção oportuna de doenças infecciosas.

4. METODOLOGIA

Quarenta (40) amostras positivas para tuberculose bovina e 40 amostras controle foram cedidas pelo Laboratório Federal de Defesa Agropecuária (LFDA) de Minas Gerais. Outras 40 amostras controle e 40 positivas para tuberculose bovina foram cedidas pela EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) Gado de Corte.

O experimento foi conduzido em três etapas para avaliar a classificação de amostras biológicas entre Controle, Brucelose e Tuberculose. Inicialmente, foi realizada a classificação binária entre Controle e Brucelose comparando duas abordagens: soro seco em estufa ("Soro Seco") e soro líquido com *background* em água deionizada ("Soro Líquido"). Em seguida, focando apenas no Soro Líquido, realizou-se a classificação binária entre Controle e Tuberculose, bem como entre Controle e Brucelose. Por fim, foi realizada a classificação multiclasse, utilizando amostras de Soro Líquido, para diferenciar simultaneamente as três condições (Controle, Brucelose e Tuberculose).

4.1 Preparo e Aquisição de Dados

Para as amostras de Soro Seco, alíquotas de 10 µl foram retiradas da parte central dos tubos das amostras de Brucelose e depositadas em um substrato de silício, com secagem em estufa a 40 °C, três vezes a cada 10 minutos (no mesmo ponto). Os espectros foram medidos em triplicata no centro e nas extremidades da gota seca, com um *background* da atmosfera realizado antes de cada aquisição (123). Já para as amostras de Soro Líquido, alíquotas de 20 µl foram coletadas da mesma região central dos tubos, e os espectros foram obtidos diretamente após a realização de um *background* utilizando água deionizada (12). Os espectros foram obtidos por espectroscopia de Infravermelho por Transformada de Fourier (FTIR), utilizando espectrômetro da marca Agilent, modelo Cary 630, no intervalo de 1800 até 900 cm^{-1} (passo de 0,5 cm^{-1}), com resolução de 4 cm^{-1} e 12 varreduras (123), utilizando um acessório de Reflectância Total Atenuada (ATR) com cristal de seleneto de zinco (ZnSe).

4.2 Análise Exploratória de Dados

Os dados foram tratados por Variação Normal Padrão (SNV) e os ruídos do Soro Líquido foram suavizados pelo filtro Savitzky-Golay (janela: 3, grau: 1). Os dados tratados foram reduzidos pela Análise de Componentes Principais (PCA). Os *scores* obtidos foram utilizados para verificar a existência de agrupamentos e a remoção de outliers utilizando o critério de T^2 de Hotelling (124). Em seguida, os dados tratados foram divididos em 70% para treino e 30% para teste, utilizando um sorteio de dados estratificados.

4.3 Aprendizado de Máquina

O classificador utilizado foi o Máquina de Vetores de Suporte (SVM) e sua acurácia de validação foi calculada utilizando o método *Leave-One-Out Cross-Validation* (LOOCV). Nesse processo, a cada iteração, uma amostra do conjunto de treino é deixada de fora, enquanto as demais passam por redução de dimensionalidade com PCA e normalização dos *scores* por SNV. O modelo é treinado com as amostras normalizadas, e a amostra deixada de fora é projetada no espaço das componentes principais (PC's) calculado durante o treino. Seus *scores* também são normalizados antes de serem utilizados na validação do modelo. Esse procedimento é repetido para todas as amostras, garantindo que cada uma fosse validada uma vez. A acurácia final de validação é obtida pela média das acurácias em todas as iterações.

Os hiperparâmetros do SVM foram ajustados testando diferentes combinações de *kernels* (linear, quadrático e cúbico) e valores de C variando de 0,01 a 0,20 (incrementos de 0,01). O modelo final foi treinado com 70% das amostras, reservando 30% para validação externa (teste). A redução de dimensionalidade e a normalização seguiram o mesmo procedimento usado no LOOCV. Os hiperparâmetros ótimos foram selecionados considerando a maior acurácia e o menor número de PC's necessários. Para avaliar o desempenho do modelo, foram construídas matrizes de confusão, e os valores de sensibilidade foram comparados entre amostras de Soro Seco e Soro Líquido. A relação sinal-ruído (SNR) foi calculada como a razão entre a amplitude máxima no intervalo de 1300-900 cm^{-1} e o desvio padrão das intensidades no intervalo de 1800 – 1750 cm^{-1}, região considerada livre de sinal e representativa do ruído de fundo no espectro. Todas as análises foram realizadas com a biblioteca Scikit-Learn (versão 1.3.0) e Python (versão 3.11.5).

5. RESULTADOS E DISCUSSÃO

5.1 Soro Seco e Soro Líquido no Diagnóstico de Brucelose Bovina

Na Fig. 1, os espectros das amostras de água deionizada (AD), soro sanguíneo líquido, soro seco em estufa e soro líquido com *background* em água deionizada são comparados quanto à intensidade de absorbância.

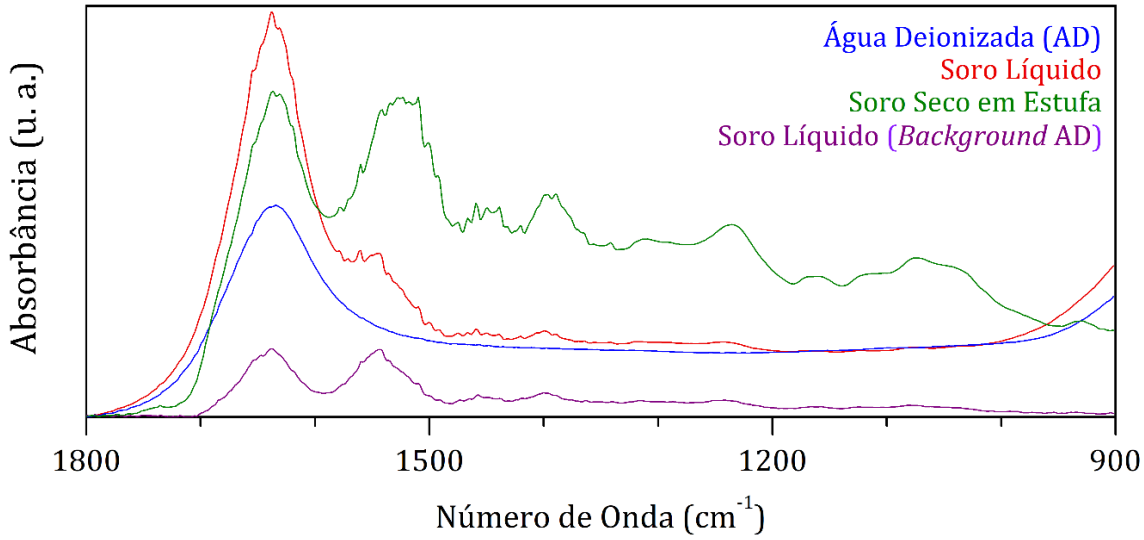


Fig. 1: Espectros FTIR de água deionizada (AD) e soro sanguíneo bovino em diferentes condições: soro sem preparação (Soro Líquido), soro seco em estufa para remoção da interferência da água (Soro Seco em Estufa) e soro medido após *background* de água deionizada (Soro Líquido – *Background* AD).

Os espectros de soro seco apresentaram maior intensidade de absorbância em relação aos demais, exceto na banda em torno de 1637 cm^{-1}, associada às vibrações amida I, onde os espectros de soro líquido foram mais intensos, indicando maior contribuição de proteínas solúveis (125). Além disso, os espectros de soro seco mostraram bandas com maior amplitude e definição, especialmente acima de 1350 cm^{-1}, devido à menor interferência da água e maior concentração de compostos sólidos. Esses fatores justificam seu uso como principal tratamento para análise de soro sanguíneo em Infravermelho por Transformada de Fourier (FTIR).

Os espectros de soro líquido com *background* em AD apresentaram menor intensidade de absorbância em comparação com os demais. Esse comportamento pode ser atribuído à baixa concentração de analito na solução aquosa, além de possíveis efeitos da manipulação espectral (*background*). Sinais de baixa intensidade tendem a ser mais vulneráveis à interferência de ruídos devido à reduzida relação sinal-ruído (SNR), exigindo, portanto, a

utilização de menor resolução espectral, um maior número de varreduras e a aplicação de filtros de suavização.

A média da SNR dos espectros de soro seco foi de 101,1 ± 26,2, enquanto os espectros de soro líquido sem a aplicação do filtro de suavização Savitzky-Golay (SG) apresentaram uma SNR média de 37,5 ± 9,6. Após a aplicação do filtro SG (janela: 3, grau: 1), a SNR dos espectros de soro líquido aumentou para 56,2 ± 15,1, representando uma melhoria de aproximadamente 50%.

Esses resultados indicam que os espectros de soro seco têm uma SNR significativamente superior ao do soro líquido, refletindo uma maior qualidade dos dados, possivelmente devido à menor interferência ou maior estabilidade dos espectros de soro seco. A aplicação do filtro SG no soro líquido foi eficaz na redução do ruído, mas ainda assim, os espectros de soro seco continuam a apresentar uma SNR muito superior, o que sugere que a preparação do soro seco oferece vantagens na obtenção de espectros com maior número de detalhes e informação espectral.

Os espectros de AD apresentaram uma única banda evidente neste intervalo espectral, em torno de 1635 cm^{-1}, associada ao modo vibracional de flexão do O–H, que coincide com a região da amida I do soro líquido. Após 1200 cm^{-1}, as intensidades de absorbância entre AD e soro líquido tornam-se semelhantes, sugerindo a interferência da água presente na amostra devido à sobreposição espectral.

Os espectros de soro líquido apresentaram maior intensidade de absorbância em relação aos espectros de soro líquido com *background* em AD. Além disso, exibiram bandas com maior amplitude, indicando maior relação sinal-ruído. Esses resultados sugerem que o uso de soro líquido sem *background* em AD pode ser uma abordagem mais eficiente e promissora no diagnóstico de doenças. A partir daqui a amostra de soro líquido com *background* em AD, objeto do estudo, será chamada simplesmente de "soro líquido".

Na Fig. 2, a média e o desvio padrão dos espectros de soro seco e soro líquido são comparados quanto às suas principais bandas. As cores indicam a quais componentes essas bandas se relacionam, e seus modos vibracionais estão indicados entre parênteses.

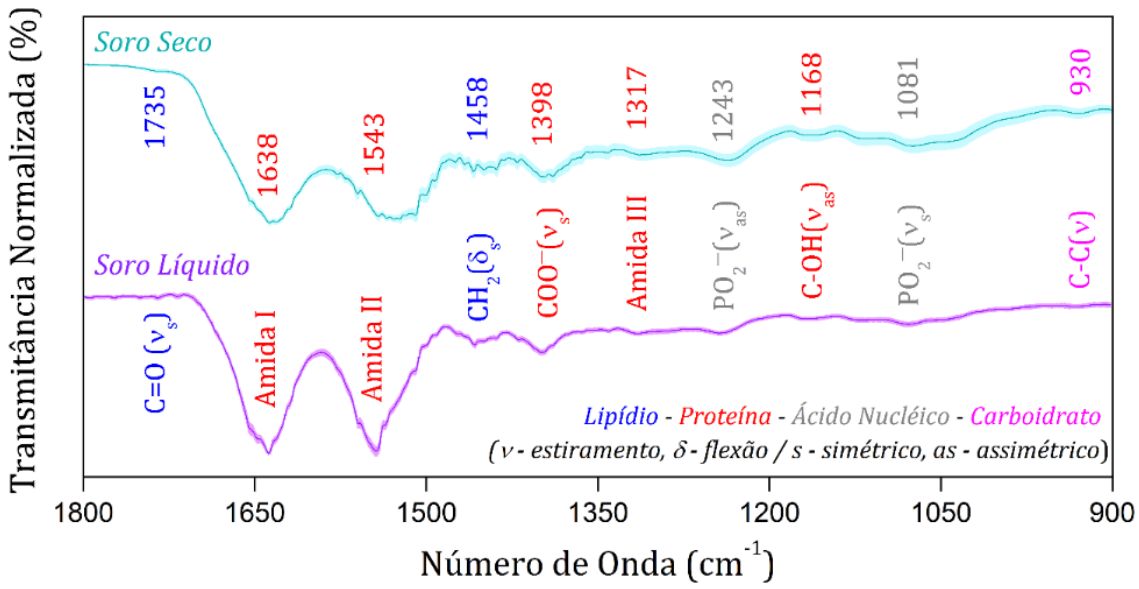


Fig. 2: Média e desvio padrão dos espectros FTIR de soro seco sem estufa (Soro Seco) e soro líquido com background em água deionizada (Soro Líquido) e suas principais bandas. As cores indicam a quais componentes essas bandas se relacionam e os seus modos vibracionais são indicados por letras gregas entre parênteses.

As principais bandas relacionadas ao soro sanguíneo foram identificadas tanto nas amostras de soro seco como nas de soro líquido (126). Nos espectros de soro seco, as bandas associadas à amida I e II são mais largas do que nas do soro líquido. A amplitude das amidas I e II, em comparação com as demais bandas, é menor no soro seco do que no soro líquido, onde essas bandas possuem maior magnitude de absorbância.

Além disso, os espectros de soro seco apresentaram maior desvio padrão do que os de soro líquido, possivelmente devido à dispersão e absorção irregulares características das amostras secas, o que pode ter causado variações nas intensidades espectrais. Em contraste, os espectros de soro líquido são mais homogêneos, possivelmente devido à diluição da matriz em água. O processo de secagem pode ter causado desnaturação nas proteínas, resultando em rearranjo das moléculas ou mudanças nas interações intermoleculares, que podem ter causado um deslocamento na posição da amida II, em torno de 1500 cm^{-1}, além de outros deslocamentos entre 1800 e 1350 cm^{-1}, observados na Fig. 2 como pequenos “ombros” próximos às bandas principais.

Assim, os espectros de soro líquido apresentaram menor intensidade de absorbância em relação aos de soro seco, com maior susceptibilidade à interferência de ruídos. No entanto, eles possuem bandas mais bem definidas, sem a interferência do processo de secagem, o que pode preservar informações, especialmente sobre as proteínas.

Na Fig. 3, as médias dos espectros das classes Controle e Brucelose para as amostras de soro seco (a) e soro líquido (b) são comparadas. O objetivo é identificar padrões ou características espectrais específicas nos dados médios que possam distinguir as classes.

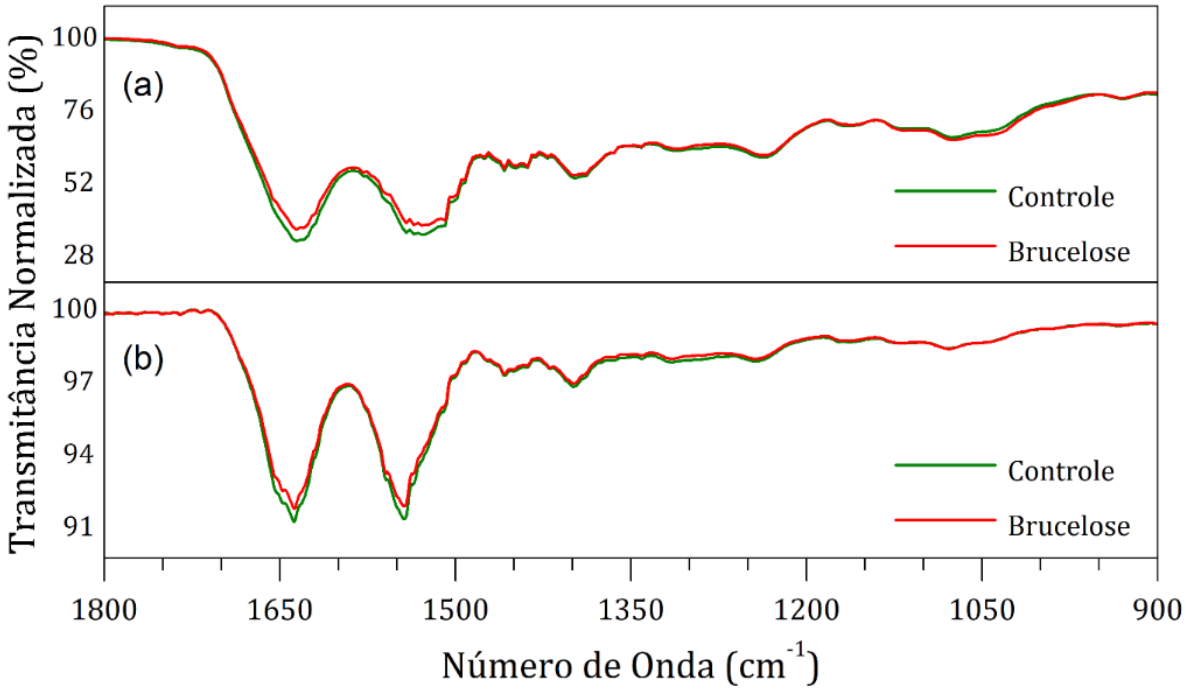


Fig. 3: Média dos espectros FTIR normalizados por SNV das amostras Controle e Brucelose comparados quanto ao tipo de preparação do soro sanguíneo: soro seco em estufa (a) e soro líquido com *background* em água deionizada (b).

Os espectros das amostras de Brucelose apresentaram menor transmitância, esse comportamento pode estar relacionado a alterações nas propriedades ópticas causadas pela resposta imune à infecção, observadas tanto no soro seco quanto no soro líquido. Essas alterações podem ser atribuídas a mudanças na composição das proteínas plasmáticas, como imunoglobulinas, albuminas ou fibrinogênio, e no conteúdo de lipídios. O processo inflamatório pode modificar a permeabilidade celular, a estrutura das células sanguíneas e aumentar a concentração de citocinas e outras substâncias inflamatórias. Além disso, durante uma infecção, alterações na viscosidade do sangue e o aumento de células sanguíneas ou de complexos imunes podem influenciar a interação das moléculas com a radiação infravermelha, resultando em mudanças na intensidade e posição das bandas (127)

Não foi possível identificar visualmente biomarcadores espectrais nas bandas que permitissem a classificação específica dos espectros de Controle e Brucelose. Esse resultado evidencia a necessidade do uso da análise multivariada e do aprendizado de máquina para

identificar padrões sutis, que possibilitem o desenvolvimento de diagnósticos precisos e confiáveis.

Na Fig. 4 os *scores* e os *loadings* das duas primeiras componentes principais dos espectros de soro seco (a, b) e soro líquido (c, d) são comparados com o objetivo de avaliar o agrupamento das classes Controle e Brucelose, bem como identificar as bandas de contribuição para a formação desses grupos. Os círculos e quadrados são as amostras de treino utilizadas na Análise de Componentes Principais (PCA), enquanto os contornos dessas figuras são as amostras de teste projetadas neste mesmo espaço. Para facilitar a interpretação dos *loadings* (b, d), a média dos espectros de soro seco e soro líquido é exibida na cor cinza.

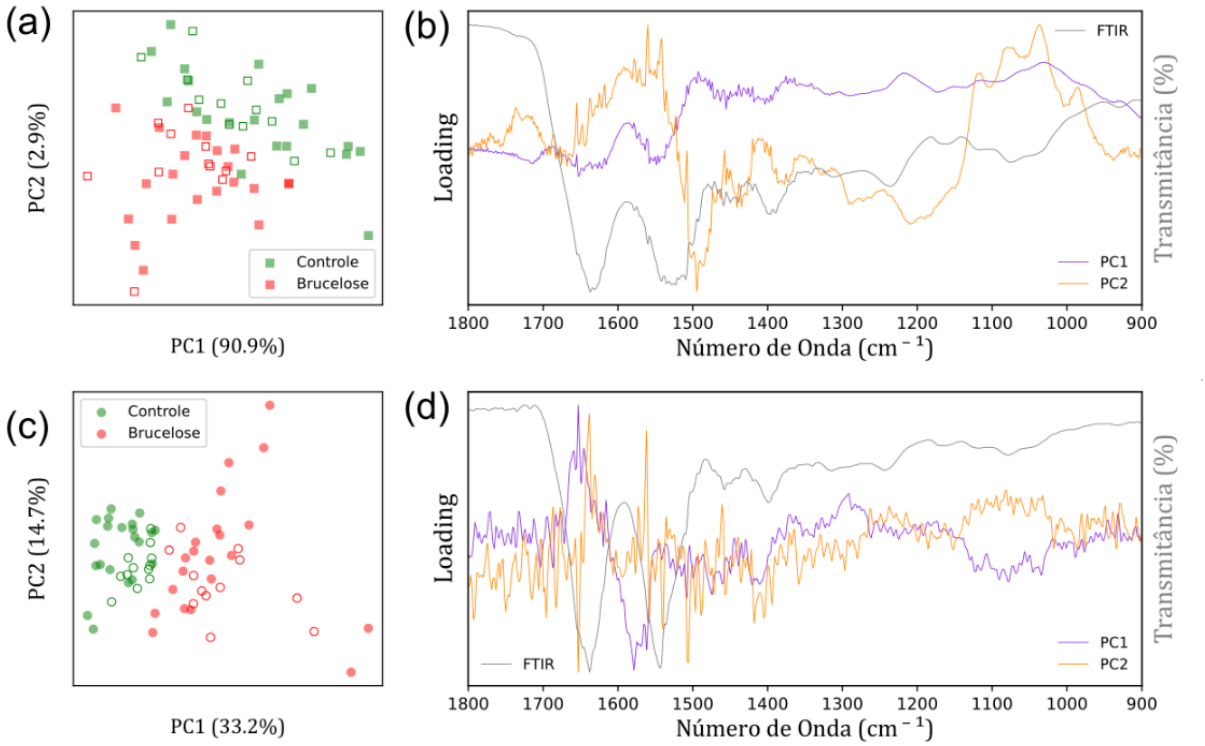


Fig. 4: *Scores* (gráfico de dispersão) e *loadings* (gráfico de linha) das amostras de soro seco em estufa (a, b) e soro líquido com *background* em água deionizada (c, d). a média de todos os espectros é exibida na cor cinza para comparação entre os pesos e as bandas do espectro das amostras.

Comparando os *scores* (a, c), observa-se a formação de grupos de forma dispersa, mas com fronteiras lineares bem definidas. As amostras de treino e teste ficaram agrupadas corretamente em suas respectivas classes. No caso do soro seco, a fronteira entre as classes estava muito próxima, o que fez com que algumas amostras de uma classe estivessem posicionadas no grupo formado pela classe oposta. A separação dos grupos ao longo da diagonal reflete a influência das duas primeiras componentes principais (PC's). No soro líquido, não houve sobreposição entre os grupos, mas um espaço vazio. Isso indica que o modelo terá facilidade em

realizar a classificação, com a separação ocorrendo quase exclusivamente ao longo da primeira componente principal (PC1).

A variância explicada acumulada pelas duas primeiras PCs foi de 93,8% para o soro seco e 47,9% para o soro líquido. Embora o soro seco tenha apresentado maior variância explicada, suas amostras não formaram grupos tão bem definidos quanto as amostras de soro líquido, possivelmente devido à perda de informações relevantes durante o processo de secagem. A baixa variância observada no soro líquido pode ser atribuída à presença de ruído nos espectros, que pode obscurecer informações mais relevantes. No entanto, a formação de grupos foi satisfatória, indicando que a informação foi preservada apesar da baixa SNR.

Os *loadings* da PC1 de soro seco (b) indicam uma contribuição equilibrada das principais bandas (Fig. 2), enquanto na PC2 as contribuições a partir de 1300 cm^{-1} se destacam pela sua magnitude. Na região das amidas, observam-se pesos com características ruidosas e de grande magnitude, que não são claramente associadas às bandas de amida I ou II.

Os *loadings* da PC1 do soro líquido (d) apresentam um aspecto mais ruidoso em todo o intervalo espectral, mas ainda é possível associar esses pesos a contribuições semelhantes às observadas no soro seco, com destaque para a maior magnitude na região das amidas, em comparação com o soro seco. Na PC2, não foi possível identificar contribuições associadas a bandas de referência.

As contribuições observadas no soro seco também estão presentes no soro líquido, com as mesmas bandas identificadas em ambas as condições. No entanto, essas bandas apresentam uma distribuição distinta nos loadings, aparecendo na PC1 em uma amostra e na PC2 na outra. Esse comportamento sugere que a preparação da amostra influencia significativamente os espectros obtidos, embora os pesos indiquem que as mesmas variáveis contribuem para os resultados em ambos os casos. Não foi possível identificar visualmente biomarcadores espectrais nos *loadings* que distinguissem de forma específica os espectros de Controle e Brucelose, devido à contribuição conjunta de diferentes componentes (lipídios, proteínas e carboidratos).

A Fig. 5 compara as acurácias obtidas por *Leave-One-Out Cross-Validation* (LOOCV) para diferentes valores do hiperparâmetro C (0,01 até 0,2 a cada 0,01) e tipos de *kernel* (linear, quadrático e cúbico), utilizando as duas primeiras componentes principais das amostras de soro seco (a) e soro líquido (b). O objetivo é identificar o valor de acurácia mais alto, o que indica o melhor desempenho para o respectivo conjunto de dados.

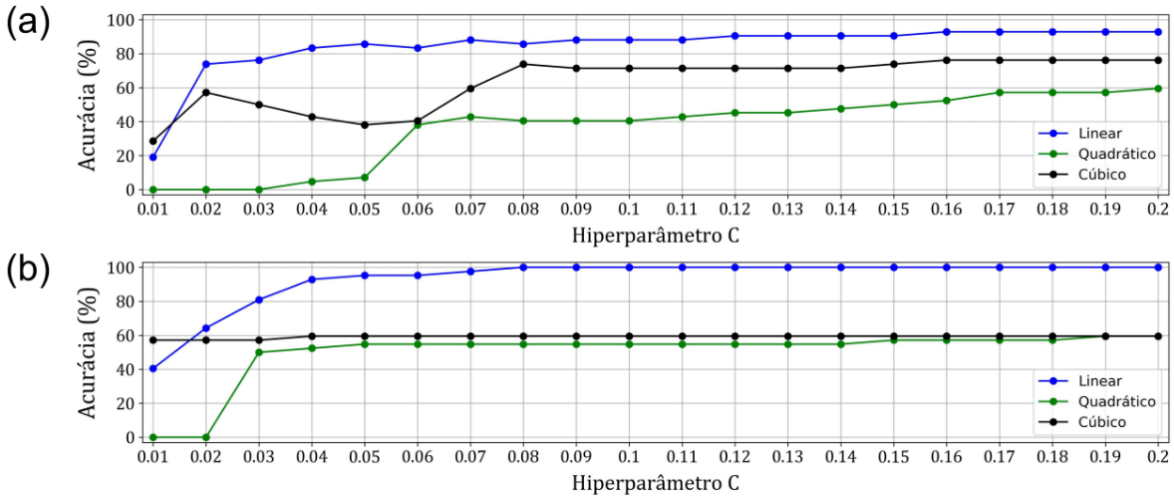


Fig. 5: Acurácias de validação obtidas por LOOCV para diferentes valores de hiperparâmetro C obtidas a partir das amostras de soro seco em estufa (a) e soro líquido com *background* em água deionizada (b), comparando o desempenho do tipo de preparação para diferentes *kernels* utilizando as duas primeiras componentes principais (PC1 e PC2).

Como esperado, o *kernel* linear do classificador Máquina de Vetores de Suporte (SVM), apresentou melhor desempenho tanto para o soro seco quanto para o soro líquido, alcançando uma acurácia de 92,9% para o soro seco com C = 0,16 (a) e 100% de acurácia para o soro líquido com C = 0,08 (b).

A Fig. 6 exibe as matrizes de confusão de validação e teste dos modelos regularizados para as amostras de soro seco (a, b) e soro líquido (c, d), comparando as previsões do modelo com os rótulos reais. O objetivo é avaliar o desempenho do modelo treinado com parâmetros ótimos em dados desconhecidos (teste).

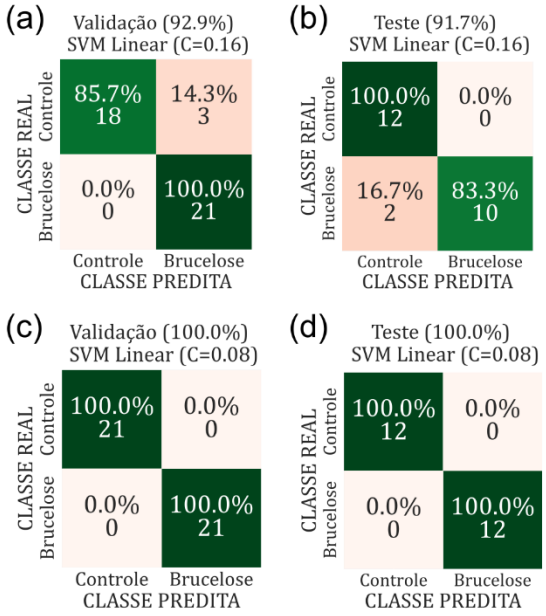


Fig. 6: Matrizes de confusão para as amostras de soro seco (a, b) e soro líquido (c, d). As matrizes da esquerda são da validação (LOOCV) e as matrizes da direita são os resultados do teste. As porcentagens na diagonal indicam especificidade e sensibilidade, respectivamente e sua média é a acurácia indicada no topo. Cada coluna representa as previsões feitas pelo classificador e cada linha representa as classes reais.

No diagnóstico de triagem, uma alta sensibilidade (taxa de verdadeiros positivos) é essencial para evitar a permanência de animais infectados no rebanho. Para o soro seco, a validação mostrou sensibilidade de 100%, enquanto o teste apresentou 83,3%. Esse desempenho reduzido pode ser explicado pela sobreposição entre os grupos de positivos e negativos, que dificultou a definição de uma fronteira de decisão clara. Por outro lado, para o soro líquido, a sensibilidade foi de 100% tanto na validação quanto no teste. Esses resultados indicam que o uso do soro líquido no diagnóstico da Brucelose bovina é promissor comparado aos testes atuais.

A Fig. 7 apresenta as fronteiras de decisão geradas pelo classificador SVM linear para o soro seco (a) e o soro líquido (b). As áreas coloridas representam as regiões de decisão do classificador, com tons mais intensos indicando maior confiança na classificação, enquanto as regiões mais claras representam zonas de maior incerteza. Os círculos e quadrados correspondem às amostras de treino utilizadas no PCA, enquanto os contornos dessas figuras representam as amostras de teste projetadas no mesmo espaço.

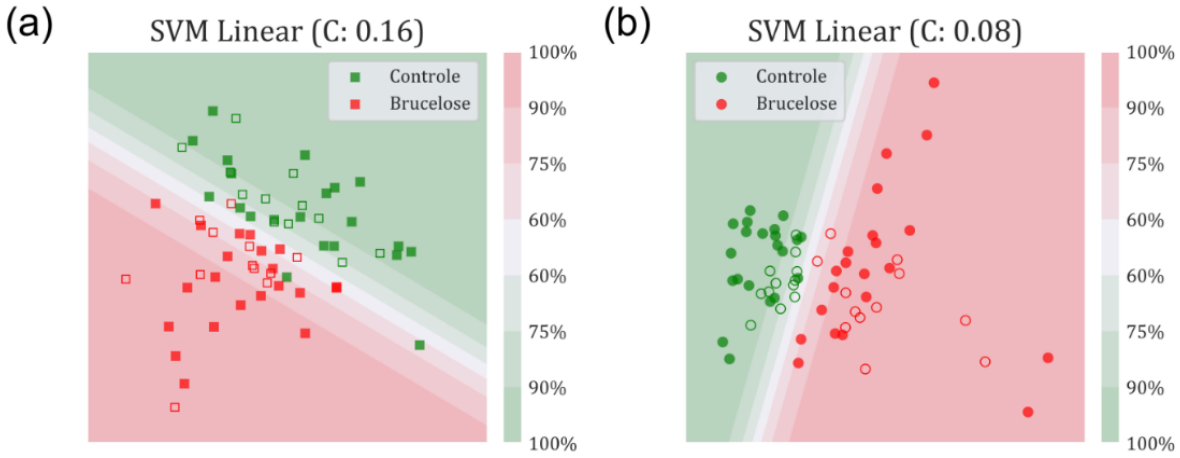


Fig. 7: Fronteiras de decisão para amostras de soro seco em estufa (a) e soro líquido com *background* em água deionizada (b). A cor mais clara indica maior incerteza sobre a previsão enquanto a mais escura indica maior certeza. A cor neutra é a margem de decisão, onde pode ocorrer mistura de amostras de classes diferentes.

Na figura, nota-se que o modelo foi regularizado adequadamente, já que ambas as fronteiras (a, b) apresentam predições com probabilidades superiores a 90% para a maior parte das amostras. Um valor maior de C resulta em uma margem mais estreita, o que aumenta a precisão das predições, mas com maior risco de *overfitting*. Por outro lado, um valor menor de C permite margens mais amplas, favorecendo a generalização, mas reduzindo a precisão nas predições. A escolha ideal do valor de C depende do conjunto de dados. No presente caso, as acurácias de validação e teste, juntamente com as probabilidades, indicam que o modelo foi ajustado adequadamente.

5.2 Classificação Binária de Brucelose e Tuberculose Bovina

Na Fig. 8 são comparadas as médias dos espectros para as classes Controle x Tuberculose (a) e Controle x Brucelose (b), ambas obtidas para soro líquido. O objetivo é identificar padrões ou características espectrais específicas que possam distinguir as classes.

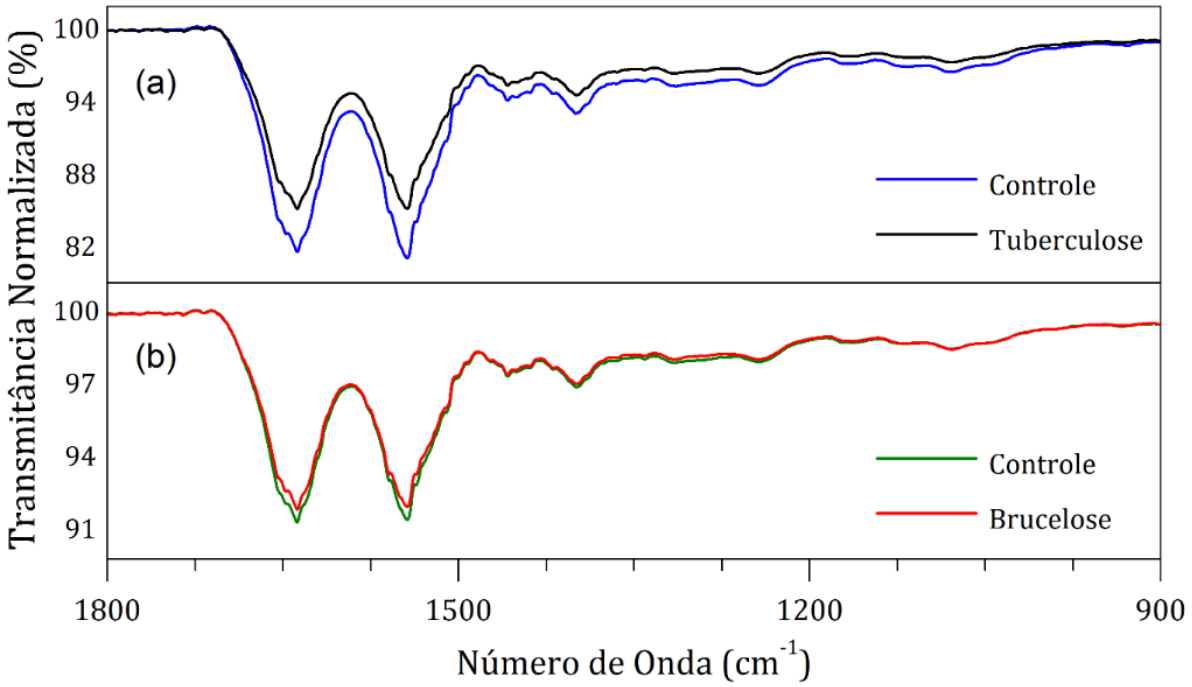


Fig. 8: Média dos espectros FTIR de soro líquido com *background* em água deionizada normalizados por SNV das amostras Controle x Tuberculose (a) e Controle x Brucelose (b).

A média da relação sinal-ruído (SNR) dos espectros dos grupos Controle e Tuberculose foi de 48,4 ± 19,5 sem a aplicação do filtro Savitzky-Golay (SG). Após a aplicação do filtro (janela: 3, grau: 1), a SNR aumentou para 70,0 ± 17,6. Em comparação, os espectros de soro líquido apresentaram uma SNR de 37,5 ± 9,6 nos dados brutos e de 56,2 ± 15,1 após o tratamento com o mesmo filtro. Esses resultados indicam que o uso do filtro SG melhora significativamente a qualidade do sinal, especialmente nos espectros dos grupos Controle e Tuberculose (+21.6), em comparação com os de soro líquido (+18).

As amostras de Controle e Tuberculose (a) apresentaram amplitudes quase duas vezes maiores do que as amostras de Controle e Brucelose (b). Este resultado sugere que o uso de amostras de soro sanguíneo em estado líquido pode estar sujeito a grandes variações na transmitância, devido a diferentes concentrações de analito no meio aquoso.

Os espectros do grupo Controle apresentaram maior transmitância em comparação com as amostras de Tuberculose (a), assim como nas amostras de Brucelose (b). Porém, essa

diferença foi observada em todas as bandas espectrais, com destaque para a região das amidas, sugerindo que outros fatores além das características específicas das doenças podem estar influenciando os espectros. Mais investigações são necessárias para identificar esses fatores.

As amostras de soro líquido apresentaram praticamente as mesmas bandas e não foi possível identificar biomarcadores espectrais que permitissem a classificação específica dos espectros de Controle e Tuberculose (a), assim como no caso da Brucelose (b). Esse resultado reafirma a necessidade do uso da análise multivariada e do aprendizado de máquina para o desenvolvimento de diagnósticos precisos e confiáveis.

Na Fig. 9 os *scores* e os *loadings* das duas primeiras componentes principais (PC1 e PC2) dos espectros de soro líquido para Controle x Tuberculose (a, b) e para Controle x Brucelose (c, d) são comparados com o objetivo de avaliar o agrupamento das classes, bem como identificar as bandas de contribuição para a formação desses grupos.

Nos *scores* (a, c) os círculos são as amostras de treino utilizadas no PCA, enquanto as circunferências são as amostras de teste projetadas neste mesmo espaço. Para facilitar a interpretação dos *loadings* (b, d), a média dos espectros é exibida em cinza.

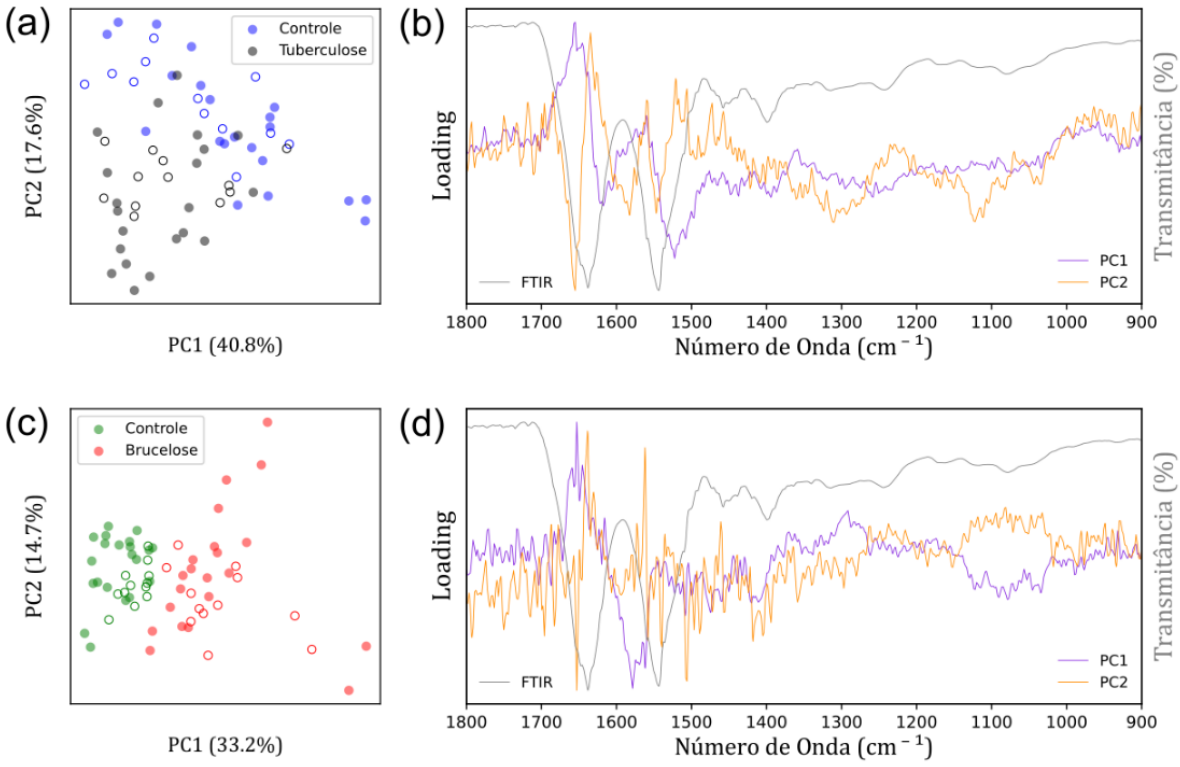


Fig. 9: *Scores* (gráfico de dispersão) e *loadings* (gráfico de linha) para as amostras de Controle x Tuberculose (a, b) e Controle x Brucelose (c, d). Nos *loadings (b, d)* a média de todos os espectros é exibida na cor cinza para comparação entre os pesos e as bandas do espectro das amostras.

Nos *scores* de Controle x Tuberculose (a), observou-se uma maior dispersão entre os grupos em comparação com os *scores* de Controle x Brucelose (b). Isso sugere uma maior variabilidade nas amostras de Tuberculose, refletindo uma maior heterogeneidade nas amostras para ambas as classes. Essa dispersão mais ampla pode indicar a influência de outros fatores que afetam o despenho da análise, que precisam ser melhor investigados.

Observa-se a formação de grupos distintos para cada classe, e as amostras de treino e teste estão agrupadas de forma consistente. No entanto, devido à grande dispersão e sobreposição entre os grupos, a fronteira de decisão não é clara, o que sugere que o modelo pode ter dificuldades para separar as classes de maneira eficiente.

A variância explicada acumulada pelas duas primeiras PC's foi de 58,4% para Controle x Tuberculose, um valor bastante próximo de 47,9% para Controle x Brucelose. Embora as variâncias sejam próximas, as amostras de Controle x Tuberculose não formaram grupos tão bem definidos quanto as amostras de Controle x Brucelose. Isso sugere que, além das características das doenças, algum outro fator pode ter afetado a distribuição das amostras de Controle x Tuberculose, dificultando sua separação adequada.

Os *loadings* da PC1 para Controle x Tuberculose (b) apresentam um padrão ruidoso ao longo de todo o intervalo espectral, o que sugere a interferência de ruídos, mas ainda assim é possível associar esses pesos a contribuições semelhantes às observadas para Controle x Brucelose (d). Na PC2, destacam-se as magnitudes nas regiões da amida I e III (1317 cm^{-1}) e na região associada ao fosfato (1081 cm^{-1}), indicando que esses pesos contribuíram na formação dos grupos, apesar da grande dispersão observada.

Não foi possível identificar biomarcadores espectrais nos *loadings* que distinguissem de forma específica os espectros de Controle x Tuberculose, bem como Controle x Brucelose, devido à contribuição conjunta de diferentes componentes (lipídios, proteínas e carboidratos). Novas análises são necessárias para identificar as regiões capazes de diferenciar uma amostra de controle de uma amostra de animal infectado, assim como as regiões ou bandas que possam distinguir essas doenças.

A Fig. 10 compara as acurácias obtidas por LOOCV para os mesmos valores de hiperparâmetros C utilizados anteriormente, considerando os *kernels* linear, quadrático e cúbico, para Controle x Tuberculose (a) e Controle x Brucelose (b), utilizando as duas primeiras PCs.

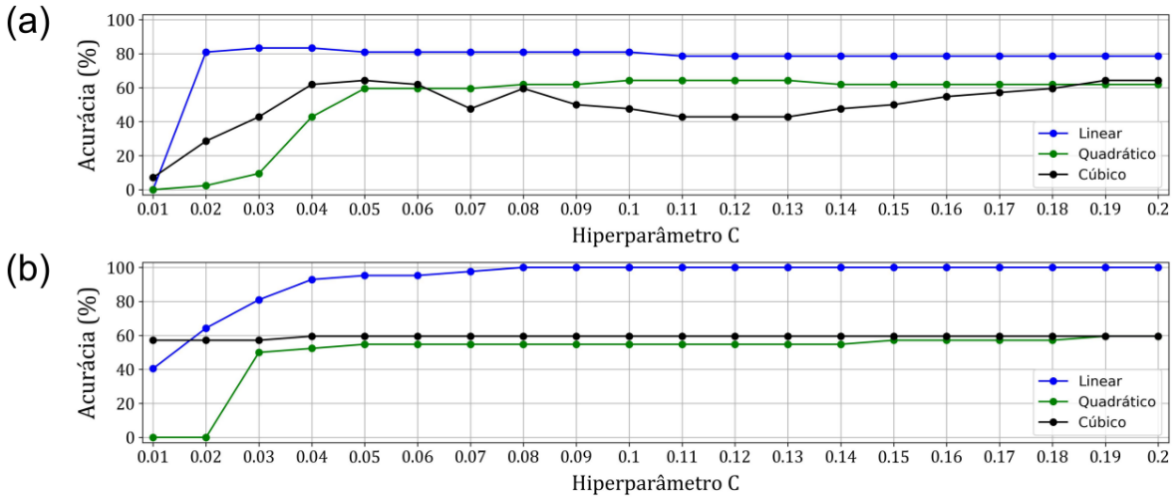


Fig. 10: Acurácias de validação obtidas por LOOCV para diferentes valores de hiperparâmetro C obtidas a partir das amostras de soro líquido com *background* em água deionizada, comparando o desempenho das amostras Controle x Tuberculose (a) e Controle x Brucelose (b) para diferentes *kernels* utilizando as duas primeiras componentes principais (PC1 e PC2).

O *kernel* linear apresentou melhor desempenho para Controle x Tuberculose (a) indicando que o conjunto de dados possui uma fronteira linear, assim como para Controle x Brucelose (b). A acurácia máxima foi de 83,3% com C igual a 0,03, um valor de hiperparâmetro próximo ao observado para Controle e Brucelose (0,08). Além disso, os valores utilizados no filtro SG (janela: 3, grau: 1) contribuíram para a definição de uma fronteira linear em ambos os casos. O uso de uma janela pequena e de um grau baixo mostrou-se eficaz para eliminar flutuações menores e destacar padrões de maior escala. Esses resultados sugerem que essa faixa de valores é apropriada para amostras de soro líquido com fundo de água deionizada e SNR variando entre 27,9 e 67,9.

A Fig. 11 exibe as matrizes de confusão de validação e teste dos modelos regularizados para as amostras de Controle x Tuberculose (a, b) e Controle x Brucelose (c, d), comparando as previsões do modelo com os rótulos reais. O objetivo é avaliar o desempenho do modelo treinado com parâmetros ótimos em dados desconhecidos (teste).

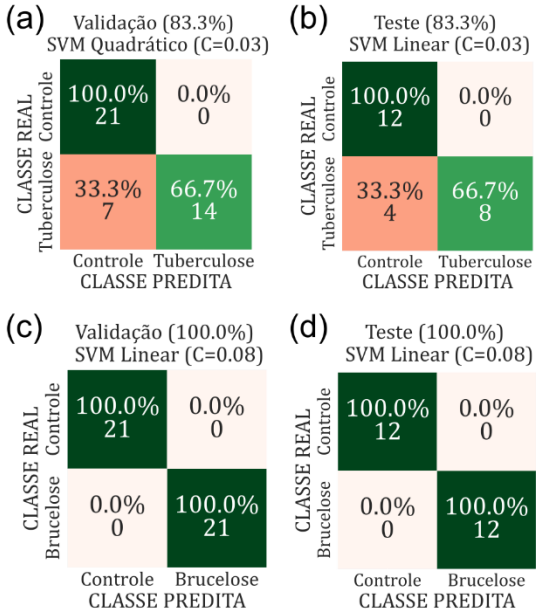


Fig. 11: Matrizes de confusão para as amostras de soro líquido com *background* em água deionizada comparando o desempenho das amostras Controle x Tuberculose (a, b) e Controle x Brucelose (c, d). As matrizes da esquerda são da validação (LOOCV) e as matrizes da direita são os resultados do teste. As porcentagens na diagonal indicam especificidade e sensibilidade, respectivamente e sua média é a acurácia indicada no topo. Cada coluna representa as previsões feitas pelo classificador e cada linha representa as classes reais.

No diagnóstico de triagem, é ideal que a sensibilidade (taxa de verdadeiros positivos) seja alta para permanência de animais infectados no rebanho. Na comparação entre Controle e Tuberculose (a, b), a sensibilidade foi de 66,7% tanto na validação quanto no teste. Embora o modelo esteja bem equilibrado, o desempenho foi abaixo do esperado, possivelmente devido à grande variação dos dados, que resultou em grupos sobrepostos e na ausência de uma fronteira clara.

Embora as amostras de Controle x Tuberculose tenham apresentado desempenho inferior às de Controle x Brucelose, os resultados sugerem que o uso de soro líquido no diagnóstico da Tuberculose bovina é promissor. No entanto, são necessários mais estudos para avaliar a influência de outros fatores que podem interferir na variância das amostras, dificultando a formação de grupos no PCA e prejudicando a robustez do modelo de aprendizado de máquina.

A Fig. 12 apresenta as fronteiras de decisão geradas pelo classificador SVM linear para Controle x Tuberculose (a) e Controle x Brucelose (b). As áreas coloridas representam as regiões de decisão do classificador, com tons mais intensos indicando maior confiança na classificação, enquanto as regiões mais claras representam zonas de maior incerteza. Os círculos e quadrados correspondem às amostras de treino utilizadas no PCA, enquanto os contornos dessas figuras representam as amostras de teste projetadas no mesmo espaço.

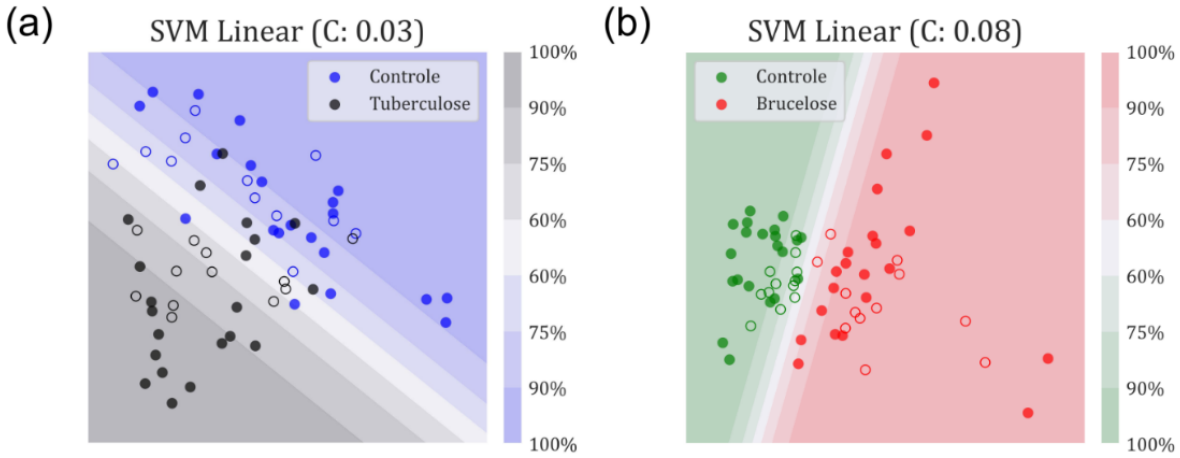


Fig. 12: Fronteiras de decisão para amostras de soro líquido com *background* em água deionizada, comparando a regularização e porcentagem de certeza de predição para as amostras Controle x Tuberculose (a) e Controle x Brucelose (b). A cor mais clara indica maior incerteza sobre a previsão enquanto a mais escura indica maior certeza. A cor neutra é a margem de decisão, onde pode ocorrer mistura de amostras de classes diferentes.

Na figura, observa-se que o modelo treinado para Controle x Tuberculose (a) apresenta uma de probabilidade de confiança de 60% maior em comparação com o modelo para Controle vs. Brucelose (b). Isso sugere que o modelo para Tuberculose permite um número maior de erros de classificação para garantir uma margem mais ampla e robusta. Embora a maioria das predições não apresente probabilidades superiores a 90%, as acurácias de validação, teste e as distribuições das probabilidades, indicam que o modelo foi ajustado de maneira adequada.

5.3 Classificação Multiclasse de Brucelose e Tuberculose Bovina

Na Fig. 13, são comparadas as médias dos espectros para as classes Brucelose, Controle e Tuberculose obtidos a partir de soro líquido. Para a classe Controle, foram utilizadas metade das amostras empregadas nas análises binárias anteriores. O objetivo dessa comparação é identificar padrões ou características espectrais específicas que possam distinguir entre as classes.

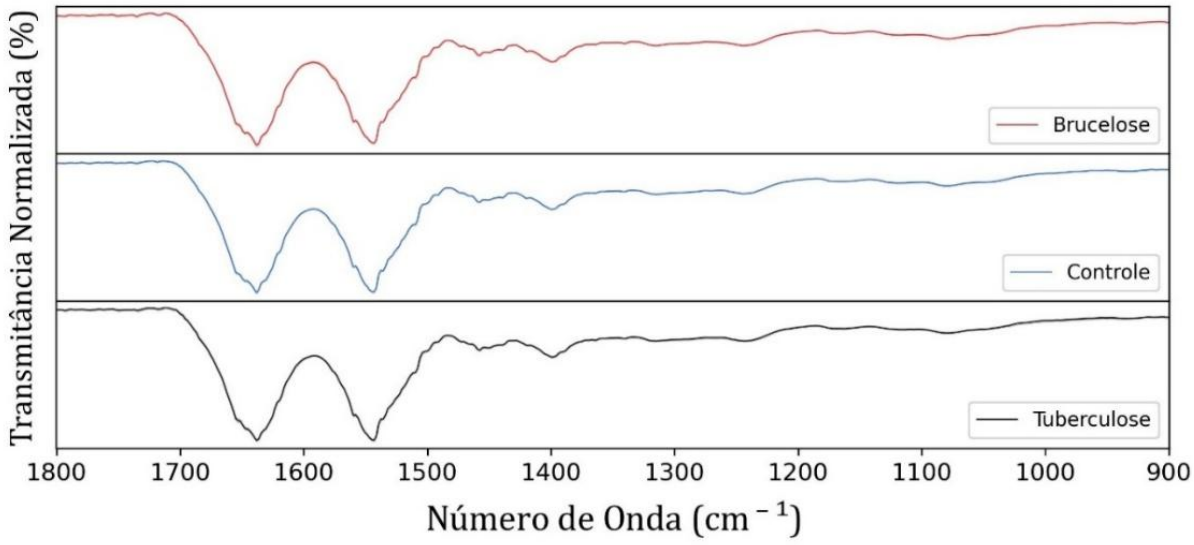


Fig. 13: Média dos espectros FTIR de soro líquido com *background* em água deionizada normalizados por SNV das amostras de Brucelose, Controle e Tuberculose.

Assim como já foi verificado, as mostras de soro líquido apresentaram praticamente as mesmas bandas e não foi possível identificar biomarcadores espectrais que permitissem a classificação específica dos espectros das classes Brucelose, Controle e Tuberculose.

Na Fig. 14 os *scores* e os *loadings* das duas primeiras componentes principais dos espectros de soro líquido para Brucelose x Controle x Tuberculose (a, b) são comparados com o objetivo de avaliar o agrupamento das classes, bem como identificar as bandas de contribuição para a formação desses grupos.

Nos *scores*, observa-se a formação de grupos distintos para as amostras Controle de Brucelose (Controle B) e Controle de Tuberculose (Controle T), confirmando uma maior heterogeneidade nas amostras de Controle T e Tuberculose. Além disso, as amostras de Tuberculose formaram subgrupos, sugerindo diferentes níveis de heterogeneidade, causados por fatores desconhecidos, que carecem de mais estudos para esclarecimento.

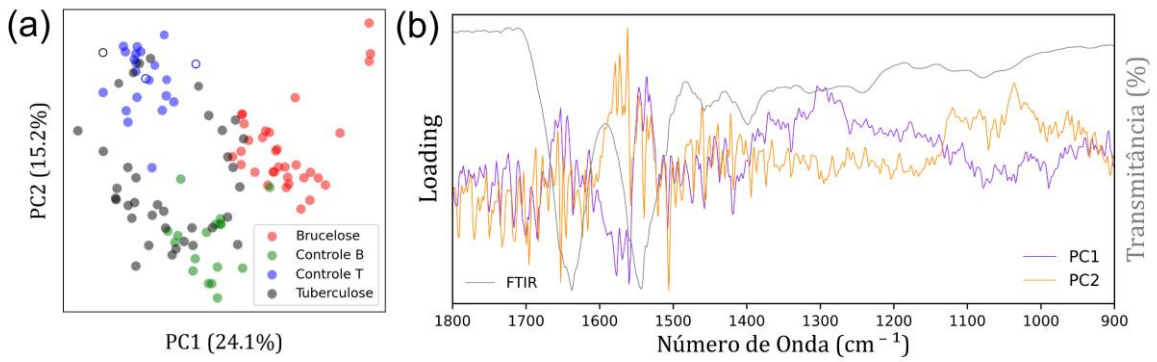


Fig. 14: *Scores* (a) e *loadings* (b) do PCA obtido pelas amostras de Brucelose, Controle e Tuberculose. Nos *scores* (a), as amostras de Controle são coloridas de acordo com a sua origem compartilhada com as amostras de animais infectados, destacando a formação de subgrupos. Nos *loadings* (b), a média de todos os espectros, apresentada em cinza, serve como referência para comparação com os pesos das variáveis espectrais.

Embora a primeira componente principal (PC1) consiga separar os grupos Controle e Brucelose, a dispersão e sobreposição dos *scores* das amostras de Tuberculose sugere que o modelo terá dificuldades para classificar adequadamente utilizando apenas as duas primeiras componentes principais (PC's). Os *loadings* serão discutidos após a definição dessas PC's. A variância explicada acumulada pelas duas primeiras PC's foi de 39,3%, um valor menor que o obtido para Controle x Brucelose (47,9%) e Controle x Tuberculose (58,4%). Esse resultado é esperado em análises de PCA multiclasse, devido à maior complexidade associada à separação de múltiplos grupos simultaneamente.

A próximas figuras comparam as acurácias obtidas por LOOCV para os valores de hiperparâmetros C previamente utilizados na classificação binária, considerando as 20 primeiras PC's. Os resultados são apresentados para os *kernels* linear (Fig. 15), quadrático (Fig. 16) e cúbico (Fig. 17).

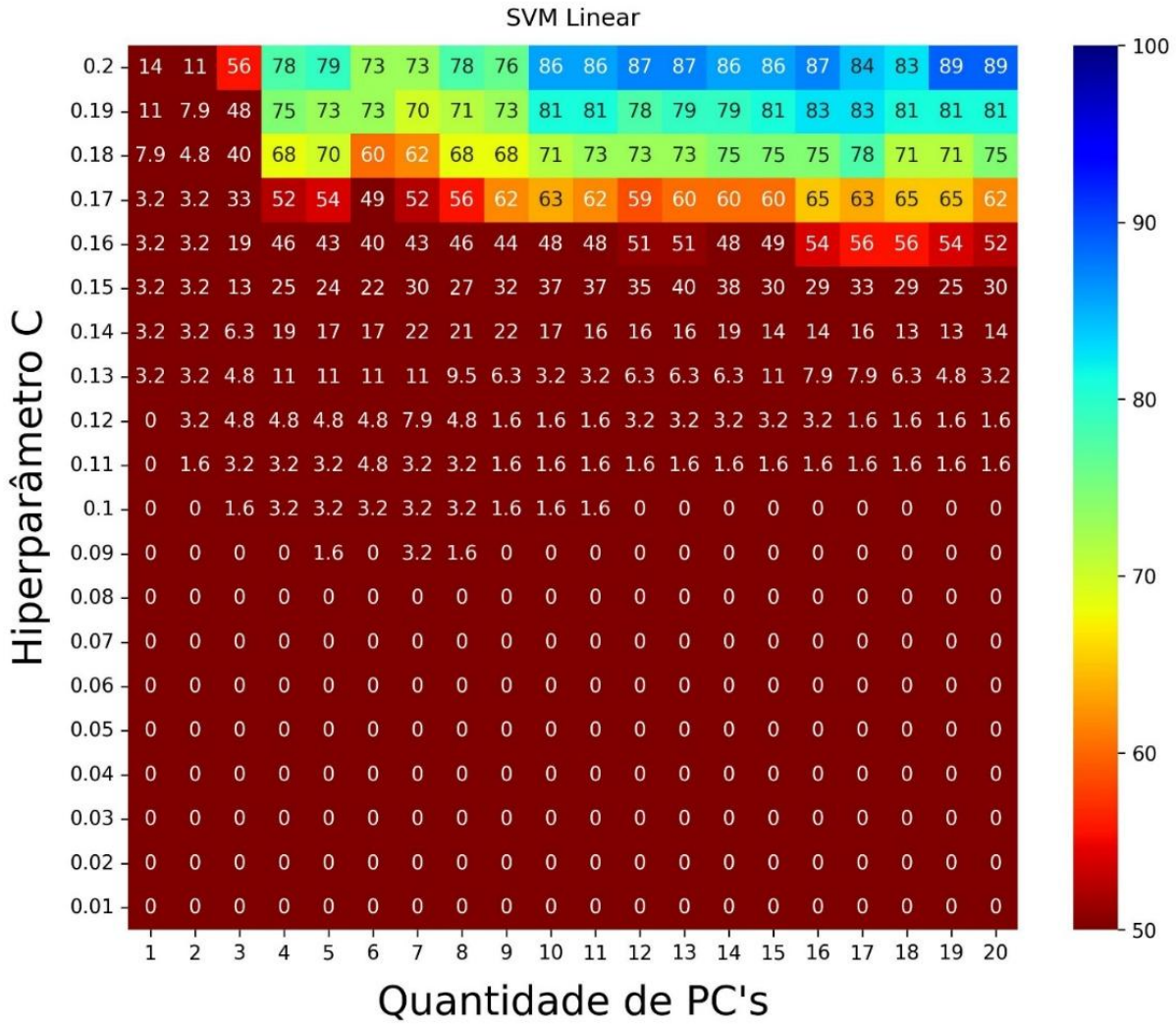


Fig. 15: Acurácias de validação obtidas por LOOCV para diferentes valores de hiperparâmetro C obtidas a partir das amostras de soro líquido com *background* em água deionizada, comparando o desempenho de diferentes quantidades de PC's, para a análise multiclasse Brucelose x Controle x Tuberculose para o *kernel* linear. O mapa de cores indica a acurácia no intervalo de 50 até 100%.

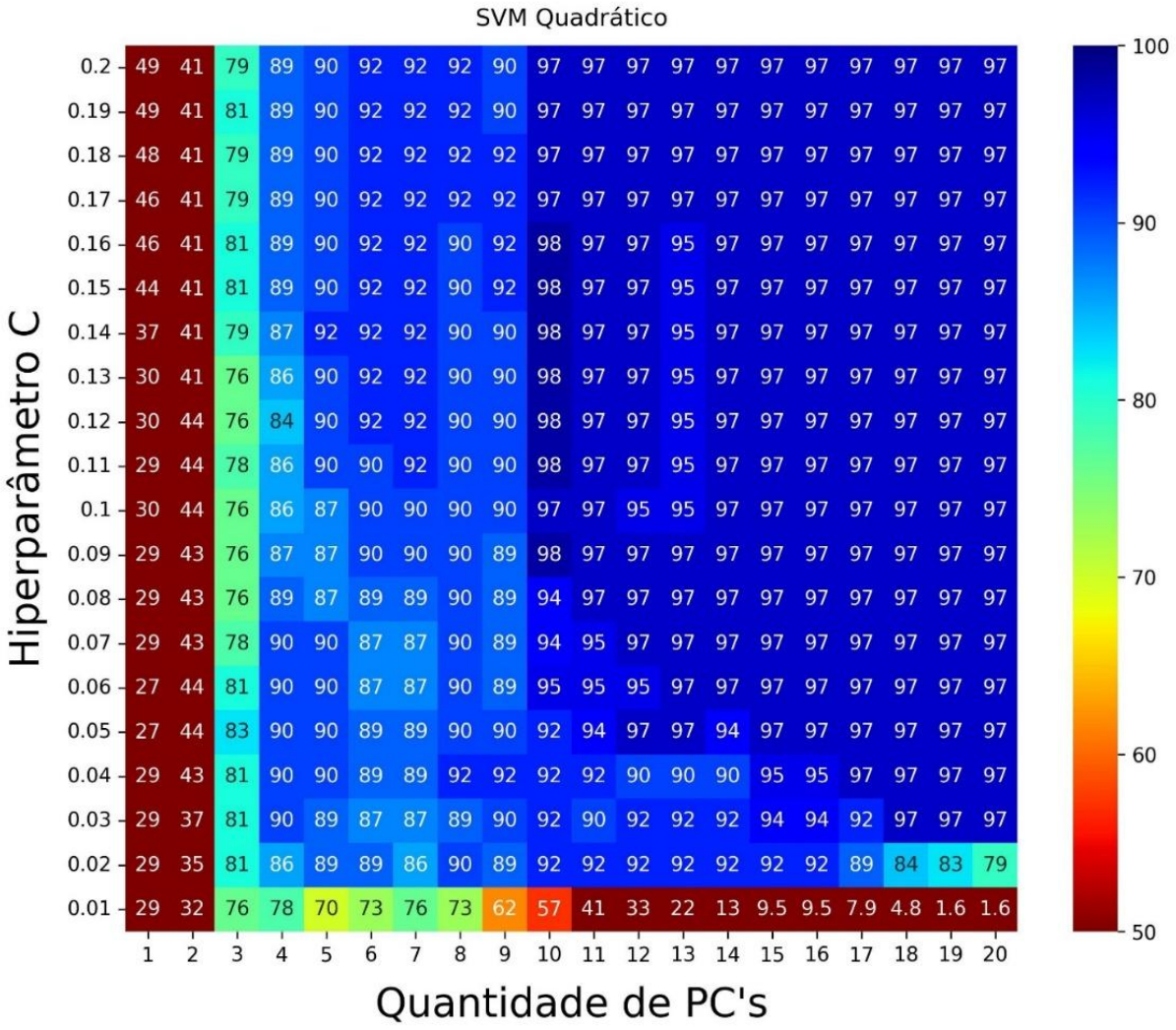


Fig. 16: Acurácias de validação obtidas por LOOCV para diferentes valores de hiperparâmetro C obtidas a partir das amostras de soro líquido com *background* em água deionizada, comparando o desempenho de diferentes quantidades de PC's, para a análise multiclasse Brucelose x Controle x Tuberculose para o *kernel* quadrático. O mapa de cores indica a acurácia no intervalo de 50 até 100%.

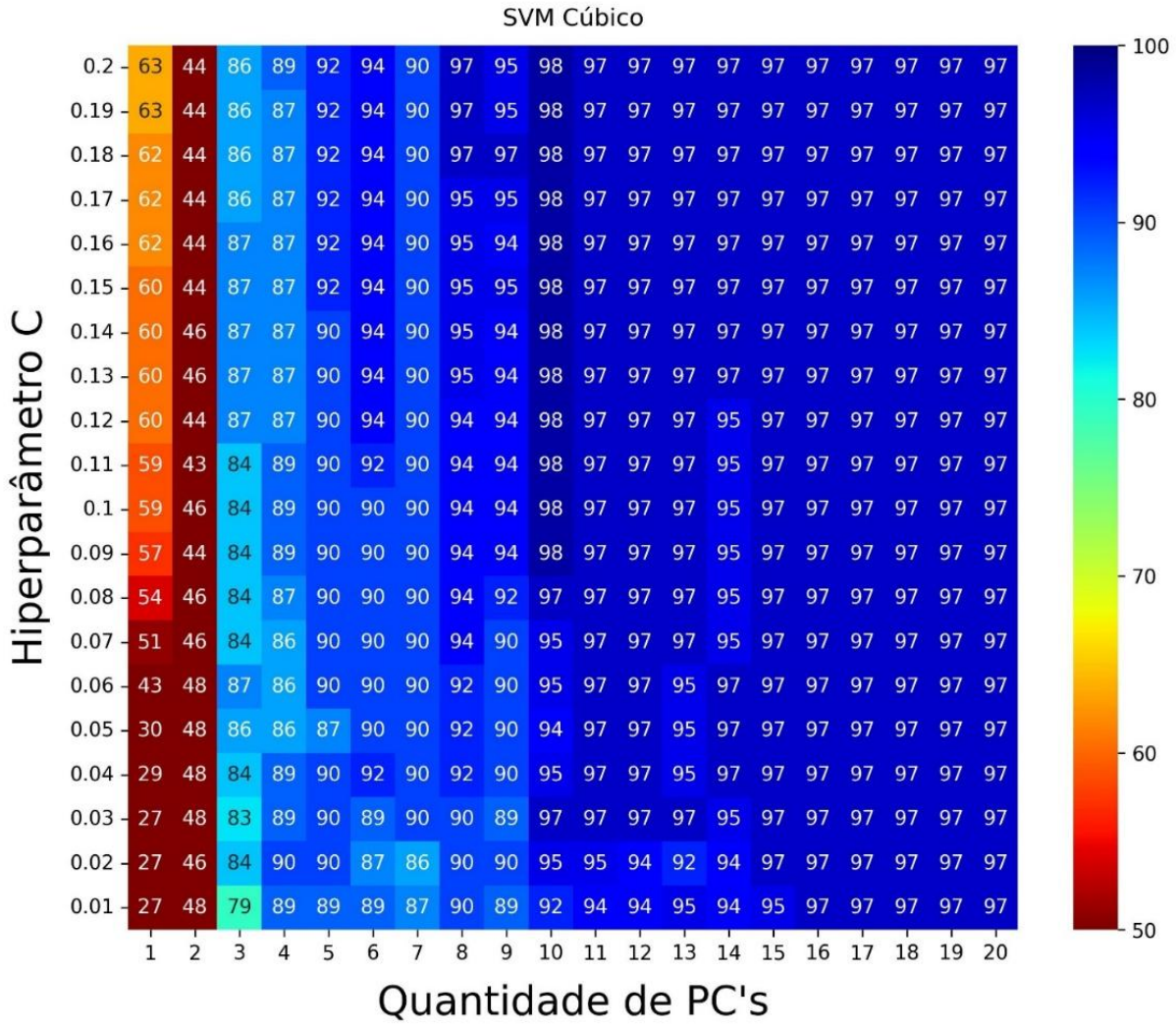


Fig. 17: Acurácias de validação obtidas por LOOCV para diferentes valores de hiperparâmetro C obtidas a partir das amostras de soro líquido com *background* em água deionizada, comparando o desempenho de diferentes quantidades de PC's, para a análise multiclasse Brucelose x Controle x Tuberculose para o *kernel* cúbico. O mapa de cores indica a acurácia no intervalo de 50 até 100%.

O *kernel* cúbico apresentou o melhor desempenho em comparação aos demais. O valor de C para este *kernel* foi escolhido com base na obtenção de uma acurácia superior a 90%, utilizando o menor número possível de componentes principais (PC's). O modelo selecionado alcançou uma acurácia de 90,5% com 4 PC's utilizando C = 0,02, um valor próximo ao observado para a comparação Controle x Brucelose (C = 0,03).

A Fig. 18 exibe as matrizes de confusão de validação (a) e teste (b) dos modelos regularizados para as amostras de Brucelose x Controle x Tuberculose, comparando as previsões do modelo com os rótulos reais. O objetivo é avaliar o desempenho do modelo treinado com parâmetros ótimos em dados desconhecidos (teste).

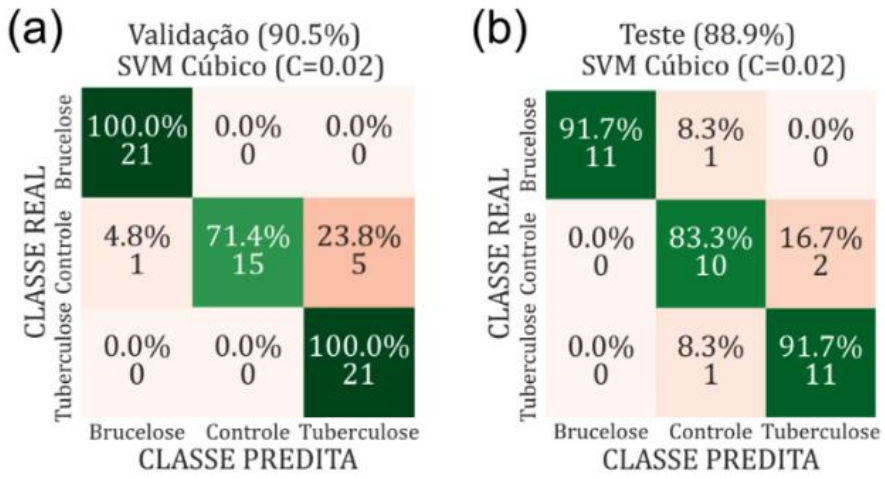


Fig. 18: Matrizes de confusão para as amostras de soro líquido com *background* em água deionizada comparando o desempenho da validação (a) e teste (b) para a análise multiclasse Brucelose x Controle x Tuberculose. A matriz da esquerda é da validação (LOOCV) e a matriz da direita são os resultados do teste. As porcentagens na diagonal indicam sensibilidade para brucelose, especificidade e sensibilidade para tuberculose, respectivamente e a média desses valores é a acurácia indicada no topo. Cada coluna representa as previsões feitas pelo classificador e cada linha representa as classes reais.

Na classificação multiclasse, a sensibilidade foi de 100% para validação e 91,7% para teste, indicando um modelo bem equilibrado. Esse desempenho superou o da classificação binária de Controle x Tuberculose, que obteve 66,7%. Esse resultado pode ser atribuído ao uso de um maior número de componentes principais, ou ainda ao fato de que a análise multiclasse contribuiu para uma maior concentração do grupo Controle T (Fig. 14).

A Fig. 19 apresenta uma matriz de fronteiras de decisão geradas pelo SVM cúbico. Cada par de PC's foi utilizado para treinar o modelo regularizado anteriormente e a acurácia obtida é exibida, juntamente com os scores normalizados. As cores representam as diferentes classes que o modelo foi treinado e as regiões delimitadas por essas cores indicam as áreas do espaço de características onde o modelo atribui uma classe particular a uma amostra desconhecida.

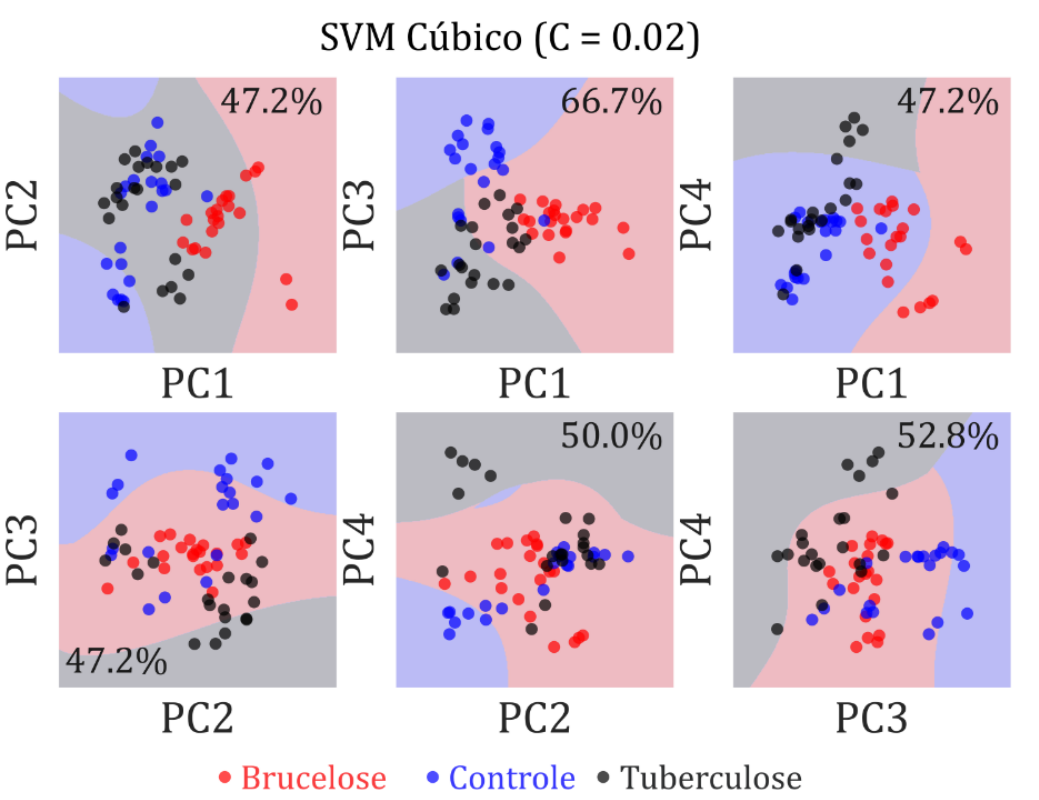


Fig. 19: Fronteiras de decisão para amostras de soro líquido com *background* em água deionizada, avaliando o desempenho de diferentes combinações de PC's na análise multiclasse Brucelose x Controle x Tuberculose. As cores indicam as classes utilizadas no treinamento do modelo, enquanto as regiões delimitadas representam as áreas do espaço de características onde o modelo atribui uma classe específica a amostras desconhecidas. As amostras usadas no treinamento do classificador estão sobrepostas no gráfico, permitindo visualizar seu agrupamento e a sua distribuição em relação às fronteiras de decisão. No alto é indicado a acurácia de validação (LOOCV) para o modelo treinado com essas combinações.

A combinação da PC1 e PC3 apresentam a maior acurácia (66,7%) valor igual a acurácia obtida pela classificação binária de Controle x Tuberculose. Essa combinação indica que a PC2 pode estar carregando a variância referente a heterogeneidade observada nas amostras. Além disso essa combinação de PC's permitiu uma melhor separação entre os grupos de cada classe apesar da sobreposição.

A Fig. 20 exibe os *scores* e *loadings* das 4 primeiras PC's para Brucelose x Controle x Tuberculose, com o objetivo de identificar as bandas de contribuição para a formação dos grupos observados na Fig. 19. Essa visualização permite compreender como as diferentes bandas contribuem para a separação e formação dos grupos no espaço de características.

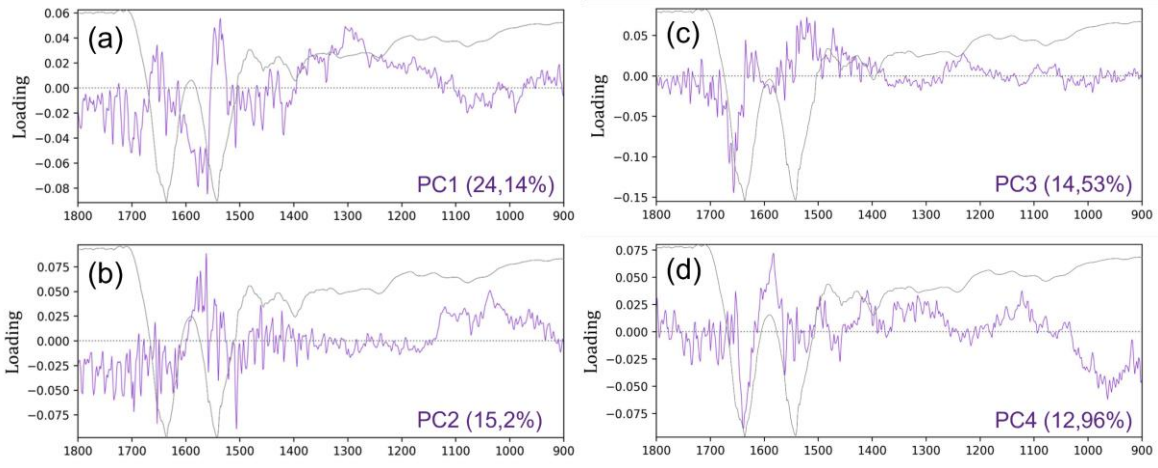


Fig. 20: *Loadings* do PCA obtido pelas amostras de Brucelose, Controle e Tuberculose. A média de todos os espectros, apresentada em cinza, serve como referência para comparação com os pesos das variáveis espectrais.

Os *scores* indicam que as PC's não são responsáveis por separar as classes de forma independente, sendo a combinação dessas componentes que favorece a formação dos grupos. Já os *loadings*, embora apresentem um padrão ruidoso, capturam variações relevantes. Observa-se que as regiões com maior peso nos *loadings* coincidem com as áreas de maior intensidade no espectro médio, reforçando a contribuição dessas bandas para as diferenças bioquímicas observadas entre as classes

6. CONCLUSÃO

Os resultados indicaram diferenças significativas entre os espectros de Soro Seco e Soro Líquido. O Soro Seco apresentou maior relação sinal-ruído (SNR), enquanto o Soro Líquido demonstrou maior homogeneidade, bandas espectrais mais bem definidas e melhor preservação das informações, especialmente no que se refere às proteínas. O soro líquido formou grupos mais definidos, alcançando uma acurácia de 100%, superando o Soro Seco (92,9%) e os métodos diagnósticos convencionais. Com sensibilidade de 100% tanto na validação quanto no teste, o Soro Líquido revelou-se mais promissor para o diagnóstico da Brucelose bovina, destacando seu potencial como ferramenta diagnóstica eficaz.

A maior dispersão observada nos *scores* entre os grupos Controle e Tuberculose, em comparação com Controle e Brucelose, sugere uma maior heterogeneidade nas amostras, demonstrando a interferência de outros fatores, que precisam ser melhor investigados. Não foi possível identificar biomarcadores espectrais específicos nos *loadings*, uma vez que lipídios, proteínas e carboidratos contribuíram conjuntamente para os resultados das duas doenças. A acurácia para a distinção entre Controle e Tuberculose foi de 83,3%, com sensibilidade de 66,7% em ambas as fases de validação e teste. Esse desempenho inferior ao esperado pode ser atribuído à grande variância nas amostras, que resultou em grupos sobrepostos e na falta de uma fronteira clara. No entanto, os dados sugerem que o soro líquido apresenta um potencial promissor para o diagnóstico da Tuberculose bovina.

Na análise de classificação multiclasse, o uso do kernel cúbico resultou em uma acurácia de 90,5%, com sensibilidade de 100% na validação e 91,7% no teste, superando a classificação binária de Controle x Tuberculose. Esse desempenho superior pode ser atribuído ao uso de um maior número de componentes principais, o que proporcionou uma classificação mais precisa e robusta.

Portanto, a espectroscopia de Infravermelho por Transformada de Fourier (FTIR), combinada com análise multivariada e técnicas de *machine learning*, mostrou-se eficaz para discriminar entre animais controle e infectados com Brucelose e Tuberculose bovina. Esta abordagem tem grande potencial para a triagem rápida de animais, dispensando o uso de estufas para remoção de água nas amostras. Além disso, permite monitorar a saúde dos animais *in loco* de forma eficiente, o que pode contribuir significativamente para o controle da propagação dessas doenças, tanto entre animais quanto entre seres humanos.

7. SUGESTÕES DE TRABALHOS FUTUROS

Identificação de Biomarcadores Espectrais Específicos: Desenvolver métodos avançados de análise espectral e aprendizado de máquina para identificar biomarcadores espectrais específicos associados à brucelose e à tuberculose bovina, melhorando a precisão e a confiabilidade do diagnóstico.

Avaliação de Outras Matrizes Biológicas: Ampliar o estudo para incluir outras matrizes biológicas, como saliva ou leite, utilizando FTIR, para avaliar a aplicabilidade do método em diferentes contextos diagnósticos, especialmente em condições de campo.

Diagnóstico Multiespectral e Multiespécies: Ampliar a abordagem para diagnóstico de outras zoonoses em diferentes espécies animais, utilizando espectroscopia combinada (FTIR e Raman) e aprendizado de máquina, para criar uma ferramenta diagnóstica universal.

Implementação de Diagnóstico *In Loco*: Desenvolver e validar dispositivos portáteis baseados em espectroscopia FTIR para uso em campo, com foco na integração de hardware e software que permita análise rápida e sem a necessidade de infraestrutura laboratorial complexa.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ANDRADE, R. S. et al. Accuracy of serological tests for bovine brucellosis: A systematic review and meta-analysis. **Preventive Veterinary Medicine**, v. 222, p. 106079, 2024.

[2] HAYOUN, M.; MUCO, E.; SHORMAN, M. Brucellosis. 2023 Apr 29. **StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing**, 2023.

[3] MCKENNA, S. L.; DOHOO, I. R. Using and interpreting diagnostic tests. **Veterinary Clinics: Food Animal Practice**, v. 22, n. 1, p. 195–205, 2006.

[4] BULASHEV, A. K. et al. Evaluation of chimeric proteins for serological diagnosis of brucellosis in cattle. **Veterinary World**, v. 14, n. 8, p. 2187, 2021.

[5] ROBI, D. T. et al. Bovine tuberculosis reactor cattle in Southwest Ethiopia: Risk factors for bovine tuberculosis. **Journal of Clinical Tuberculosis and Other Mycobacterial Diseases**, v. 37, p. 100492, 2024.

[6] COLLINS, Á. B. et al. Prevalence of Mycobacterium bovis in milk on dairy cattle farms: An international systematic literature review and meta-analysis. **Tuberculosis**, v. 132, p. 102166, 2022.

[7] TODD, E. C. D. Mycobacterium bovis. Em: SMITHERS, G. W. (Ed.). **Encyclopedia of Food Safety (Second Edition)**. Oxford: Academic Press, 2024. p. 189–200.

[8] WISEMAN, J.; CASSIDY, J.; GORMLEY, E. The problem that residual Mycobacterium bovis infection poses for the eradication of bovine tuberculosis. **The Veterinary Journal**, p. 106266, 2024.

[9] DOU, J. et al. Rapid discrimination of Brucellosis in sheep using serum Fourier transform infrared spectroscopy combined with PCA-LDA algorithm. **Photodiagnosis and Photodynamic Therapy**, v. 42, p. 103567, 2023.

[10] KOCHAN, K. et al. Infrared spectroscopy of blood. **Applied spectroscopy**, v. 75, n. 6, p. 611–646, 2021.

[11] CAMERON, J. M. et al. Exploring pre-analytical factors for the optimisation of serum diagnostics: Progressing the clinical utility of ATR-FTIR spectroscopy. **Vibrational Spectroscopy**, v. 109, p. 103092, 2020.

[12] SALA, A. et al. Rapid analysis of disease state in liquid human serum combining infrared spectroscopy and “digital drying”. **Journal of Biophotonics**, v. 13, n. 9, p. e202000118, 2020.

[13] DE OLIVEIRA, L. F. et al. Seroprevalence and Risk Factors for Bovine Brucellosis in Minas Gerais State, Brazil. **Semina Ciências Agrárias**, 2016.

[14] QUINTERO, A. F. et al. Evaluation of Two Rapid Immunochromatographic Tests for Diagnosis of Brucellosis Infection in Cattle. **Open Veterinary Journal**, 2018.

[15] SANTOS, R. L. et al. Economic Losses Due to Bovine Brucellosis in Brazil. **Pesquisa Veterinária Brasileira**, 2013.

[16] DINE DJIBRIL, A. S. Farmers' Perceptions of Bovine Brucellosis in Benin. **Veterinary World**, 2024.

[17] HOLT, H. et al. Epidemiology of Brucellosis in Cattle and Dairy Farmers of Rural Ludhiana, Punjab. **Plos Neglected Tropical Diseases**, 2021.

[18] ZHANG, N. et al. Brucellosis Awareness and Knowledge in Communities Worldwide: A Systematic Review and Meta-Analysis of 79 Observational Studies. **Plos Neglected Tropical Diseases**, 2019.

[19] ALVES DINIZ, J. V. et al. Brucellosis and Bovine Tuberculosis in Dairy Farms in the State of Acre, Brazil. **Acta Veterinaria Brasilica**, 2021.

[20] CLEMENTINO, I. J.; DE AZEVÊDO, S. S. Bovine Brucellosis: Epidemiological Situation in Brazil and Disease Control Initiatives. **Semina Ciências Agrárias**, 2016.

[21] EOH, H. et al. Expression and validation of D-erythrulose 1-phosphate dehydrogenase from Brucella abortus: a diagnostic reagent for bovine brucellosis. **Journal of Veterinary Diagnostic Investigation**, 2010.

[22] KUMAR, V. et al. Serum Based Polymerase Chain Reaction and Enzyme Linked Immunosorbent Assays for Diagnosis of Bovine Brucellosis. **Indian Journal of Animal Research**, 2018.

[23] LIM, J. J. et al. Evaluation of recombinant 28 kDa outer membrane protein of Brucella abortus for the clinical diagnosis of bovine brucellosis in Korea. **Journal of Veterinary Medical Science**, 2012.

[24] XIN, T. et al. Limitations of the BP26 Protein-Based Indirect Enzyme-Linked Immunosorbent Assay for Diagnosis of Brucellosis. **Clinical and Vaccine Immunology**, 2013.

[25] IBARRA, M. Comparison of Diagnostic Tests for Detecting Bovine Brucellosis in Animals Vaccinated With S19 and RB51 Strain Vaccines. **Veterinary World**, 2023.

[26] NOVAK, A. et al. Development of a Novel Glycoprotein-based Immunochromatographic Test for the Rapid Serodiagnosis of Bovine Brucellosis. **Journal of Applied Microbiology**, 2022.

[27] AMARAL SOUSA, A. K. et al. Bovine Brucellosis in Slaughterhouses Controlled by Federal and Municipal Inspection Services in the State of Maranhão, Brazil. **Arquivos Do Instituto Biológico**, 2019.

[28] YIN, D. et al. A Multiepitope Fusion Protein-Based P-Elisa Method for Diagnosing Bovine and Goat Brucellosis. 2021.

[29] BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa SDA n.º 10, de 3 de março de 2017. Estabelece o Regulamento Técnico do Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal - PNCEBT. Diário Oficial do Brasil, Brasília, DF. Ministério da Agricultura, Pecuária e Abastecimento. . 20 jun. 2017.

[30] CHANG, Y. et al. Evaluating the effectiveness of badger vaccination combined with cattle test-and-removal in managing Bovine Tuberculosis: Insights from a two-host and multi-route transmission model. **Preventive Veterinary Medicine**, p. 106386, 2024.

[31] RAMANUJAM, H.; PALANIYANDI, K. Bovine tuberculosis in India: The need for One Health approach and the way forward. **One Health**, v. 16, p. 100495, 2023.

[32] GUIMARAES, A. M.; ZIMPEL, C. K. Mycobacterium bovis: from genotyping to genome sequencing. **Microorganisms**, v. 8, n. 5, p. 667, 2020.

[33] HASHIMOTO, K. et al. Complementary Vibrational Spectroscopy. **Nature Communications**, v. 10, n. 1, 2019.

[34] LI, X. et al. Fingerprinting a Living Cell by Raman Integrated Mid-Infrared Photothermal Microscopy. **Analytical Chemistry**, v. 91, n. 16, p. 10750–10756, 2019.

[35] MUELLER, N. S. et al. Surface-Enhanced Raman Scattering and Surface-Enhanced Infrared Absorption by Plasmon Polaritons in Three-Dimensional Nanoparticle Supercrystals. **Acs Nano**, v. 15, n. 3, p. 5523–5533, 2021.

[36] DORFMAN, G. et al. Vibrationally Mediated Photodissociation of Jet-Cooled CH3CF2Cl: a Probe of Energy Flow and Bond Breaking Dynamics. **The Journal of Physical Chemistry A**, v. 106, n. 36, p. 8285–8290, 2002.

[37] LIEBL, K. et al. Explaining the Striking Difference in Twist-Stretch Coupling Between DNA and RNA: A Comparative Molecular Dynamics Analysis. **Nucleic Acids Research**, p. gkv1028, 2015.

[38] MOILANEN, D. E. et al. Water Inertial Reorientation: Hydrogen Bond Strength and the Angular Potential. **Proceedings of the National Academy of Sciences**, v. 105, n. 14, p. 5295–5300, 2008.

[39] CHANG, W. et al. Tracking Molecular Structure Deformation of Nitrobenzene and Its Torsion–vibration Coupling by Intense Pumping CARS. **Chinese Physics B**, v. 25, n. 11, p. 114210, 2016.

[40] LI, Y. et al. Tool Condition Monitoring for Cavity Milling Based on Bispectrum Analysis and Bayesian Optimized SVM. 2023.

[41] SANCHÉZ-LOZANO, M. et al. Theoretical Vibrational Raman and Surface-Enhanced Raman Scattering Spectra of Water Interacting with Silver Clusters. **ChemPhysChem**, v. 15, n. 18, p. 4067–4076, 2014.

[42] ŞUVAR, N.-S. et al. **Analysis of thermal degradation behavior for some hydraulic oils, using FTIR-TGA coupling**. . Em: MATEC WEB OF CONFERENCES. EDP Sciences, 2020.

[43] CHEN, K. et al. Improved Savitzky–Golay-method-based fluorescence subtraction algorithm for rapid recovery of Raman spectra. **Applied optics**, v. 53, n. 24, p. 5559–5569, 2014.

[44] SALMI, T. et al. Computational Vibrational and Electronic Spectroscopy of the Water Nitric Oxide Complex. **The Journal of Physical Chemistry A**, v. 114, n. 14, p. 4835–4842, 2010.

[45] CRAVEN-JONES, J.; KUDENOV, M. W.; DERENIAK, E. L. Tunable Interference Contrast Using a Variable Wollaston Prism. **Optical Engineering**, v. 51, n. 1, p. 013002, 2012.

[46] KAPLAN, S. G. et al. Comparison of Near-Infrared Transmittance and Reflectance Measurements Using Dispersive and Fourier Transform Spectrophotometers. **Metrologia**, v. 39, n. 2, p. 157–164, 2002.

[47] PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198–219, 2003.

[48] WEDDING, B. B. et al. Non-Destructive Prediction of 'Hass' Avocado Dry Matter via FT-NIR Spectroscopy. **Journal of the Science of Food and Agriculture**, v. 91, n. 2, p. 233–238, 2010.

[49] KOREPANOV, O. A. et al. Polyvinylpyrrolidone as a Stabilizer in Synthesis of AgInS2 Quantum Dots. **Nanomaterials**, v. 12, n. 14, p. 2357, 2022.

[50] SMITH, B. C. **Fundamentals of Fourier transform infrared spectroscopy**. [s.l.] CRC press, 2011.

[51] AZZAM, R. M. A.; MAHMOUD, F. A. Symmetrically Coated Pellicle Beam Splitters for Dual Quarter-Wave Retardation in Reflection and Transmission. **Applied Optics**, v. 41, n. 1, p. 235, 2002.

[52] ZHANG, H. et al. Effects of Angular Misalignment in Interferometric Detection of Distributed Polarization Coupling. **Measurement Science and Technology**, v. 20, n. 9, p. 095112, 2009.

[53] LERCH, P. et al. Assessing Noise Sources at Synchrotron Infrared Ports. **Journal of Synchrotron Radiation**, v. 19, n. 1, p. 1–9, 2011.

[54] SANTORO, G. et al. Infrared Synchrotron Radiation From Bending Magnet and Edge Radiation Sources for the Study of Orientation and Conformation in Anisotropic Materials. **Review of Scientific Instruments**, v. 82, n. 3, 2011.

[55] BLUM, M.; JOHN, H. Historical Perspective and Modern Applications of Attenuated Total Reflectance – Fourier Transform Infrared Spectroscopy (ATR-FTIR). **Drug Testing and Analysis**, v. 4, n. 3–4, p. 298–302, 2011.

[56] WANG, Z.; TAKAHASHI, H. Development of Mid-Infrared Absorption Spectroscopy for Gemstone Analysis. **Minerals**, v. 13, n. 5, p. 625, 2023.

[57] ABRAMOVICH, A.; SHULZINGER, A. Diagnostic and Analysis of Human Sperm Characteristics Using Fourier Transform Infrared Spectroscopy. **Open Journal of Urology**, v. 05, n. 06, p. 97–101, 2015.

[58] ZHANG, S. et al. Fiber Optic Probe-based ATR-FTIR Spectroscopy for Rapid Breast Cancer Detection: A Pilot Study. **Journal of Biophotonics**, v. 16, n. 11, 2023.

[59] ACHTENBERG, K.; MIKOŁAJCZYK, J.; BIELECKI, Z. Two-Channel Detecting Sensor with Signal Cross-Correlation for FTIR Instruments. **Sensors**, v. 22, n. 22, p. 8919, 18 nov. 2022.

[60] SHEBERSTOV, K. F. et al. SAN Plot: A Graphical Representation of the Signal, Noise, and Artifacts Content of Spectra. **Magnetic Resonance in Chemistry**, v. 58, n. 5, p. 466–472, 2019.

[61] CHRISTENSEN, J. B. et al. Intrinsic Spectral Resolution Limitations of QEPAS Sensors for Fast and Broad Wavelength Tuning. **Sensors**, v. 20, n. 17, p. 4725, 2020.

[62] HASHIMOTO, K.; IDEGUCHI, T. Phase-Controlled Fourier-Transform Spectroscopy. **Nature Communications**, v. 9, n. 1, 2018.

[63] BRANGULE, A.; ŠUKELE, R.; BANDERE, D. Herbal Medicine Characterization Perspectives Using Advanced FTIR Sample Techniques –

Diffuse Reflectance (DRIFT) and Photoacoustic Spectroscopy (PAS). **Frontiers in Plant Science**, v. 11, 2020.

[64] GUERRERO-PÉREZ, M. O.; PATIENCE, G. S. Experimental Methods in Chemical Engineering: Fourier Transform Infrared Spectroscopy—FTIR. **The Canadian Journal of Chemical Engineering**, v. 98, n. 1, p. 25–33, 2019.

[65] MARGENOT, A. J.; PARIKH, S. J.; CALDERÓN, F. J. Fourier-transform Infrared Spectroscopy for Soil Organic Matter Analysis. **Soil Science Society of America Journal**, v. 87, n. 6, p. 1503–1528, 2023.

[66] MERRILL, R. A.; BARTICK, E. G. Analysis of Pressure Sensitive Adhesive Tape: I. Evaluation of Infrared ATR Accessory Advances. **Journal of Forensic Sciences**, v. 45, n. 1, p. 93–98, 2000.

[67] MILEWSKA, A. et al. In-line Monitoring of Protein Concentration With MIR Spectroscopy During UFDF. **Engineering in Life Sciences**, v. 23, n. 2, 2022.

[68] CANEVAROLO JR, S. V. Técnicas de caracterização de polímeros. **Artliber, São Paulo**, v. 430, n. 2004, 2004.

[69] ENGEL, J. et al. Breaking with trends in pre-processing? **TrAC Trends in Analytical Chemistry**, v. 50, p. 96–106, 2013.

[70] NOONAN, K. Y. et al. Rapid classification of simulated street drug mixtures using Raman spectroscopy and principal component analysis. **Applied Spectroscopy**, v. 63, n. 7, p. 742–747, 2009.

[71] OCHIENG, P. J. et al. Adaptive Savitzky–Golay Filters for Analysis of Copy Number Variation Peaks from Whole-Exome Sequencing Data. **Information**, v. 14, n. 2, p. 128, 2023.

[72] OXBY, P. W. An Optimal Weighting Function for the Savitzky-Golay Filter. **arXiv preprint arXiv:2111.11667**, 2021.

[73] HAMRIN, T. H. et al. Performance of regional oxygen saturation monitoring by near-infrared spectroscopy (NIRS) in pediatric inter-hospital transports with special reference to air ambulance transports: a methodological study. **Journal of Clinical Monitoring and Computing**, v. 32, p. 841–847, 2018.

[74] QUARANTA, G.; CARBONI, B.; LACARBONARA, W. Damage detection by modal curvatures: numerical issues. **Journal of Vibration and Control**, v. 22, n. 7, p. 1913–1927, 2016.

[75] BARTON, S. J.; WARD, T. E.; HENNELLY, B. M. Algorithm for optimal denoising of Raman spectra. **Analytical methods**, v. 10, n. 30, p. 3759–3769, 2018.

[76] SCHULZE, H. G. et al. Smoothing Raman spectra with contiguous single-channel fitting of Voigt distributions: an automated, high-quality procedure. **Applied Spectroscopy**, v. 73, n. 1, p. 47–58, 2019.

[77] WANG, L. et al. Multisource uncertain dynamic load identification fitted by Legendre polynomial based on precise integration and the Savitzky–Golay filters. **International Journal for Numerical Methods in Engineering**, v. 123, n. 20, p. 4974–5006, 2022.

[78] SIM, K. et al. Signal-to-noise ratio estimation on SEM images using cubic spline interpolation with Savitzky–Golay smoothing. **Journal of microscopy**, v. 253, n. 1, p. 1–11, 2014.

[79] JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 2016.

[80] PUTRI, R. H. K. Multivariate Statistical Analysis of Coal in Dahor Formation, Borneo Island, Indonesia: A Comparative Study Utilizing Principal Component Analysis (PCA). **Journal of Earth and Marine Technology (Jemt)**, v. 3, n. 2, p. 88–97, 2023.

[81] SANTOS, P. R. B. et al. A Multivariate Approach to Analyze the Spatial-Temporal Variation of Limnological Parameters of the Reservoir of the Curuá-Una Hydroelectric Plant. **Revista Ibero-Americana De Ciências Ambientais**, v. 11, n. 6, p. 479–491, 2020.

[82] SINHA, A. et al. Principal Component Analysis Approach for Comprehensive Screening of Tomato Germplasm for Polyhouse Condition. **Journal of Experimental Agriculture International**, p. 67–72, 2021.

[83] WANG, Z. et al. Assessments of Surface Water Quality Through the Use of Multivariate Statistical Techniques: A Case Study for the Watershed of the Yuqiao Reservoir, China. **Frontiers in Environmental Science**, v. 11, 2023.

[84] YE, N.; PARMAR, D.; BORROR, C. M. A Hybrid SPC Method With the Chi-Square Distance Monitoring Procedure for Large-scale, Complex Process Data. **Quality and Reliability Engineering International**, v. 22, n. 4, p. 393–402, 2005.

[85] GUH, R. On-line Identification and Quantification of Mean Shifts in Bivariate Processes Using a Neural Network-based Approach. **Quality and Reliability Engineering International**, v. 23, n. 3, p. 367–385, 2006.

[86] BONNINI, S.; ASSEGIE, G. M. Advances on Permutation Multivariate Analysis of Variance for Big Data. **Statistics in Transition New Series**, v. 23, n. 2, p. 163–183, 2022.

[87] PARA, M. G. L.; RODRIGUEZ-LOAIZA, P. Application of the Multivariate T2Control Chart and the Mason–Tracy–Young Decomposition Procedure to the Study of the Consistency of Impurity Profiles of Drug Substances. **Quality Engineering**, v. 16, n. 1, p. 127–142, 2003.

[88] LUTS, J. et al. A tutorial on support vector machine-based methods for classification problems in chemometrics. **Analytica chimica acta**, v. 665, n. 2, p. 129–145, 2010.

[89] FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações**. [s.l.] Editora da UNICAMP, 2015.

[90] DEGADWALA, DR. S. Theoretical Evaluation of Machine Learning and Deep Learning Applications in Various Domain. **International Journal of Scientific Research in Computer Science Engineering and Information Technology**, v. 10, n. 3, p. 567–575, 2024.

[91] TABARES-SOTO, R. et al. A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data. **Peerj Computer Science**, v. 6, p. e270, 2020.

[92] DE LIMA, M. D.; BARBOSA, R. Methods of authentication of food grown in organic and conventional systems using chemometrics and data mining algorithms: A review. **Food Analytical Methods**, v. 12, p. 887–901, 2019.

[93] XU, Y.; GOODACRE, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. **Journal of Analysis and Testing**, v. 2, n. 3, p. 249–262, 2018.

[94] ROSENBERG, L. H. et al. Multivariate Meta-Analysis of Proteomics Data From Human Prostate and Colon Tumours. **BMC Bioinformatics**, v. 11, n. 1, 2010.

[95] KORJUS, K.; HEBART, M. N.; VICENTE, R. An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. **Plos One**, v. 11, n. 8, p. e0161788, 2016.

[96] ANIL, A. K. P. An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models. **Journal of Systems Engineering and Information Technology (Joseit)**, v. 2, n. 2, p. 77–84, 2023.

[97] LING, K. S. Modeling Tenant's Credit Scoring Using Logistic Regression. **Sage Open**, v. 13, n. 3, 2023.

[98] SEHRA, S.; FLORES, D.; MONTANEZ, G. D. Undecidability of Underfitting in Learning Algorithms. 2021.

[99] MAIONE, C. et al. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. **Computers and Electronics in Agriculture**, v. 121, p. 101–107, 2016.

[100] GU, B. Optimizing Large-Scale Hyperparameters via Automated Learning Algorithm. 2021.

[101] FEURER, M.; HUTTER, F. Hyperparameter Optimization. 2019.

[102] YU, Z.; WANG, Y.; WANG, Y. A Support Vector Machine and Particle Swarm Optimization Based Model for Cemented Tailings Backfill Materials Strength Prediction. **Materials**, 2022.

[103] VAN RIJN, J. N. et al. Fast Algorithm Selection Using Learning Curves. 2015.

[104] WENZEL, F. et al. Bayesian Nonlinear Support Vector Machines for Big Data. 2017.

[105] VERMA, M.; THIRUMALAISELVI, A.; RAJASANKAR, J. Kernel-Based Models for Prediction of Cement Compressive Strength. **Neural Computing and Applications**, 2016.

[106] WANG, B.; GONG, N. Z. Stealing Hyperparameters in Machine Learning. 2018.

[107] VAN RIJN, J. N.; HUTTER, F. Hyperparameter Importance Across Datasets. 2018.

[108] BAHMANI, M. To Tune or Not to Tune? An Approach for Recommending Important Hyperparameters. 2021.

[109] EICHELBERGER, R. K.; SHENG, V. S. Does One-Against-All or One-Against-One Improve the Performance of Multiclass Classifications? **Proceedings of the Aaai Conference on Artificial Intelligence**, v. 27, n. 1, p. 1609–1610, 2013.

[110] CHAO, C.; HORNG, M.-H. The Construction of Support Vector Machine Classifier Using the Firefly Algorithm. **Computational Intelligence and Neuroscience**, v. 2015, p. 1–8, 2015.

[111] ELSHEWEY, A. M. et al. Bayesian Optimization With Support Vector Machine Model for Parkinson Disease Classification. **Sensors**, 2023.

[112] DURLIK-POPIŃSKA, K. et al. Correlations between autoantibodies and the atr-ftir spectra of sera from rheumatoid arthritis patients. **Scientific Reports**, v. 11, n. 1, p. 17886, 2021.

[113] HUANG, J. et al. Vibrational spectroscopic investigation of blood plasma and serum by drop coating deposition for clinical application. **International journal of molecular sciences**, v. 22, n. 4, p. 2191, 2021.

[114] PLATA-NAZAR, K. et al. Evaluation of clinical usefulness of serum neopterin determination in children with bacterial infections. **Acta Biochimica Polonica**, v. 62, n. 1, p. 133–137, 2015.

[115] AWAN, M. et al. Serum lipid variation in patients with dengue virus infection and associated risks of cardio vascular disorder. **Albus Scientia**, v. 2022, n. 2, p. 1–4, 2022.

[116] KUMAR, P. et al. Occurrence of malaria positive cases and their association with serum creatinine and blood urea in different age group. 2021.

[117] AL-SALAHY, M. et al. Parasitaemia and its relation to hematological parameters and liver function among patients malaria in Abs, Hajjah, Northwest Yemen. **Interdisciplinary perspectives on infectious diseases**, v. 2016, n. 1, p. 5954394, 2016.

[118] CHECHET, G.; AMINU, R. Trypanosoma congolense: Prophylactic Potentials Of Antiserum And Adjuvant In Experimental Mice. **Nigerian Veterinary Journal**, v. 40, n. 1, p. 19–34, 2019.

[119] ANAS, M. et al. Is procalcitonin a reliable indicator of sepsis in spinal cord injury patients: an observational cohort study. **European Spine Journal**, v. 32, n. 5, p. 1591–1597, 2023.

[120] YU, Y.; LI, H. Diagnostic and prognostic value of procalcitonin for early intracranial infection after craniotomy. **Brazilian Journal of Medical and Biological Research**, v. 50, n. 5, p. e6021, 2017.

[121] SINGH, D. et al. Understanding the Pathogenesis of Endothelial Injury Induced by Bluetongue Virus in Experimentally Infected Sheep. **Journal of Pure & Applied Microbiology**, v. 18, n. 3, 2024.

[122] KHAN, S. et al. Analysis of tuberculosis disease through Raman spectroscopy and machine learning. **Photodiagnosis and Photodynamic Therapy**, v. 24, p. 286–291, dez. 2018.

[123] BAKER, M. J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. **Nature protocols**, v. 9, n. 8, p. 1771–1791, 2014.

[124] BEHDAD, S.; MASSUDI, R.; PAKDEL, A. Non-destructive diagnosis of Inflammatory Bowel Disease by near-infrared spectroscopy and aquaphotomics. **Scientific Reports**, v. 14, n. 1, p. 15895, 2024.

[125] SANDT, C. Identification and classification of proteins by FTIR microspectroscopy. A proof of concept. **Biochimica et Biophysica Acta (BBA)-General Subjects**, v. 1868, n. 10, p. 130688, 2024.

[126] LI, H. et al. Comparison of serum from lung cancer patients and from patients with benign lung nodule using FTIR spectroscopy. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 306, p. 123596, fev. 2024.

[127] TOKGOZ, G. et al. Spectrochemical and explainable artificial intelligence approaches for molecular level identification of the status of critically ill patients with COVID-19. **Talanta**, v. 279, p. 126652, nov. 2024.

ANEXOS

A1: Manuscrito submetido

Improving Bovine Brucellosis Diagnostics: Fast, Accurate Detection via blood serum FTIR Spectroscopy and ML Techniques

Journal:	*ACS Omega*
Manuscript ID	ao-2025-00504g
Manuscript Type:	Article
Date Submitted by the Author:	17-Jan-2025
Complete List of Authors:	Franca, Thiago; UFMS, Physics Institute Lacerda, Miller; UFMS, Physics Institute Calvani, Camila; UFMS, Physics Institute Arruda, Kelvyn; UFMS Maranni, Ana; UFMS Nicolodelli, Gustavo; UFSC, PHysics KARTHIKEYAN, SIVAKUMARAN; Dr Ambedkar Goverment Arts College, Department of Physics Marangoni, Bruno; UFMS, Physics Institute Nascimento, Carlos; UFMS Cena, Cicero; UFMS, Physics Institute

Improving Bovine Brucellosis Diagnostics: Fast, Accurate Detection via blood serum FTIR Spectroscopy and ML Techniques

Thiago Franca[1], Miller Lacerda[1], Camila Calvani[1], Kelvyn Arruda[1], Ana Maranni[1], Gustavo Nicolodelli[3], Sivakumaran Karthikeyan[2], Bruno Marangoni[1], Carlos Ramos[4]*, Cicero Cena[1]*

*carlos.nascimento@ufms.br; *cicero.cena@ufms.br

[1]UFMS – Universidade Federal de Mato Grosso do Sul, Optics and Photonic Lab (SISFOTON-UFMS), Campo Grande-MS, Brazil.

[2]Department of Physics , Dr. Ambedkar Government Arts College, Chennai 600039 ,Tamilnadu, India.

[3]UFSC – Universidade Federal de Santa Catarina, Florinópolis-SC, Brazil.

[4]UFMS – Universidade Federal de Mato Grosso do Sul, Campo Grande-MS, Brazil.

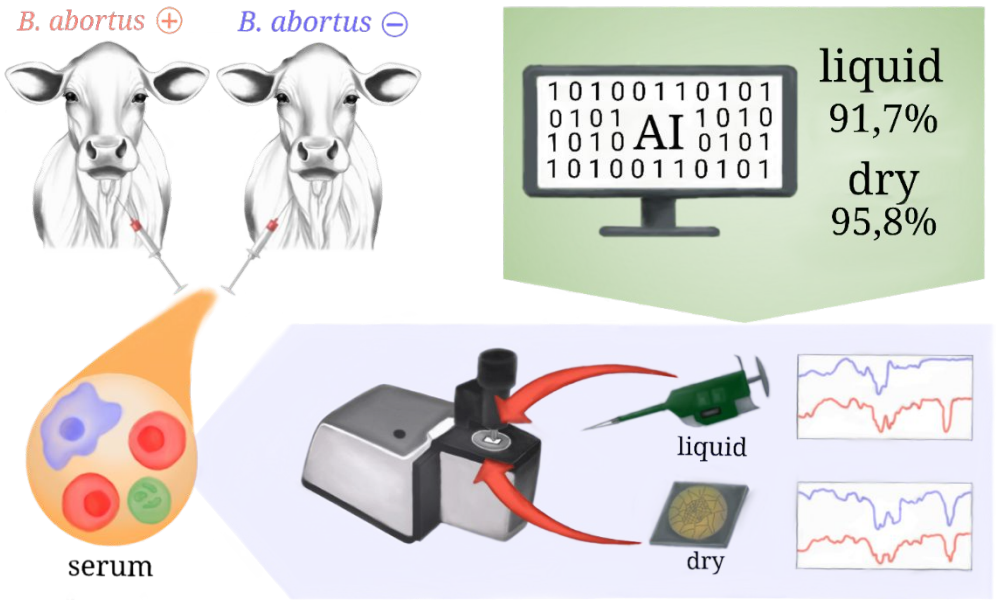
ABSTRACT

Diagnosing bovine brucellosis is a major challenge due to its significant economic impact, causing losses in meat and dairy production, and its potential to transmit to humans. In Brazil, disease control relies on diagnosis, animal culling, and vaccination. However, existing diagnostic tests, despite their quality, are time-consuming and prone to false positives and negatives, complicating effective control. There is a critical need for a low-cost, fast, and accurate diagnostic test for large-scale use. Spectroscopy techniques combined with machine learning show great promise for improving diagnostic tests. Here we explore the potential use of FTIR spectroscopy and machine learning algorithms to provide a rapid, accurate, and cost-effective diagnostic method for Brucella abortus. This study explored the use of FTIR spectroscopy on bovine blood

serum in liquid and dried form to develop a new photodiagnosis method. Initially, the FTIR data was pre-treated using the Standard Normal Deviate method to remove baseline deviations. Principal component analysis (PCA) was then applied to observe clustering tendencies, and further selection of principal components improved clustering. Using Support Vector Machine (SVM) algorithms, the predictive models achieved an overall accuracy of 95.8%. This new methodology delivers results in about 5 minutes, compared to the 48 hours required for standard diagnostic methods. These findings demonstrate the viability of this approach for diagnosing bovine brucellosis, potentially enhancing disease control programs in Brazil and beyond.

Keywords: Photodiagnosis; Diagnostic test; FTIR spectroscopy; Machine learning; Bovine brucellosis.

Graphical Abstract



INTRODUCTION

Bovine brucellosis is a highly infectious zoonotic disease caused primarily by the bacterium *Brucella abortus*, which affects cattle and can also impact other livestock and humans. The disease causes reproductive failures in cattle, such as abortions, stillbirths, and infertility, leading to significant economic losses in the livestock industry [1]. The disease management and control strategies depend on the quick and accurate identification of infected animals [1,2]. Despite advances in bovine brucellosis diagnostic methods, we are still seeking for an easily implementation diagnostic method, portable for field applications that may offer faster and more reliable results [2,3].

Brucellosis triggers a complex immune response in cattle, characterized by a robust activation of both innate and adaptive immunity. Briefly, the innate immune response is the body's first line of defense against infections and injuries. It provides a rapid, non-specific response using physical and chemical barriers, immune cells (such as macrophages, neutrophils, and natural killer cells), and proteins. This response also involves inflammation to isolate and combat pathogens and uses pattern recognition receptors to detect common features of pathogens. It is essential for immediate defense and for activating the more specific adaptive immune response. The adaptive immune response is a specific defense mechanism that develops after exposure to a pathogen. It involves the activation of lymphocytes (B cells and T cells) that recognize specific antigens. B cells produce antibodies that neutralize pathogens, while T cells destroy infected cells or help other immune cells [4-6].

As a result, this immune response causes increased levels of pro-inflammatory cytokines (e.g., TNF-α, IL-1, and IL-6) [7]. The serum of infected cattle

typically shows elevated levels of specific antibodies against *Brucella* antigens. The detection of these antibodies is crucial for diagnosing the disease, although cross-reactivity with other infections can complicate interpretations [2]. The number of acute phase proteins (APPs) such as haptoglobin and serum amyloid A gets higher. These proteins serve as biomarkers for inflammation and can help assess the severity of the infection [8]. The immune response may also cause alterations in serum electrolytes and total protein levels, reflecting the systemic effects of the infection and the body's response to it [9].

This scenario is fruitful for using molecular optical spectroscopy to investigate biofluid properties, whose characteristics can lead to a data analysis protocol to find a new photodiagnostic method. Molecular optical spectroscopy, such as Fourier Transformed Infrared (FTIR) and Raman spectroscopy are very sensitive to small changes in the sample composition and have been using to the development of many photodiagnosis methods [10-13]. The main advantage of FTIR or Raman lies in the robustness of the technique, fast response (few seconds), easy implementation, without the need for extensive preparation, making them less labor-intensive, and low cost. At the same time, the importance of developing new methods for diagnosing brucellosis lies in the disease's impact on public health, agriculture, and economic stability.

Serum samples from Brucella-infected sheep were investigated by Fourier Transform Infrared (FTIR) spectroscopy combined with Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA). According to the study, the spectral range of 3700–3090 cm^{-1} and 3000–2800 cm^{-1} ranges - usually assigned to lipids molecules - achieved 100% sensitivity, specificity, and accuracy. The FTIR-PCA-LDA method outperformed traditional diagnostic methods, such as indirect ELISA and serum

fluorescence polarization assay, in terms of sensitivity, specificity, and accuracy. These findings suggest that the FTIR-PCA-LDA method provides high diagnostic accuracy for distinguishing Brucella-infected sheep from healthy sheep. The classification models were validated using independent sample sets. The performance of the models were described by using Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) parameters [14].

Bovine brucellosis was successfully diagnosed by using UV-vis spectroscopy combined with machine learning (ML) algorithms. Due to the nature of the UV-vis spectroscopy to enhance the detection of sample differentiation antigen for *B. abortus* infection was used diluted in saline containing phenol. Then 2 µL of serum-antigen mixture was characterized in the 200-300 nm range, after 3 minutes of incubation at room temperature. After data pre-processing, the results obtained from principal component analysis (PCA) didn´t suggest any tendency of group separation; to improve such characteristics the authors applied a Feature Selection Recursive Feature Elimination (RFE) algorithm to select the principal components (PCs) that most contributed for group separation. After the selection of three PCs (PC3, PC4, and PC14) that contributed most to group classification, the best prediction model was built using the k-nearest Neighbor (KNN) algorithm with the cosine function, achieving an overall accuracy of 95.9% in the Leave-One-Out Cross-Validation (LOOCV) test. The analytical sensitivity was 92%, and the specificity was 100%. In external validation, the model showed an accuracy of 84.4%, sensitivity of 75%, and specificity of 94%. The proposed UV-vis/ML method effectively distinguished between positive and negative samples for antibodies against B. abortus in cattle, reducing the diagnostic time to about 5 minutes,

compared to the 48 hours required by traditional methods like BAAT, 2-ME, and SAT [15].

Despite the great potential demonstrated by optical spectroscopy combined with machine learning algorithms for disease diagnosis, the method also faces limitations. If the disease does not cause significant alterations in the sample to be analyzed, usual strategies involving PCA analysis and regular machine learning algorithms, such as KNN, SVM (support vector machine), and LDA (linear discriminant analysis), may not achieve high accuracy. For instance, in an animal model study to diagnose cutaneous leishmaniasis in female BALB/c mice, blood serum analysis by FTIR spectroscopy followed by PCA analysis and the SVM algorithm only achieved 72% accuracy [12].

Research has indicated that optical spectroscopy can be used to identify specific biomarkers related to brucellosis, such as changes in serum proteins or metabolic profiles. This could lead to the development of point-of-care diagnostic tools that are both effective and easy to use. Combining optical spectroscopy with existing molecular methods, such as PCR, could enhance diagnostic accuracy and provide a comprehensive approach to brucellosis detection [16]. In summary, advancing diagnostic methods for brucellosis is crucial for effective disease management and public health safety. Optical spectroscopy offers a novel approach that could improve the speed, sensitivity, and specificity of brucellosis diagnosis, addressing current limitations in traditional methods. In this study, FTIR data was analyzed using machine learning algorithms to develop predictive models. By applying these techniques, we achieved over 90% accuracy in sample classification, significantly reducing the subjectivity inherent in current diagnostic tests, which rely on human interpretation.

2. MATERIALS AND METHODS

2.1. Sample description, preparation, and data collection.

Eighty bovine blood serum samples were provided by the Federal Agricultural Defense Laboratory-MG, which receives samples from all over Brazil for diagnosis under the National Program for the Control and Eradication of Brucellosis and Animal Tuberculosis (PNCEBT) [17]. Therefore, the samples were not originally collected by LFDA for research purposes. The blood serum originated from cows of various breeds, sexes, weights, and ages. The samples were divided into forty serum samples from infected animals and forty from a control group. The infected group was determined as positive for antibodies against *B. abortus*, according to reference tests recommended by Brazilian legislation (BAAT, 2-ME and SAT) [16].

The samples were analyzed in both liquid and dried forms in the range of 1800−900 cm^{-1} range, with a resolution of 4 cm^{-1}, and 12 scans, by using a Fourier Transformed Infrared Spectrometer (FTIR), Agilent – Cary 630 model – with an attenuated total reflectance (ATR) accessory.

Dried samples were prepared by casting blood serum onto a silicon (SiO2) substrate. A thick sample was obtained after three steps of 20 µl deposition followed by a drying process at 40°C per 10 minutes. The average spectra of the dried samples were collected in three different spots (at the center and two edges of the drop) to mitigate compositional heterogeneity [18], atmospheric background was obtained before each acquisition. Liquid samples were left to thaw at room temperature, and then 20 µl was directly deposited on the Attenuated Total Reflectance accessory (ATR) at

the FTIR spectrometer, deionized water was used to obtain the background before each spectra acquisition.

2.2. Data pre-treatment and prediction model development.

The entire data analysis was performed using the Scikit-Learn library (version 1.3.0) in the Python programming language (version 3.11.5) [19]. First, the FTIR raw data were pre-processed using Standard Normal Variate (SNV). SNV is a well-known normalization process that enhances data quality by reducing experimental noise and systematic variations, allowing for more accurate and consistent spectral analysis. Practically, SNV normalizes spectroscopic data by adjusting each spectrum so that its mean is 0 and its standard deviation is 1, facilitating the comparison between different samples and the extraction of relevant features for analytical models [20]. Due to noise in the spectra of liquid samples, we also applied an IFFT (Inverse Fast Fourier Transform) filter to smooth the curve [21].

Second, the FTIR-SNV spectra of each group were divided into 60% for training and 40% for testing and submitted to a Principal Component Analysis (PCA) [22]. Here, we apply Hotelling's T^2 to remove outliers by measuring the distance of a sample from the multivariate mean, considering the covariance of the variables across the entire sample set [23]. The PCA is an important analysis to reduce the dimensionality of large datasets - such as spectral data - while preserving as much variability as possible. This is achieved by transforming the original data into a new set of uncorrelated variables called principal components. As a result, we obtain information regarding the data variance; (i) the score plot, provides information about the distribution and clustering of samples, revealing potential separation and grouping patterns of the new

data at the principal component space. The separation or grouping depends on the data variance related to each PC; (ii) The contribution of each PC and its meaning can be revealed by analyzing the loading plot, which provides insights into the contribution of each original variable to the principal components' variance, helping to identify which variables are most influential in the data's variance and contribute for group separation when it occurs.

Once the potential separation of the groups is explored and revealed by PCA analysis we can build a prediction model to automatize the sample classification and perform validation tests for the method. In the next step, PCA data (60% samples for each group) are used to train Support Vector Machine (SVM) algorithms, which explore diverse functions (Linear, Quadratic, and Cubic) to construct hyperplanes responsible for separating the groups and then classify new samples according to their classes. The algorithm can optimize the hyperplane characteristics to enhance the accuracy, the hyperparameter C, which manages the trade-off between maximizing the margin and minimizing classification errors, was adjusted to balance fitting the training data well while preventing overfitting. Fine-tuning of this hyperparameter was crucial, as their optimal values depend on the dataset's characteristics. The model was regularized using linear and polynomial kernels (degrees 2 and 3) and hyperparameters (C = 1, 10, or 100) [24].

The performance of the SVM was evaluated using Leave-One-Out Cross-Validation (LOOCV) [25]. In this method, one sample is withdrawn from the dataset, and the prediction model is built using the remaining data. The model is then tested on the withdrawn sample. This process is repeated until each sample has been removed and tested. The final accuracy is calculated as the average of the accuracy values obtained

from each iteration. Finally, a validation test is performed by using the best SVM, function, and hyperparameters found in the LOOCV test by using the remaining 40% of the samples. The results are summarized in a confusion matrix, showing how the samples are classified according to their original label and the overall accuracy of the model.

3. RESULTS AND DISCUSSION

Figure 1 shows the FTIR-SNV spectra for bovine blood serum samples. The blue-colored trace represents the control group (non-infected by *B. abortus*), and the red-colored trace represents the *B. abortus* infected group (from here called Brucellosis group). The dried sample (upper spectra set) and the liquid sample (lower spectra set) exhibit significant differences, particularly in the 1650 to 1500 cm^{-1} range. These differences are primarily due to strong bands associated with C=C and C=O stretching, as well as N-O stretching and N-H bending vibrational modes from amide I and amide II, respectively [11,12, 26].

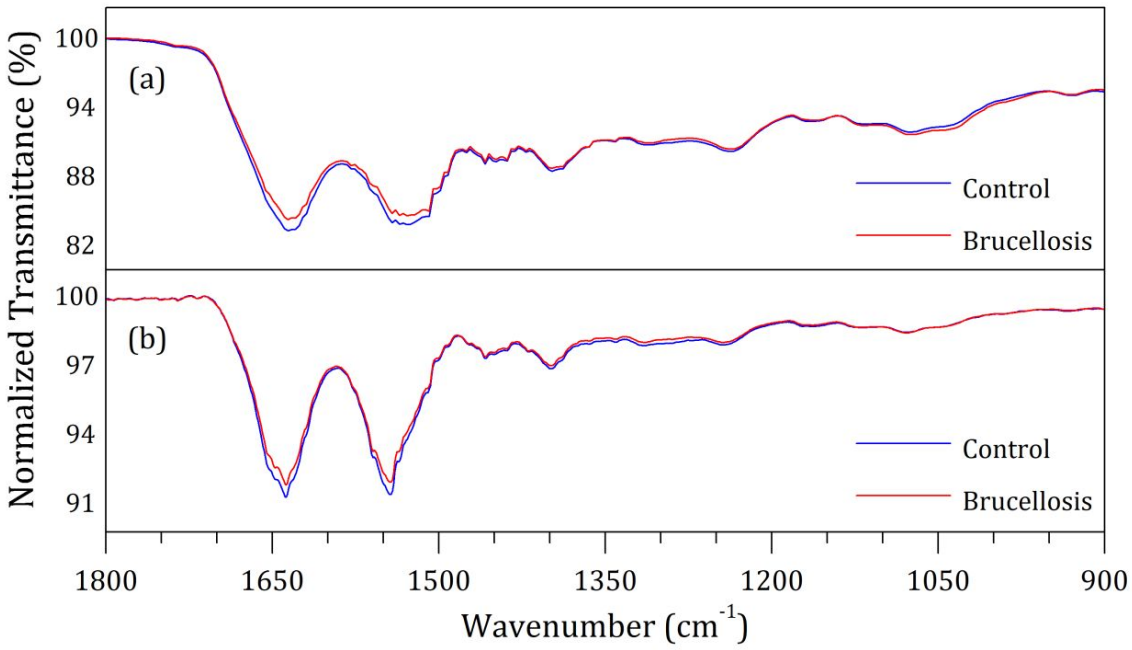


Figure 1: Superimposed FTIR-SNV spectra for bovine blood serum control group (blue-colored trace), and the *B. abortus* infected group (red-colored trace) – called Brucellosis. The dried sample (upper spectra set) and the liquid sample (lower spectra set). For each kind of sample, the Control and Brucellosis group spectra were superimposed to direct comparison.

Both sample groups exhibit the same vibrational bands with closely matching positions and relative intensities. The spectra of the dried samples are approximately six times more intense than those of the liquid samples, primarily due to concentration differences. Notably, the liquid samples display a distinctive pattern: two strong bands in the 1650 to 1500 cm^{-1} range, while the weaker bands in the 1200 to 1000 cm^{-1} range are almost imperceptible.

Since dried samples yield a more intense signal, we anticipated that this data set could offer more precise characteristics for sample characterization and differentiation between groups. In contrast, the low-intensity spectra of the liquid samples may complicate the differentiation process, as subtle variations in the spectra

could be attributed to inherent data quality issues rather than actual compositional differences.

More importantly, no significant difference between the spectra groups for each kind of sample can't be observed. The same bands with similar characteristics are present for both groups. The main prominent band identified for both groups was assigned for Amide I and Amide II from proteins, around 1638 and 1543 cm^{-1}, respectively; followed by minor vibrational bands at 1458 cm^{-1} (CH_2 symmetric bending) from lipids; 1398 cm^{-1} (COO^- symmetric stretching) and 1317 cm^{-1} (Amide III) from proteins; 1243 cm^{-1} (PO_2^- asymmetric stretching) from phosphate band; 1168 cm^{-1} (C-OH asymmetric stretching) from proteins; 1081 cm^{-1} (PO_2^- symmetric stretching) from phosphate band; and 930 cm^{-1} (C-C stretching) from carbohydrates [11,12, 27].

The FTIR-SNV spectra for dried samples and the FTIR-SNV-IFFT spectra for liquid samples were analyzed by principal component analysis (PCA). Figure 2 summarizes the main results found for dried, Figure 2(a-b), and liquid, Figure 2(c-d), samples. The normalized score plot for FTIR-SNV spectra of bovine dried blood serum, Figure 2(a), demonstrates a clear tendency for separation using only two principal components (PCs), which account for 92.3% of the data variance. Both the control group (blue circles) and the Brucellosis group (red circles) exhibit high dispersion along both PC1 and PC2, reflecting the significant intra-group variance due to the inherent variability of the animals studied. Despite this, the variance between the control and Brucellosis groups reveals a distinct separation trend along the opposite diagonal. This separation trend indicates a promising condition for building a predictive model using SVM algorithms.

By analyzing the loading plot in Figure 2(b), we can correlate the data variance with the original data, supporting the chemical significance of the analysis. The PC1 loading shows a significant contribution from the amide I and II bands between 1650 and 1500 cm-1, followed by contributions around 1400, 1200, and 1050 cm-1. Additionally, the PC2 loading also highlights a major contribution from the amide I and II bands, with further contributions between 1500 and 1300 cm-1 and a notable contribution in the 1150 to 950 cm-1 range.

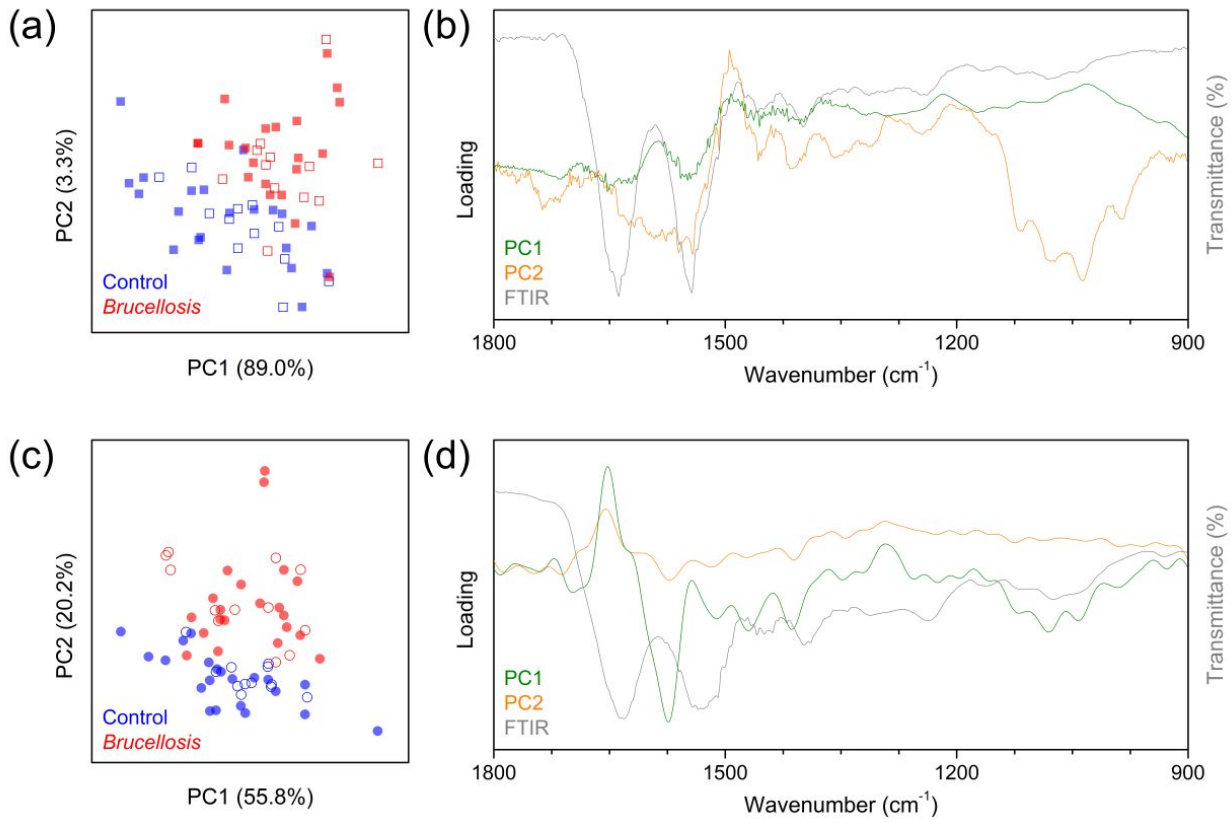


Figure 2: Principal Component Analysis results from (a-b) FTIR-SNV dried sample spectra, and (c-d) FTIR-SNV-IFFT liquid sample spectra. Brucella abortus infected group (red circles) and Control (non-infected) group (blue circles). Score plot on the left side and loading plot on the right side.

The normalized score plot for FTIR-SNV-IFFT spectra of bovine liquid blood serum, Figure 2(c), also demonstrates a clear tendency for separation using only two principal components (PCs), which account for 76% of the data variance. The expected high dispersion along both PC1 and PC2 for both the control group (blue circles) and the Brucellosis group (red circles) was also observed. Despite this, the variance between the control and Brucellosis groups reveals a distinct separation trend along the diagonal. In the same way, this separation trend suggests a promising possibility for building a predictive model using SVM algorithms.

By analyzing the loading plot in Figure 2(d), we can observe that the same data variance present in the dried sample is evident, although the intensity has changed considerably. Significant differences are noted in the 1500 to 1700 cm-1 range, where prominent protein bands, including amide I and amide II, are notable. All contributions to data variance are strongly correlated with the FTIR spectra and the changes in the molecular composition of the blood serum. These changes are due to the immune response involving new molecules associated with macrophages, neutrophils, lymphocytes, antibodies, antigens, and cytokine cells [4-9].

The separation frontier observed in Figures 2(a) and (c) cannot be approximated by a purely linear function, as evidenced by the intermixing of red and blue points within each other's regions. This suggests the need to evaluate quadratic and cubic functions, in addition to adjusting hyperparameters. Adjusting the hyperparameter C in an SVM controls the trade-off between margin width and classification error, thereby influencing the complexity and smoothness of the decision hyperplane. However, when the score plot shows this clear tendency of group

separation very promising results are expected for machine learning algorithms build successful prediction models [26, 27].

Figure 3 shows the accuracy values obtained in the LOOCV and validation tests using the SVM algorithm with different functions and hyperparameter (C) values. No combination of SVM function, hyperparameter, or sample nature achieved 100% accuracy. For all tests, we limited the number of principal components (PCs) to two, consistent with the score plot shown in Figure 2. Overall, both dried and liquid samples provided a method for brucellosis diagnosis with accuracy above 90%.

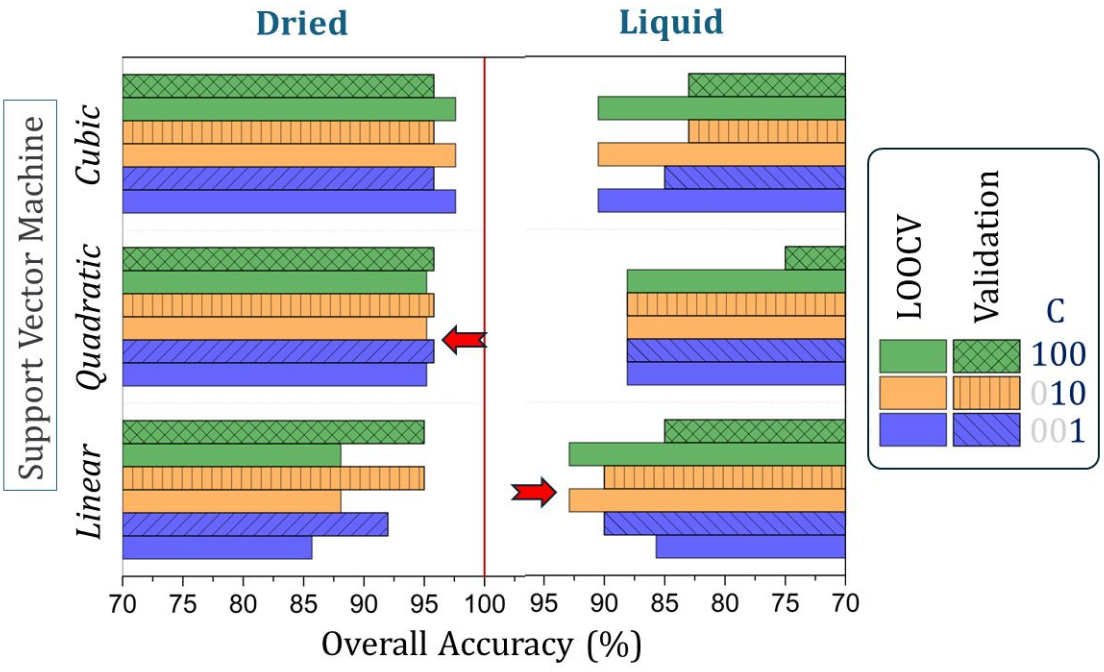


Figure 3: Overall accuracy obtained in the LOOCV (full color) and validation (line pattern) tests for Support vector machine (SVM) algorithms. Only the first two PCs were explored according to the function (Linear, Quadratic, and Cubic) and the hyperparameter C (1, 10, and 100) in the 1800 to 900 cm^{-1} range for dried and liquid blood serum samples. The highest overall accuracy was around 97% for Cubic SVM, with any C parameter function for dried samples.

We have an unbalanced group of samples (60/40), which may result in slight deviations in overall accuracy between LOOCV and validation tests. However, our goal is to develop a more stable prediction model, which can be achieved by using a low number of PCs with high chemical significance. This approach offers simpler mathematical interpretation and, importantly, similar overall accuracy in both LOOCV and validation tests – here, we consider the best prediction models to be those that exhibit the smallest accuracy difference between Leave-One-Out Cross-Validation (LOOCV) and the validation test. Additionally, the accuracy from the validation test must not exceed that obtained from LOOCV. For such conditions, the best prediction model for dried samples was found by using Quadratic SVM with any hyperparameter C, which exhibits an overall accuracy of around 95% for LOOCV and validation tests, Figure 4(a). Linear SVM with hyperparameter C=100 for liquid samples was shown to be the best choice, with an overall accuracy of around 92% for LOOCV and validation test, Figure 4(b).

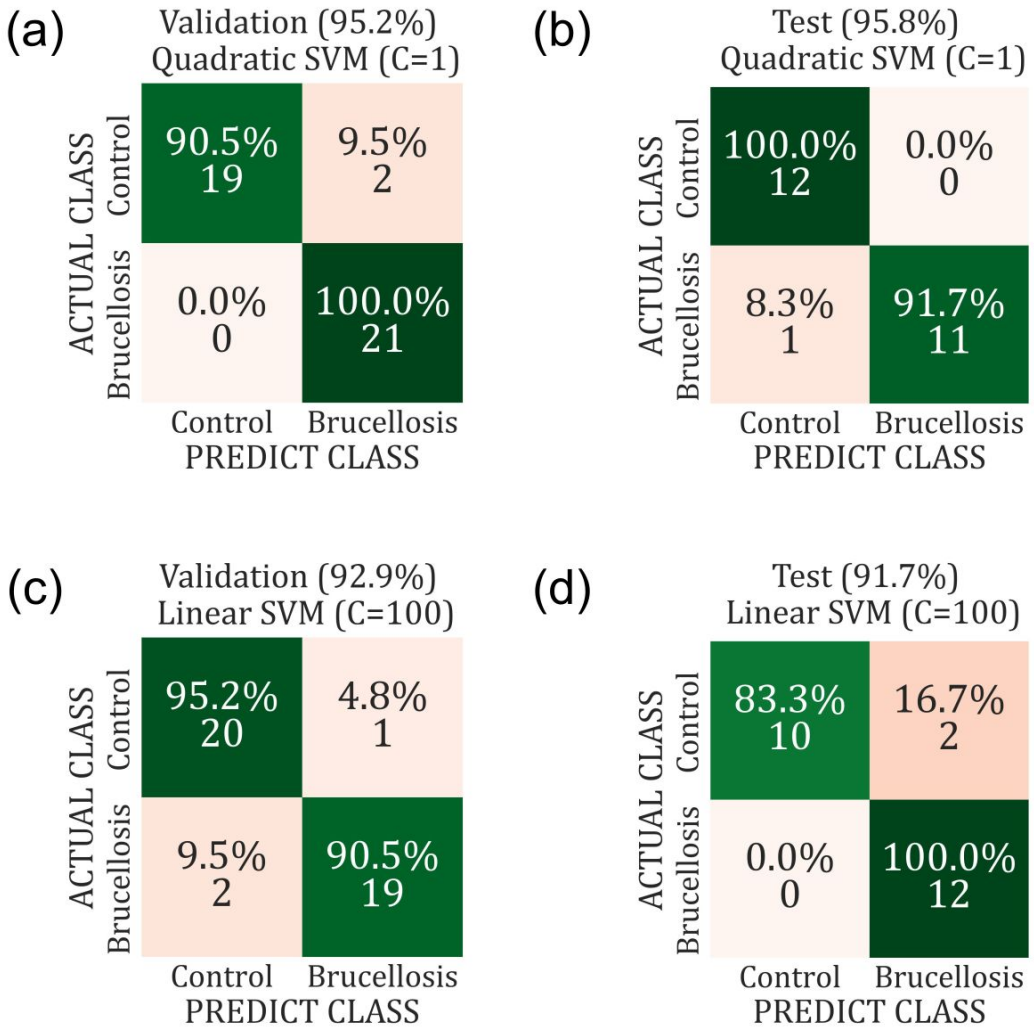


Figure 4: Confusion matrix for the performance of SVM algorithm using 2-PCs for dried and liquid blood serum samples. Classification results were obtained in the LOOCV test (right side) with a 60% data set and the External validation (left side) test with a 40% data set. PCA data from FTIR blood serum spectra in the 1800 to 900 cm^{-1} range.

The PCA analysis results clearly suggest a significant difference between the two groups analyzed. This leads to an easy interpretation of the differences caused by the immune response to B. abortus infection. Additionally, it facilitates the construction of a predictive model using machine learning that is simpler and more

intuitive, such as models based on the SVM algorithm, without requiring a more robust data analysis for this task. The correlation with the original data observed from the PCA analysis further reinforces the results obtained in the supervised validation tests presented in the right column of Figure 4. The control group used in this study consists of animals tested for brucellosis exclusion but monitored for overall animal health, not necessarily free from other diseases. The robustness of the proposed method now requires analysis with other disease groups that have similar immune response mechanisms to verify if this data analysis routine can be safely implemented as a routine test.

The current diagnostic tests take up to a week, leading to empiric antibiotic prescriptions. To avoid future problems rapid genomic tests are being developed to identify pathogens and resistance genes within a day to guide more targeted antibiotic use and improve antimicrobial stewardship [28]. Here we present a potential contribution to implementing photodiagnosis tests to help in this important task for disease control and surveillance.

Bovine brucellosis is an inherently chronic disease, characterized by a slow progression of clinical signs following infection, which may take months or even years to manifest. In some instances, such as in infected male cattle, clinical signs might not appear at all. The primary challenge in diagnosing bovine brucellosis lies in identifying asymptomatic animals to facilitate their removal from herds, thereby reducing disease transmission and enabling herd sanitation over time, ultimately achieving a "disease-free" status [29].

In the present study, samples were collected from asymptomatic animals. According to current Brazilian legislation, females aged 24 months and older, vaccinated

up to 8 months of age, must undergo diagnostic testing for bovine brucellosis, regardless of symptomatology [17]. Consequently, routine herd management in Brazil mandates that all females meeting this age requirement be subjected to serological tests for the disease. Animals identified as reactive must then be culled or subjected to sanitary slaughter. Thus, this study aimed to evaluate the proposed diagnostic method under the same conditions as those employed in routine bovine brucellosis diagnosis.

Moreover, as in any infectious disease, the stage of infection is directly correlated with the serum antibody titer. In this study, all samples underwent traditional serological tests (BAAT, 2-ME, and SAT), which are established and recommended for bovine brucellosis diagnosis [17]. In samples classified as positive, antibodies were consistently detected using conventional methods, indicating an adequate humoral immune response in infected animals.

For samples classified as negative, it is plausible that some originated from animals in the very early stages of infection (less than approximately 15 days), during which a detectable humoral immune response has not yet developed. In such cases, further studies involving experimental infections and longitudinal animal monitoring could provide insights into the proposed diagnostic method's capacity for early detection. However, given the pathogen's nature and associated biological risks, such studies can only be conducted in biosafety level 3 (BSL-3) [30], facilities equipped for large animal experimentation unavailable in Brazil. Alternatively, experiments could be conducted using mice in appropriately equipped facilities, which are more widely available. Nonetheless, the humoral immune response in mice does not necessarily reflect that of bovines, and any conclusions drawn under such conditions would be speculative at best.

4- CONCLUSION

The use of FTIR spectroscopy combined with machine learning enables the differentiation between serum samples (oven-dried) from control animals and those infected with bovine brucellosis, achieving 95.8% accuracy, 91.7% sensitivity, and 100% specificity. For the analysis of liquid blood serum, the results were 91.7% accuracy, 100% sensitivity, and 83.3% specificity. Both dried and liquid serum results surpassed those of gold standard tests. This approach shows potential for screening animals without the need for oven drying to remove water from the samples. Consequently, it allows for quick and efficient in situ health monitoring of animals, facilitating better control of bovine brucellosis spread in both animals and humans.

Ethical Approval

The present study does not require approval of the Animal Use Ethics Committee (CEUA), once the samples were obtained from a repository of the Federal Agricultural Defense Laboratory-MG, Brazil.

Consent to Publish

All participants were informed and consent for publication was obtained.

Author Contribution Statement

All authors have equal contributions for the study

Funding

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), code 001. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), code 403651/2020-5; 302525/2022-0; 440214/2021-1. Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT), code 91/2023; 360/2022.

Competing interest statement

The authors declare that there is no conflict of interest.

Data availability statement

Data will be shared upon request.

References

[1] Ashenafi Kiros, Hagos Asgedom, Reta Duguma Abdi. A Review on Bovine Brucellosis: Epidemiology, Diagnosis and Control Options. ARC Journal of Animal and Veterinary Sciences. Vol 2, Issue 3, p. 8-21, 2016.

[2] Ducrotoy MJ, Muñoz PM, Conde-Álvarez R, Blasco JM, Moriyón I. A systematic review of current immunological tests for the diagnosis of cattle brucellosis. Prev Vet Med. Vol 151, p.57-72, 2018.

[3] Rafaella Silva Andrade, Marina Martins de Oliveira, Júlio Sílvio de Sousa Bueno Filho, Fernando Ferreira, Jacques Godfroid, Andrey Pereira Lage, Elaine Maria Seles Dorneles. Accuracy of serological tests for bovine brucellosis: A systematic review and meta-analysis. Preventive Veterinary Medicine, Vol. 222, 106079, 2024.

[4] Gomes MTR, Campos PC, de Almeida LA, Oliveira FS, Costa MMS, Marim FM, Pereira GSM and Oliveira SC (2012) The role of innate immune signals in immunity to Brucella abortus. Front. Cell. Inf. Microbio. 2:130

[5] López-Santiago R, Sánchez-Argáez AB, De Alba-Núñez LG, Baltierra-Uribe SL and Moreno-Lafont MC (2019) Immune Response to Mucosal Brucella Infection. Front. Immunol. 10:1759

[6] Jinke Yang, Yue Wang, Yuanpan Hou, Mengyao Sun, Tian Xia, Xin Wu. Evasion of host defense by Brucella. Cell Insight, Vol. 3, Issue 1, 100143, 2024.

[7] Alcina V. Carvalho Neta, Juliana P.S. Mol, Mariana N. Xavier, Tatiane A. Paixão, Andrey P. Lage, Renato L. Santo. Pathogenesis of bovine brucellosis. The Veterinary Journal, vol. 184, p. 146–155, 2010.

[8] Dorneles EM, Teixeira-Carvalho A, Araújo MS, Sriranganathan N, Lage AP. Immune response triggered by Brucella abortus following infection or vaccination. Vaccine, vol. 33(31), p.3659-66, 2015.

[9] Pereira PCM. Interaction between infection, nutrition and immunity in tropical medicine. J Venom Anim Toxins incl Trop Dis. Vol 9(2), p.163–73, 2003.

[10] BAKER, Matthew J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature protocols, v. 9, n. 8, p. 1771-1791, 2014.

[11] Eliana C.A. de Brito, Thiago Franca Thalita Canassa, Simone S. Weber Anamaria M.M. Paniago, Cicero Cena. Paracoccidioidomycosis screening diagnosis by FTIR spectroscopy and multivariate analysis. Photodiagnosis and Photodynamic Therapy, Vol. 39, 102921, 2022.

[12] G Pacher, T Franca, M Lacerda, NO Alves, EM Piranda, C Arruda, C Cena. Diagnosis of Cutaneous Leishmaniasis Using FTIR Spectroscopy and Machine Learning: An Animal Model Study. ACS Infectious Diseases. Vol 10, issue 2, 2024.

[13] Stefano Fornasaro, Fatima Alsamad, Monica Baia, Luís A. E. Batista de Carvalho, Claudia Beleites, Hugh J. Byrne, Alessandro Chiadò, Mihaela Chis, Malama Chisanga, Amuthachelvi Daniel, Jakub Dybas, Gauthier Eppe, Guillaume Falgayrac, Karen Faulds, Hrvoje Gebavi, Fabrizio Giorgis, Royston Goodacre, Duncan Graham, Pietro La Manna, Stacey Laing, Lucio Litti, Fiona M. Lyng, Kamilla Malek, Cedric Malherbe, Maria P. M. Marques, Moreno Meneghetti, Elisa Mitri, Vlasta Mohaček-Grošev, Carlo Morasso, Howbeer Muhamadali, Pellegrino Musto, Chiara Novara, Marianna Pannico, Guillaume Penel, Olivier Piot, Tomas Rindzevicius, Elena A. Rusu, Michael S. Schmidt, Valter Sergo, Ganesh D. Sockalingum, Valérie Untereiner, Renzo Vanna, Ewelina Wiercigroch, and Alois Bonifacio. Surface Enhanced Raman Spectroscopy for Quantitative Analysis: Results of a Large-Scale European Multi-Instrument Interlaboratory Study. Analytical Chemistry 2020 92 (5), 4053-4064.

[14] DOU, Jingrui et al. Rapid discrimination of Brucellosis in sheep using serum Fourier transform infrared spectroscopy combined with PCA-LDA algorithm. Photodiagnosis and Photodynamic Therapy, v. 42, p. 103567, 2023.

[15] BS de Rezende, T Franca, MAB de Paula, HPK Cleveland, C Cena, Carlos A. do Nascimento Ramos. Turning chaotic sample group clusterization into organized ones by feature selection: Application on photodiagnosis of Brucella abortus serological test. Journal of Photochemistry and Photobiology B: Biology 247, 112781, 2023.

[16] MA. MOLAVI, HS. SAJJADI2, AA. NEJATIZADE. EFFECTIVE METHODS FOR APPROPRIATE DIAGNOSIS OF BRUCELLOSIS IN HUMANS AND ANIMALS. Online Journal of Animal and Feed Research. Vol 4, Issue 3, p 60-66. 2014.

[17] BRASIL. Ministério da Agricultura, Pecuária e Abastecimento, Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT). Brasília: IN MAPA/SDA/DSA, 10/2017, 2017.

[18] James M. Cameron, Holly J. Butler, David S. Palmer, Matthew J. Baker. Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying. Journal of Biophotonics, vol. 11, issue 4, e201700299, 2018.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. ´ Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[20] Jasper Engel, et al., Breaking with trends in pre-processing, TrAC Trends Anal. Chem. 50 (2013) 96–106.

[21] E L Kosarev, E Pantos. Optimal smoothing of 'noisy' data by fast Fourier transform. Journal of Physics E: Scientific Instruments, Volume 16, Number 6, 1983.

[22] I.T. Jolliffe, J. Cadima, Principal Component Analysis: a review and recent developments, Phil. Trans. R. Soc. A 374 (2016) 20150202

[23] D.R. Jensen, D.E. Ramirez. Use of Hotelling's T^2: Outlier Diagnostics in Mixtures. International Journal of Statistics and Probability, Vol. 6, issue 6, 2017.

[24] V.K. Chauhan, K. Dahiya, A. Sharma, Problem formulations and solvers in linear SVM: a review, Artif. Intell. Rev. 2018.

[25] Syed, Ali R. “A review of cross-validation and adaptive model selection.” 2011.

[26] Aline E Casaril, Carlos G Santos, Bruno S Marangoni, Sandro M Lima, Luis HC Andrade, Wagner S Fernandes, Jucelei OM Infran, Natália O Alves, Moacir DGL Borges, Cicero Cena, Alessandra G Oliveira. Intraspecific differentiation of sandflies specimens by optical spectroscopy and multivariate analysis. Journal of biophotonics 14 (4), e202000412, 2021.

[27] IC Oliveira, T Franca, G Nicolodelli, CP Morais, B Marangoni, G Bacchetta, Debora MBP Milori, Charline Z Alves, Cicero Cena. Fast and Accurate Discrimination of Brachiaria brizantha (A.Rich.) Stapf Seeds by Molecular Spectroscopy and Machine Learning. ACS Agricultural Science & Technology 1 (5), 443-448, 2021.

[28] Bruce, A, Adam, KE, Buller, H, Chan, KW & Tait, J. Creating an innovation ecosystem for rapid diagnostic tests for livestock to support sustainable antibiotic use', Technology Analysis and Strategic Management, vol 34, issue 11, 2022.

[29] V.C. Neta, J.P.S. Mol, M.N. Xavier, T.A. Paixão, A.P. Lage, R.L. Santos. Pathogenesis of bovine brucellosis. Vet. J., 184 (2010), pp. 146-155.

[30] Biosafety in microbiological and biomedical laboratories. Centers for Disease Control and Prevention (U.S.); National Institutes of Health (U.S.); June 2020, Pages in Document : xxvi,574 numbered pages. Series : HHS publication ; no. (CDC) 300859.

A2: Publicação em evento
Acesso: 10.1109/SBFotonIOPC62248.2024.10813483

FTIR spectroscopy of biofluids for disease diagnosis: Data quality influence on group separation

Thiago Franca
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
t.franca@ufms.br

Ana Maranni
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
ana.maranni@ufms.br

Camila Calvani
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
calvancamila@gmail.com

Miller Lacerda
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
miller.lacerda@ufms.br

Ronaldo Correa
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
ronaldoimecc@gmail.com

Cicero Cena
UFMS- Universidade Federal de Mato Grosso do Sul
Campo Grande-MS, Brazil
cicero.cena@ufms.br

***Abstract*—Fourier Transform Infrared (FTIR) Spectroscopy, particularly using the Attenuated Total Reflection (ATR) technique, has become a powerful tool in the analysis of biofluids for diagnostic purposes. The effectiveness of this technique in disease diagnosis relies heavily on the quality of the acquired spectral data. This study investigates the impact of key experimental parameters, such as the number of scans and spectral resolution, on the quality of FTIR spectra. By analyzing blood serum samples from different groups, we explore how these parameters influence the performance of Principal Component Analysis (PCA) and Support Vector Machine (SVM) classifiers in clustering tasks. Our findings indicate that while higher resolution does not necessarily enhance clustering, an increased number of scans can significantly improve classification performance when combined with lower-resolution settings. This work underscores the importance of optimizing FTIR data acquisition parameters to build robust predictive models in clinical diagnostics.**

Keywords— FTIR spectroscopy, biofluids, principal component analysis, experimental parameters.

I. INTRODUCTION

Fourier Transform Infrared (FTIR) Spectroscopy, particularly through the Attenuated Total Reflection (ATR) technique, is increasingly utilized to characterize biofluids in the development of innovative diagnostic methods – photodiagnosis methods. This approach enables the extraction of distinct molecular fingerprints from biological samples, such as blood serum and saliva, by analyzing the vibrational molecular modes. ATR-FTIR spectra have been successfully employed to differentiate between healthy and diseased states by identifying specific spectral variations in biofluids [1-4]. The technique shows rapid data acquisition, low cost, ability to analyze raw samples, and minimal sample volume requirements, which enables as a promising tool for non-invasive diagnostics in clinical settings [5].

Various factors, including the design of the interferometer, detector type, cooling methods, optical path length, sample preparation, and the selected spectral resolution influence the quality of FTIR spectra. Different FTIR spectrometers, depending on the manufacturer, provide spectra with varying levels of sensitivity, resolution, and signal-to-noise ratio. Optimizing these parameters is crucial for acquiring high-quality, reproducible spectra. The design of the interferometer, the performance of the detector, sample handling techniques, and data processing all contribute to the overall reliability and accuracy of FTIR measurements[6].

In FTIR spectroscopy, the number of scans refers to the total number of individual interferograms co-added or averaged to produce the final spectrum. Increasing the number of scans enhances the signal-to-noise ratio (SNR) by amplifying the desired signal while reducing random noise, resulting in a cleaner spectrum with more defined peaks and baselines. Additionally, a larger number of scans improves spectral stability and repeatability, reducing the impact of random fluctuations and aiding in the detection of subtle changes in the sample. The optimal number of scans varies depending on the application. For routine analysis, where speed is a priority, a moderate number of scans are recommended. However, in applications requiring high-quality spectra, such as analyzing complex mixtures or detecting trace components, a larger number of scans is often necessary to achieve the desired SNR and spectral stability [7].

Another important parameter is the spectral resolution, which refers to the ability to distinguish between closely spaced spectral features. Higher resolution allows finer spectrum detail detection, resulting in sharper and more defined peaks. This is essential for accurately identifying and quantifying compounds, as overlapping peaks can obscure critical information. Factors such as the optical path difference in the interferometer and the properties of the light source influence spectral resolution. Insufficient resolution compared to the linewidth of spectral features can lead to significant information loss, compromising the reliability of analyses [8].

Therefore, optimizing both parameters scan and resolution is essential for enhancing the quality and interpretability of FTIR spectra, particularly in complex mixtures where precise component identification is necessary to develop a reliable photodiagnostic method.

The FTIR spectra for group classification necessitate robust data analysis, often employing machine learning algorithms and multivariate analysis. In this context, data quality is paramount for effective clustering in Principal

2024 SBFoton International Optics and Photonics Conference (SBFoton IOPC) | 979-8-3503-8817-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/SBFOTONIOPC62248.2024.10813483

Component Analysis (PCA) and for influencing the performance of Support Vector Machine (SVM) classifiers. High-quality data ensures that PCA accurately captures the underlying structure of the dataset by identifying the principal components that represent the most variance. Poor data quality, characterized by noise, outliers, or missing values, can distort PCA results, leading to misleading data representations and incorrect clustering. Such misrepresentation affects SVM model training, as SVMs rely on PCA-derived features for classification. If these features are compromised due to low data quality, the SVM's predictive accuracy and generalization capabilities are diminished. Thus, ensuring high data quality is critical for both PCA analysis and the subsequent performance of SVM classifiers, directly impacting the reliability and interpretability of clustering results [9,10].

In this study, we explore how the optimization of scan parameters and resolution affects data quality, which is crucial before developing and applying methodologies for photodiagnosis using FTIR and machine learning. We examine the impact of these two parameters on sample clustering in PCA analysis, aiming to provide better data for future studies.

II. MATERIALS AND METHODS

A. *Sample description*

Nellore (Bos indicus) female blood serum was obtained from two different groups (pregnant and non-pregnant) as described elsewhere [11], after Animal Use Ethics Committee approval under protocol 1273/2023.

B. *FTIR data acquisition and pre-processing*

The female bovine blood serum was thawed at room temperature before the measurements. A small drop of around 20 µL of serum is placed directly onto the ATR crystal of an Agilent spectrometer, model Cary 630. The measurements were taken in 2000 to 900 cm^{-1} range, using deionized water as a background. FTIR-ATR spectra were acquired by combing scans of 64, 32, and 12 with resolutions of 16, 08, and 04 cm^{-1}.

Then, data pre-processing was conducted by normalizing the data using the Modified Standard Normal Variate (MSNV) method, which adjusts data dispersion and eliminates experimental misalignment. MSNV is calculated by subtracting the median of each data point and then dividing the result by the Median Absolute Deviation (MAD), thus normalizing the data set. The MSNV equation can be expressed as:

$$MSNV(xi) = (xi - Mediana(x))/(MAD(x)) \quad (1)$$

where xi represents the individual data value, Median(x) is the median of the data set x, and MAD(x) is the median of the absolute differences from the median, providing a robust measure of data dispersion.

C. *Principal Component Analysis - PCA*

The FTIR-ATR-MSNV data was submitted to Principal Component Analysis (PCA). PCA is a statistical technique employed to reduce the dimensionality of large datasets while preserving the maximum amount of variability inherent in the data. This is achieved by transforming the original variables into a new set of uncorrelated variables known as principal components. These components are ordered such that the first principal component captures the highest amount of variance, meaning it explains the most significant portion of the data's variability. Each subsequent principal component accounts for progressively less variance and is orthogonal to the previous ones [12].

In the context of PCA, variance refers to the degree of information or variability contained within the dataset. The primary objective of PCA is to maximize the variance captured by each principal component, ensuring that the most informative aspects of the data are retained in the reduced dimensional space.

The score plot, a graphical representation of the observations in the space defined by the principal components, serves as a critical tool for interpreting PCA results. Each point in the score plot represents an observation from the original dataset, plotted according to its values on the selected principal components. The interpretation of this plot is based on the clustering of points indicating similarity between observations based on the principal components, suggesting underlying patterns or groupings within the data. Conversely, a clear separation between groups in the score plot suggests that the principal components effectively distinguish between different classes or categories within the dataset. Additionally, outliers, which appear as points distant from the main cluster, may represent data points with unique characteristics that differ significantly from the rest of the dataset [13].

The axes of the score plot, typically corresponding to the first two or three principal components, often display the percentage of explained variance, which indicates how much of the data's variability is captured by these components. A higher percentage of explained variance suggests that the principal components provide a more accurate and meaningful representation of the original dataset

III. RESULTS AND DISCUSSION

The FTIR spectra obtained from liquid blood serum using an ATR accessory are presented in Figure 1. Figure 1(a) displays the average raw FTIR spectra, while Figure 1(b) depicts the average FTIR-MSNV spectra.

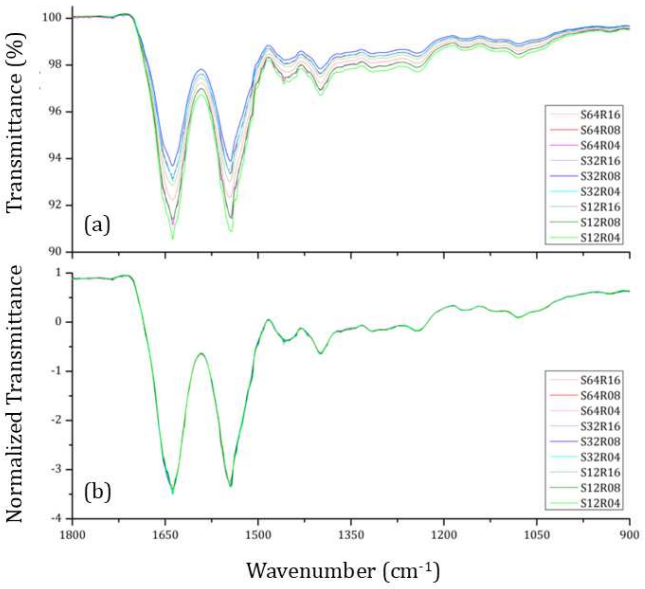


Fig. 1. *Blood serum FTIR spectra from liquid samples using ATR accessory. (a) raw spectra, and (b) normalized spectra with Modified Standard Normal Variate (MSVN). Different Scans: (i) 64 with resolution 16, 08, and 4 cm^{-1} ; (ii) 32 with resolution 16, 08, and 4 cm^{-1}; (iii) 12 with resolution 16, 08, and 4 cm^{-1}*

The two prominent vibrational bands observed are attributed to the Amide I and II groups associated with protein molecules. The weaker bands in the 1500 to 950 cm^{-1} range correspond to the vibrational modes of carbohydrates, lipids, and proteins present in the blood serum. A comprehensive analysis of these bands and their significance is provided in our previous related study [10].

In this study, we focus on identifying subtle features in the FTIR spectra that can complicate data analysis for sample classification using multivariate methods, such as Principal Component Analysis (PCA) combined with machine learning algorithms like Support Vector Machine (SVM). A close examination of the FTIR spectra in Figure 1(a) reveals the presence of minor shoulders in the two major bands within the 1650 to 1500 cm^{-1} range. These subtle shoulders, observed in a few scans, appear to be independent of resolution, suggesting a possible relationship with the signal-to-noise ratio (SNR).

Figure 2 shows the score plot from PCA results obtained using FTIR-MSNV spectra as input data. The impact of these minor shoulders in the FTIR spectra on clustering performance is evident. The score plot clearly demonstrates that high resolution does not necessarily lead to optimal clustering performance, regardless of the number of scans used. However, increasing the number of scans appears to improve the clustering of the blue group, with the best results achieved when a high number of scans is combined with low resolution

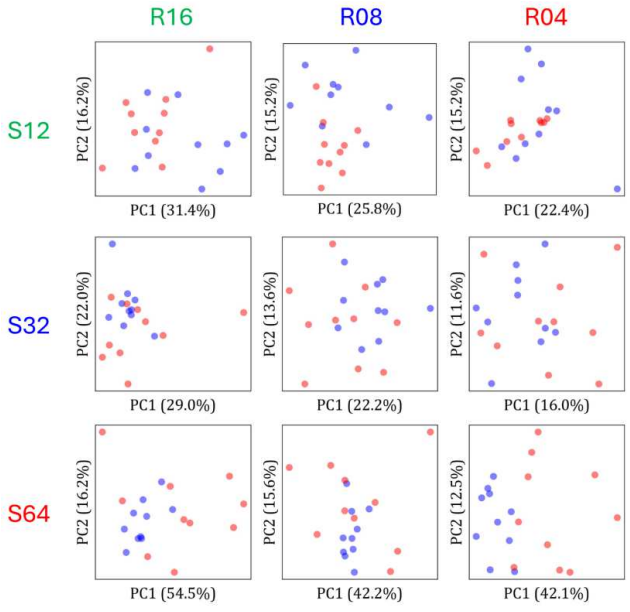


Fig. 2. *Blood serum FTIR spectra from liquid samples using ATR acessory. (a) raw spectra, and (b) normalized spectra with Modified Standard Normal Variate (MSVN). Diferent Scans: (i) 64 with resolution 16, 08 , and 4 cm^{-1} ; (ii) 32 with resolution 16, 08 , and 4 cm^{-1}; (iii) 12 with resolution 16, 08 , and 4 cm^{-1}.*

The quality of FTIR spectra is crucial for developing robust predictive models using machine learning algorithms. In addition to adhering to protocols for sample analysis [14], it is essential to monitor the equipment's performance during data acquisition by exploring various setup configurations, especially when working with liquid samples. This is particularly important because the signal from liquid samples is typically weaker compared to that from dried samples.

IV. CONCLUSION

This study highlights the critical role of FTIR spectral quality in the successful application of machine learning algorithms for disease diagnosis. Through the careful optimization of scan number and resolution, we demonstrated that higher spectral resolution does not always correlate with better clustering performance. Instead, a greater number of scans, particularly when paired with lower resolution settings, can enhance the ability of PCA to distinguish between different sample groups. These findings provide valuable insights for the design of FTIR-based diagnostic protocols, emphasizing the need for tailored data acquisition strategies to ensure reliable and accurate disease classification.

ACKNOWLEDGMENT

The Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), code 001. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), code 403651/2020-5; 302525/2022-0; 440214/2021-1. Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT), code 91/2023; 360/2022.

REFERENCES

[1] G. Pacher, T. Franca, M. Lacerda, N.O. Alves, E.M. Piranda, C. Arruda and C. Cena, "Diagnosis of cutaneous leishmaniasis using FTIR spectroscopy and machine learning: an animal model study," ACS Infect Dis, vol. 10, 2024

[2] Y.G. Marangoni-Ghoreyshi, T. Franca, J. Esteves, A. Maranni, K.D.P. Portes, C. Cena and C. Leal, "Multi-resistant diarrheagenic Escherichia coli identified by FTIR and machine learning: a feasible strategy to improve the group classification," RSC Adv, vol. 13, pp. 24909–24917, 2023

[3] E.C.A. de Brito, T. Franca, T. Canassa, S.S. Weber, A.M.M. Paniago and C. Cena, "Paracoccidioidomycosis screening diagnosis by FTIR spectroscopy and multivariate analysis," Photodiagnosis Photodyn Ther, vol. 39, 2022.

[4] G. Larios, M. Ribeiro, C. Arruda, S.L. Oliveira, T. Canassa, M.J. Baker, B. Marangoni, C. Ramos and Cícero Cena, "A new strategy for canine visceral leishmaniasis diagnosis based on FTIR spectroscopy and machine learning," J. of Biophotonics, vol. 14, 2021.

[5] C.I. Oliveira, T. Franca, G. Nicolodelli, C.P. Morais, B. Marangoni, G. Bracchetta, D.M.B.P. Milori, C.Z. Alves and C. Cena, "Fast and Accurate discrimination of Brachiaria brizantha (A.Rich.) staff seeds by molecular spectroscopy and machine learning," Agricultural Science & Technology, vol. 1,2021.

[6] H. Yahong, H. Lujia, Y. Yumei, L. Yanfei and L. Xian, "Key factors in FTIR spectroscopic analysis of DNA: the sampling technique, pretreatment temperature and sample concentration," Analytical Methods, vol. 10, pp. 24-36-2443, 2018.

[7] I. Barra, L. Khiari, S.M. Haefele, R. Sakrabani and F. Kebede, "Optimizing setup of scan number in FTIR spectroscopy using the moment distance index and PLS regression: application to soil spectroscopy," Sci. Rep., vol. 11, 13358, 2021.

[8] Y. Qin, J. Tong, X. Li, X. Han and M. Gao, "The Effect of Spectral Resolution on the Quantification of OP-FTIR Spectroscopy," Photonics, vol. 10, p. 475, 2023.

[9] S. Halouska and R Powers, "Negative impact of noise on the principal component analysis of NMR data," J. of Mag. Ressonance, vol. 178, p. 88-95, 2006.

[10] B.S. de Rezende, T. Franca, M.A.B. de Paula, H.P.K. Cleveland, C. Cena and C.A. do Nascimento Ramos, “Turning chaotic sample group clusterization into organized ones by feature selection: Application on photodiagnosis of Brucella abortus serological test,” J Photochem Photobiol B, vol. 247, 2023.

[11] W. Reis, T. Franca, C. Calvani, B. Marangoni, E. Costa e Silva, A. Nobre, G. Netto, G. Macedo and C.Cena, “Enhancing early identification of high-fertile cattle females using infrared blood serum spectra and machine learning,” Sci Rep, vol. 14, p. 19446, 2024.

[12] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” vol. 13, 2016.

[13] A.E. Casaril, C. Santos, B. Marangoni, S.M. Lima, L.H.C. Andrade, W.S. Fernandes, J.O.M. Infran, N.O. Alves, M.D.G.L. Borges, C. Cena and A.G. Oliveira, “Intraspecific differentiation of sandflies specimens by optical spectroscopy and multivariate analysis,” J Biophotonics, vol. 14, 2021.

[14] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. Butler, K. Dorling, P. Fielden, S. Fogarty, N. Fullwood, K. Heys, C. Hughes, P. Lasch, P. Martin-Hirsch, B. Obinaju, G. Sockalingum, J. Sulé-Suso, R. Strong, M. Walsh, B. Wood, P. Gardner and F. Martin, “Using Fourier transform IR spectroscopy to analyze biological materials,” Nat Protoc, vol. 9, no. 8, pp. 1771–1791, 2014