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

Curso de FÍSICA – BACHARELADO
Trabalho de Conclusão de Curso

Análise de soro sanguíneo para o fotodiagnóstico de brucelose bovina: Espectroscopia FTIR e análise multivariada

Kelvy Cezar Cordeiro Arruda

Orientador: *Prof. Cícero Rafael Cena da Silva*

Trabalho de Conclusão de Curso apresentado ao curso de Física Bacharelado do Instituto de Física (INFI), da Universidade Federal de Mato Grosso do Sul (UFMS).

Campo Grande – MS
Janeiro/2024

"Segui minha diretriz e venci, se a vida tá a minha mercê, merci"
(BK)

AGRADECIMENTOS

Agradeço a minha mãe e meu pai por todo o apoio, amor e carinho que me deram.

Agradeço a minha namorada Raquel Terêncio e ao meu amigo Victor Fidelis pela companhia e lealdade.

Agradeço aos meus professores que me deram exemplos de profissionalismo. Em especial ao meu orientador Cícero Cena pela atenção e paciência.

Agradeço a mim mesmo por sempre estar comigo e nunca ter perdido a fé em mim.

RESUMO

A brucelose é uma zoonose que gera grandes percas econômicas para países subdesenvolvidos devido aos danos causados a saúde dos animais bovinos infectados. Tendo em vista que é uma doença insidiosa, prejudicial ao ser humano e não existem vacinas totalmente eficientes, é necessário a criação de novas técnicas para erradicação da brucelose. O diagnóstico é uma boa alternativa para evitar a disseminação e atualmente a espectroscopia está sendo utilizada com sucesso no diagnóstico doenças [22],[21].

Utilizando a espectroscopia é possível extrair informações bioquímicas das amostras [23]. Porém, apenas a espectroscopia não é suficiente para criar um diagnóstico prático, necessitando da ajuda da análise multivariada e aprendizagem de máquina para criação de modelos preditores utilizando os dados dos espectros [24]. Então, essa pesquisa tem como objetivo: verificar a eficácia da espectroscopia no infravermelho por transformada de Fourier (FTIR) em combinação com análise multivariada e aprendizagem de máquina para a análise de soro sanguíneo bovino. A partir daí, criar um modelo preditor.

Utilizando 80 amostras de soro sanguíneo bovino, divididas em dois grupos: 40 animais infectados com *Brucella Abortus* e 40 animais não infectados. Os espectros no infravermelho médio foram obtidos por um espectrômetro de Infravermelho com transformada de Fourier e foram tratados utilizando *Standard Normal Variante (SNV)* e a *Principal Component Analysis (PCA)* e o algoritmo *Support Vector Machine (SVM)* foi escolhido como melhor classificador com uma acurácia de 94,6% na validação cruzada *Leave One Out Cross Validation (LOOCV)* e *91,7%* de acurácia na validação externa.

ABSTRACT

Brucellosis is a zoonosis that generates huge economic losses for underdeveloped countries due to the damage caused to the health of infected cattle. Given that it is an insidious disease, harmful to humans and that there are no fully efficient vaccines, it is necessary to create new techniques to eradicate brucellosis. Diagnosis is a good alternative to prevent dissemination and spectroscopy is currently being used successfully to diagnose diseases [22], [21].

Using spectroscopy, it is possible to extract biochemical information from samples [23]. However, spectroscopy alone is not enough to create a practical diagnosis, requiring the help of multivariate analysis and machine learning to create predictive models using spectral data [24]. Therefore, this research aims to: verify the effectiveness of Fourier transform infrared spectroscopy (FTIR) in combination with multivariate analysis and machine learning for the analysis of bovine blood serum. From there, to create a predictive model.

Using 80 bovine blood serum samples, divided into two groups: 40 animals infected with Brucella Abortus and 40 uninfected animals. The mid-infrared spectra were obtained by a Fourier Transform Infrared spectrometer and were treated using Standard Normal Variance (SNV) and Principal Component Analysis (PCA) and the Support Vector Machine (SVM) algorithm was chosen as the best classifier with an accuracy of 94.6% in Leave One Out Cross Validation (LOOCV) and 91.7% accuracy in external validation.

LISTA DE FIGURAS E TABELAS

SUMARIO

1. INTRODUÇÃO

Os países que fabricam produtos de origem animal enfrentam grandes dificuldades criadas por zoonoses, entre essas doenças a brucelose sendo endêmica no Brasil [15]. A brucelose bovina é uma doença de origem bacteriana causada geralmente por *Brucella Abortus* tendo como seu principal hospedeiro o gado bovino. Nos bovinos a principal característica é o aborto, que ocorre no final da gestação devido ao desenvolvimento de placentite e a queda da produção de leite [1],[2]. A transmissão mais comum da brucelose é feita por meio da vaca prenha infectada que expele os materiais do aborto ou secreções contaminando o pasto, a água e os fômites. A brucelose é uma doença ocupacional em áreas endêmicas que afeta diversos tratadores que lidam com os animais e é responsável por gerar grandes danos à saúde animal e à bovinocultura do Brasil [2]. Além disso, o descuido com a fiscalização do comércio de animais e má higienização das quintas aumentam a propagação da doença [1],[2].

O diagnóstico por meio do isolamento bacteriológico do patógeno é o padrão ouro. Como os sintomas e sinais clínicos não são específicos o diagnóstico baseia-se em resultados clínicos epidemiológicos e sorológicos [2]. Segundo o Programa Nacional de Controle e Erradicação da Brucelose e da Tuberculose Animal (PNCEBT), para o diagnóstico os testes mais comuns são: Teste de Soroaglutinação com Antígeno Acidificado Tamponado (AAT), o teste do 2-Mercaptoetanol (2-ME), Teste de Polarização Fluorescente (TPF), o Teste de Fixação do Complemento (TFC) e o Teste do Anel do Leite (TAL) sendo o AAT e o 2-ME usados para triagem, o TPF e o TFC como confirmatórios e o TAL como ferramenta de vigilâncias epidemiológica [12]. Algumas informações bem estabelecidas sobre esses testes serão dispostas a seguir.

O teste do Antígeno Acidificado Tamponado é o teste oficial de despistagem para a brucelose bovina. É um teste barato e simples de se fazer, porém é muito sensível ao gado, por essa razão deve ter testes complementares para evitar sacrifícios de animais não infectados [2],[12].O teste do 2 Mercaptoetanol apesar de mostrar bons resultados para o gado, mostra dificuldades no período inicial da doença e pode gerar resultados falso negativo [2],[12]. O teste de Polarização por

Fluorescência foi agregado ao programa posteriormente sendo um teste rápido, fácil e confiável e tendo uma especificidade superior aos testes usuais, todavia é um teste que pode ser caro devido ao custo dos reagentes e dos instrumentos utilizados para fazê-lo e necessita de uma mão de obra treinada e cuidadosa [2],[14]. O teste de Reação de Fixação de Complemento mostra alta especificidade sofrendo menos influência dos anticorpos das vacinas do que os testes AAT e 2-ME e com um custo ligeiramente baixo. Sendo referência para outros testes sorológicos e foi utilizado por vários países para erradicação da brucelose, porém é uma técnica muito trabalhosa e exige uma mão de obra laboratorial especializada [2],[12],[13]. O Teste do Anel de Leite à base de amostras de leite a granel é um teste de rastreio sendo utilizado no programa como um modo de fiscalizar as propriedades e garantir que essas estejam livres da doença, contudo não mostra eficiência em testes individuais apresentando uma taxa considerável de resultados falso positivo [2],[15].

Levando em consideração as informações sobre os testes aplicados atualmente e a situação endêmica da brucelose, a criação de novas técnicas para um diagnóstico mais rápido, fácil e barato se mostra necessário. Observando que a espectroscopia no infravermelho próximo se mostrou eficiente para classificação de produtos em diversas áreas incluindo agricultura e setor de alimentos [16].

2. OBJETIVO

O presente trabalho teve como objetivo estudar o potencial da espectroscopia no infravermelho por transformada de Fourier com o auxílio da análise de componentes principais (PCA) e aprendizagem de máquina para diagnosticar brucelose bovina. Utilizando a PCA para analisar e tratar os dados espectrais e a partir desses resultados treinar um algoritmo de aprendizagem de máquina para classificar amostras de soro sanguíneo bovino.

3. REVISÃO BIBLIOGRÁFICA

3.1 Espectroscopia no Infravermelho por Transformada de Fourier (FTIR)

A espectroscopia é uma técnica de análise que se baseia na interação entre a radiação eletromagnética e a matéria. A maioria dos compostos orgânicos e inorgânicos absorve várias frequências na região do infravermelho médio. Essa região de frequências é associada com energias capazes de alterar apenas os níveis vibracionais das substâncias. A espectroscopia no infravermelho por transformada de Fourier é um procedimento que visa obter o espectro infravermelho de forma mais simples e rápida do que pelo método convencional [3].

3.1.1 Espectro Infravermelho

A natureza da luz pode ser bem explicada por duas teorias: a teoria de ondas e a teoria corpuscular. A luz é uma onda eletromagnética com velocidade $c \approx 2{,}99\times10^{8}$ m/s. Para qualquer onda eletromagnética a frequência ν e comprimento de onda λ se conectam pela relação $c=\nu\lambda$. O espectro eletromagnético é o conjunto de todas as possíveis frequências (ou comprimentos de onda) da radiação eletromagnética [3]. A energia E referente a uma radiação eletromagnética é proporcional a frequência dessa onda pela relação $E=h\nu$, onde h é a constante de Planck. Se várias radiações infravermelhas com diferentes frequências incidem em uma molécula, um padrão de absorção de energia será formado. A representação gráfica das intensidades de absorção (ou transmissão) no infravermelho em função da frequência é chamado de espectro infravermelho [3].

A parte do espectro eletromagnético denominada infravermelho está situada entre os comprimentos de onda associados a luz visível, aproximadamente de 400 a 800 nm, e os comprimentos de onda associados a micro-ondas, que são maiores que 1 mm. O intervalo de interesse deste trabalho é o infravermelho médio ou infravermelho vibracional que é normalmente aceito entre 2,5 μm e 25 μm. Contudo é mais comumente se referir a região vibracional do

infravermelho em termos de número de onda. Os números de onda são apresentados em cm^{-1} assim o infravermelho vibracional vai de 4000 a 400 cm^{-1} [3].

3.1.2 Espectrômetro de Infravermelho.

O instrumento que permite a obtenção do espectro de absorção de moléculas ativas na região do infravermelho médio é o espectrômetro de infravermelho. Dentre os aparelhos usualmente utilizados encontramos dois tipos distintos. O espectrômetro dispersivo, que conta com uma rede de difração e um monocromador para manipular o feixe que passa pela amostra e assim produzir o espectro já no domínio da frequência . E o espectrômetro de transformada de Fourier (FTIR), que utiliza uma transformação matemática dos dados para converter um sinal de interferência em espectro infravermelho no domínio da frequência. O FTIR é mais utilizado pois apresenta maior sensibilidade [3].

A figura 1 apresenta um diagrama esquemático de um espectrômetro FTIR. Nesse diagrama a radiação infravermelha incide em um interferômetro, nesse interferômetro há um divisor de feixes que o separa em dois feixes perpendiculares. Um feixe tem sua direção desviada em 90° e o outro mantém a direção original. O feixe desviado é refletido por um espelho fixo (sua posição é mantida fixa) e volta para o divisor de feixe, o outro feixe é refletido por um espelho móvel e volta para o divisor de feixe. Então quando os dois feixes se encontram novamente no divisor o resultado é um feixe com padrões de interferência [3].

Quando o feixe recombinado passa pela amostra as moléculas constituintes do meio absorvem de forma simultânea todas as frequências normalmente encontradas no seu espectro infravermelho. Esse feixe recombinado contém toda a energia radiada provinda da fonte e é obtido em um intervalo de comprimento de ondas de 4000 a 400 cm-1. Então esse sinal (interferograma) chega ao detector e é comparado com o interferograma de referência (obtido pelo material do divisor de feixes), que é subtraído do sinal obtido da amostra via software - o interferograma é um sinal que está no domínio do tempo e que contém todas as frequências que formam o espectro do

infravermelho-. Finalmente, o computador aplica a transformada de Fourier fornecendo o espectro infravermelho FTIR [3].

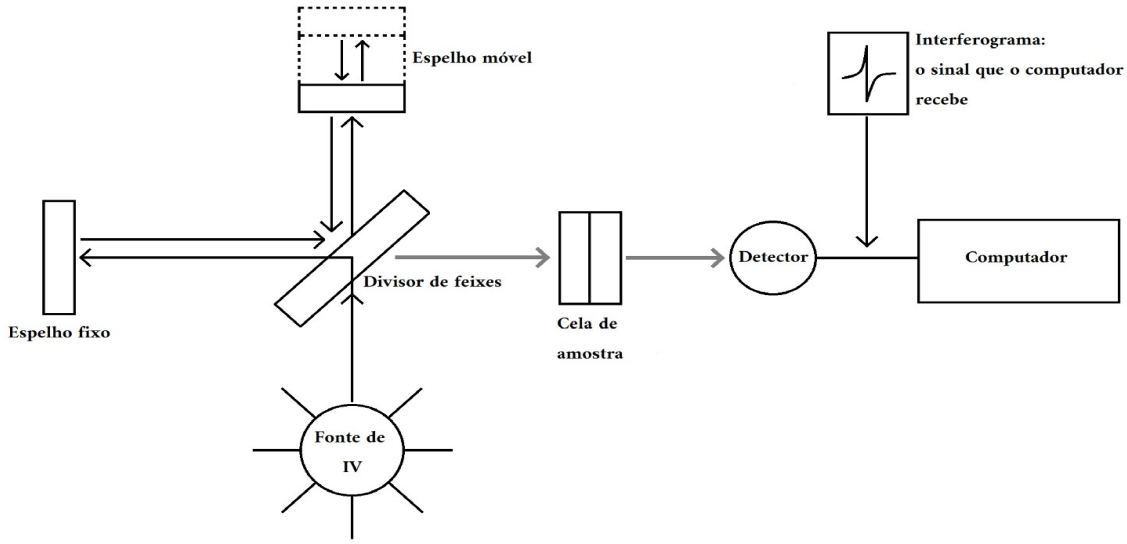

Figura 1: Diagrama do espectrômetro de transformada de Fourier.
Fonte: Adaptado de PAIVA, Donald.2010 [3].

Nesse caso a transformada de Fourier é empregada para transformar os dados do domínio temporal para o domínio da frequência. A transformada de Fourier decompõe o sinal complexo das intensidades em função do tempo em vários sinais de intensidade em função das frequências e assim nos dando as frequências de absorção contidas no sinal do interferograma. Isso gera um espectro virtual idêntico aos gerados por espectrômetros com instrumentos dispersivos [9].

3.1.3 Refletância Total Atenuada (ATR)

É uma técnica proposta por Fahrenfort e Harrick em 1960 desenvolvida para obter espectros de infravermelho semelhantes a espectros de transmissão. A ATR é baseada na propriedade de uma onda chamada de reflexão interna total dando origem a uma onda evanescente [25]. A reflexão

interna total ocorre quando o meio incidente é mais denso e o ângulo de incidência é superior ao ângulo crítico [9]. Essa técnica de amostragem é interessante para o presente estudo pois permite analisar as amostras no estado líquido sem a necessidade de preparação e utilizando uma pequena quantidade de amostra para cada medição [25].

3.1.4 Interação da radiação eletromagnética com a matéria

Quando uma radiação eletromagnética interage com uma molécula o resultado depende da energia dessa radiação. Como já foi dito, a energia de um fóton é proporcional a frequência ou inversamente proporcional ao comprimento de onda da onda eletromagnética associada a esse fóton. É bem estabelecido que uma molécula não pode deter qualquer valor de energia, então quando um fóton tem exatamente a energia necessária para promover essa molécula para um nível energético maior verifica-se a absorção desse fóton pela molécula [3].

Portanto, se radiações com comprimentos de onda na região do infravermelho interagirem com uma molécula, alguns comprimentos de onda serão absorvidos e outros não. Se após a interação com a molécula, for efetuado a aferição das intensidades de cada uma dessas radiações o resultado será um gráfico de absorção de radiação no infravermelho e esse é denominado espectro no infravermelho. Duas moléculas diferentes não podem produzir o mesmo espectro [3].

No infravermelho médio a energia das radiações são capazes de mudar apenas os estados vibracionais ou rotacionais de uma molécula. Nesse processo a molécula absorve as radiações aumentando assim a sua amplitude da vibração. Contudo apenas as moléculas com ligações que tem um momento de dipolo variando com o tempo são capazes de absorver essa radiação. A Figura 2 mostra os principais modos de movimento vibracional de uma molécula no infravermelho, que são divididos em dois grupos: estiramento e dobramento. Os tipos de estiramento mais comuns são: simétrico e assimétrico, e os tipos de dobramento mais comuns são: *scissoring* (tesoura), *wagging* (balanço), *rocking* (rotação) e *twisting* (torção) [3].

A força de ligação e a massa dos átomos vão interferir na frequência de absorção no infravermelho. Isso é facilmente visualizado com uma molécula heteronuclear vibrando no modo estiramento que tem seu movimento análogo a um sistema de duas massas conectadas por uma mola. Aqui duas coisas são importantes. A primeira é que ligações mais fortes, vibram mais do que ligações mais fracas com as mesmas massas. E a segunda é que para o mesmo tipo de ligação, átomos maiores vibram em frequências mais baixas que átomos mais leves. As radiações que compõem a região do infravermelho médio correspondem a faixa de frequências vibracionais das ligações na maioria das moléculas [3].

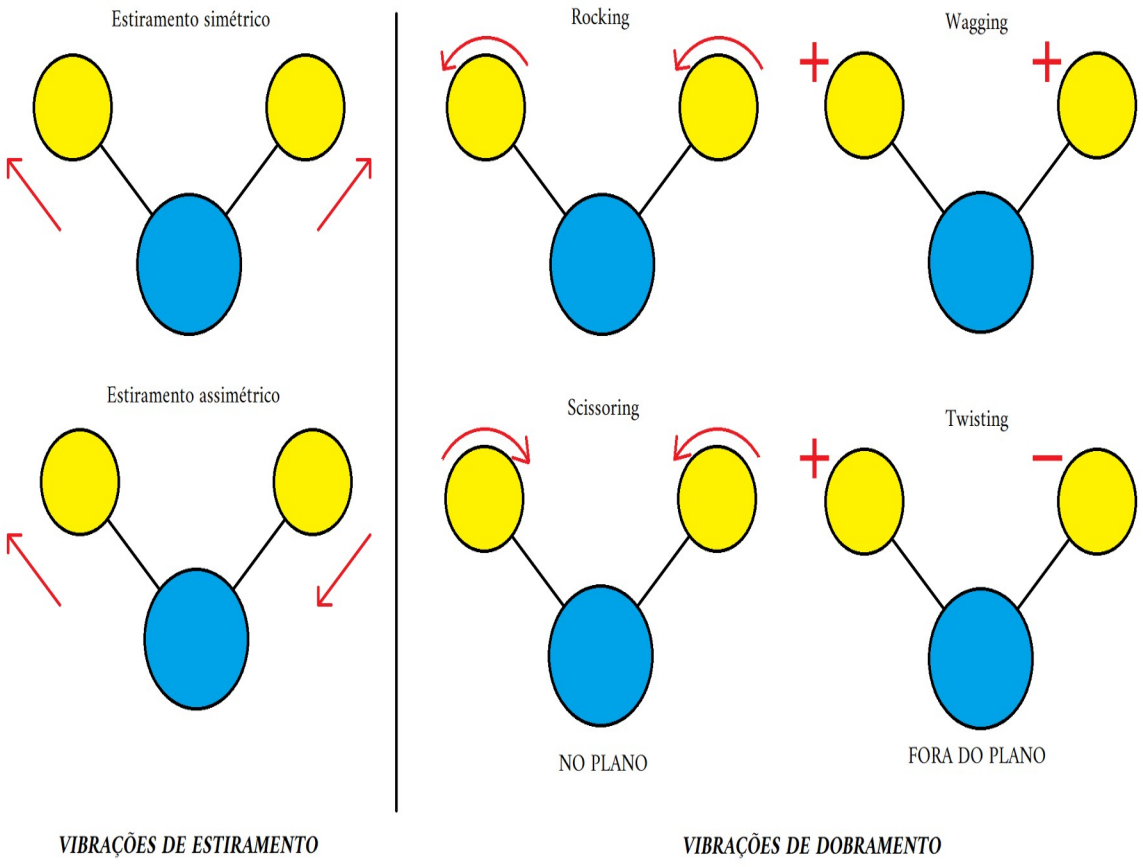

Figura 2: Modos vibracionais ativos no infravermelho.
Fonte: Adaptado de PAIVA, Donald.2010 [3].

3.2 Análise Multivariada e Aprendizado de Máquina

Na região do infravermelho do espectro, a quantidade de variáveis (intensidades de radiação para cada número de onda) é gigantesca, por esse motivo é necessário a análise multivariável feita algoritmos embasados em métodos estatísticos. A aprendizagem de máquina é um campo de estudo que busca desenvolver sistemas que melhoram através da experiência e entender as leis estatísticas algoritmo-teóricas que ditam o comportamento dos sistemas de aprendizagem [10].

A aprendizagem de máquina é dividida em três tipos. O primeiro tipo é a aprendizagem supervisionada em que o algoritmo por meio dos dados de entrada aperfeiçoa a sua capacidade de distinção devido a propriedade essenciais de cada classe, mesmo que as semelhanças e as diferenças entras as classes sejam muitas, é importante ressaltar que cada exemplo de dados de entrada é associado a um rótulo de classificação conhecido [26].

O segundo tipo é o algoritmo não supervisionado, esse tipo é focado para resolução de problemas com uma variedade de graus de liberdade, nesse caso o algoritmo se adapta para obter um resultado melhor não associando um dado de entrada com o resultado mas criando seu próprio caminho para se aperfeiçoar [26].

E o terceiro é a aprendizagem por reforço cujo o algoritmo se aperfeiçoa por meio de recompensas ou não, dependendo do seu resultado em se aproximar de uma função almejada [26].

As técnicas de aprendizado de máquina tem sido um fator cada vez mais relevante em diversas áreas principalmente nas áreas da saúde [6] e em áreas como psicologia, biologia e antropologia. Em tais áreas é comum trabalhar com várias variáveis simultaneamente; podemos querer considerar várias características de uma planta ou indicar diferenças raciais por diferenças das medidas dos crânios [11]. Ao usar algoritmos para estudar diversas variáveis essa técnica é denominada análise multivariada [27].

3.2.1 Pré Processamento de Dados

Para conseguir interpretar os dados do espectro no infravermelho próximo é necessário aplicar um procedimento de normalização dos espectros para evitar problemas de variância dos dados. Os efeitos resultantes de variações físicas do sistema são então contornados pela transformação matemática chamada Standard Normal Variante (SNV), deixando assim apenas as variações químicas das amostras, que é o desejado [20].

Essa transformação matemática é aplicada utilizando a Equação 1.

$$x_{i,j}^{SNV} = \frac{x_{i,j} - u_i}{s} \tag{1}$$

Onde x_i^{snv} é o valor corrigido pelo SNV para o i-ésimo comprimento de onda do espectro. O x_i é o i-ésimo comprimento de onda do espectro. O u_i é a média das intensidades nos diferentes comprimentos de onda x_i do espectro. E s é o desvio padrão das intensidades dos diferentes comprimentos de onda x_i do espectro dado pela Equação 2.

$$s = \sqrt{\frac{\sum_{j=1}^{p} (x_{i,j} - \bar{x}_i)^2}{p-1}} \tag{2}$$

O SNV, em diversos casos [21], é utilizado como o primeiro passo para a técnica de tratamento dos dados chamada de *Analysis Component Principal* (PCA), ou em português análise de componentes principais [4].

3.2.2 Análise de Componentes Principais (PCA)

A análise de componentes principais é uma técnica estatística que faz uma transformação linear em um conjunto de variáveis correlacionadas gerando um conjunto de variáveis não correlacionadas, chamados de componentes principais (PCs) [5],[4]. Os principais objetivos da PCA são diminuir a dimensão dos dados, revelar padrões ocultos e facilitar a visualização dos resultados. A PCA visa reduzir o conjunto das variáveis que estão sendo investigadas para um número muito menor e gerenciável, essa redução facilita a representação dos dados e examinando esse novo conjunto se pode encontrar relações entre as variáveis originais e os componentes principais [4].

A ideia principal da PCA é encontrar uma nova base que maximize a variação dos dados, isto é, encontrar novas variáveis que são combinações lineares daquelas do conjunto de dados originais e que não estejam correlacionadas entre si. O primeiro elemento da base, ou o primeiro PC, é o eixo que tem a maior variação de dados, o segundo eixo é o que tem a segunda maior variação dos dados e assim por diante. Encontrar esses PCs se resume a um problema de autovalor e autovetor encontrados a partir de uma matriz de covariância ou uma matriz de correlação dos dados originais [5].

As variáveis são características medidas em cada observação, por exemplo, absorção ou transmitância de uma amostra, dispostas como colunas na matriz. As observações são conjuntos de valores para as variáveis, por exemplo, o conjunto de comprimentos de ondas da medida de uma amostra, representadas como linhas na matriz [4].

Após a normalização dos dados (SNV) é construída duas matrizes: a matriz de covariância e a matriz de correlação dos dados. A matriz de covariância é construída a partir dos dados originais e visa entender como duas variáveis mudam juntas. Já a matriz de correlação é construída a partir dos dados padronizados com o objetivo de capturar as relações lineares entre os pares de variáveis independente das suas escalas. Utilizando uma dessas duas matrizes encontram-se os autovalores e

autovetores. Os autovetores representam as direções dos eixos do espaço dos dados e os autovalores representam a variância dos dados nessas direções [4].

Após encontrar os autovetores e autovalores, que formam os componentes principais, é escolhido alguns PCs e os dados padronizados são projetados neles. Isso é feito multiplicando a matriz de dados padronizados pelo autovetor que representa o PC escolhido, o que caracteriza uma transformação linear dos dados para um espaço que tem como base os PCs escolhidos. Essa projeção dos dados resulta em um conjunto de pontos no espaço dos componentes principais selecionados chamados de scores. Esses scores representam a contribuição de cada componente principal selecionada para explicar a variância dos dados, o que gera uma representação compacta e informativa dos mesmos [4].

A interpretação dos PCs envolve entender a contribuição de cada variável original na formação dos PCs, para entender como isso acontece é necessário introduzir o conceito de carga do componente principal. As cargas, ou loadings, indicam a direção e a magnitude da contribuição de cada variável original para a criação de um determinado PC e são definidas como coeficientes [4].

Isso facilita na seleção de variáveis relevantes para encontrar padrões nos dados, no entanto o conhecimento do contexto, por exemplo, entender quais fenômenos as variáveis representam dentro do cenário estudado é necessário para uma boa interpretação das cargas e consequentemente uma escolha de PCs mais significativos. É importante destacar que os dados no caso da espectroscopia são as bandas dos espectros [4].

3.2.3 Análise Supervisionada

Um algoritmo de aprendizagem de máquina é o *Support Vector Machines* (SVM), ou máquina de vetores de suporte, esse algoritmo de aprendizagem de máquina supervisionado é empregado como classificador e sua ideia principal é buscar determinar um hiperplano ótimo que separe as classes no espaço de características utilizando uma combinação linear de funções

parametrizadas por vetores de apoio. A classificação pode ser linear ou não linear, quando o SVM consegue encontrar um hiperplano linear que separa os grupos essa é chamada uma classificação linear, porém quando isso não é possível o SVM plota os dados em um espaço de dimensão muito maior recorrendo a um processo chamado truque de *kernel* reincidindo no caso da classificação linear [18].

O SVM classifica uma amostra por meio de uma função que atribui valores para a amostra indicando em que lado do hiperplano ela se encontra. A margem é a distância entre as amostras mais próximas de cada classe e o hiperplano logo um aspecto que apresenta o sucesso do algoritmo é ter uma margem vultosa. Os vetores de suporte são as amostras mais próximas do hiperplano e são muito importante uma vez que o SVM leva em consideração apenas esses vetores para definir o hiperplano – uma característica que é computacionalmente interessante [8].

Um hiperparâmetro importante para o SVM é o parâmetro C que controla um equilíbrio entre o tamanho da margem de decisão e classificação correta de uma classe. Um parâmetro C maior significa uma margem menor e diminui erros de classificação por outro lado um parâmetro C menor significa uma margem maior em troca de alguns erros de classificação [8].

Para a classificação não linear a função de decisão dá origem as funções básicas chamadas de kernel que permitem mapear os dados em um espaço com uma dimensionalidade muito maior [7],[18]. Uma função de kernel muito utilizada é a Radial Basis Fuction (RBF) que mapeia os dados em um espaço de dimensão infinita utilizando uma função de base radial. Nessa função de Kernel existe um parâmetro γ (gamma) que constitui a largura da banda de influência das amostras vizinhas [19].

3.2.4 Validação Cruzada.

Um problema da aprendizagem de máquina é quando os algoritmos tendem a entrar em sobreajuste. Quando o classificador consegue classificar com excelência o grupo de dados inseridos

para treiná-lo mas não consegue classificar um conjunto novo de dados da mesma população é dito que o modelo entrou em sobreajuste. É uma técnica que avalia a capacidade de um modelo de aprendizado de máquina em generalizar sua capacidade de predição a partir de um conjunto de dados e evitar o sobreajuste [17].

Agregar mais dados da mesma população utilizada para construir esse modelo seria a melhor forma de avaliá-lo, porém isso é inviável no primeiro momento, então primeiramente se mede o desempenho preditivo do sistema utilizando a validação cruzada [17].

O subtipo de validação cruzada utilizado nesse trabalho é o *leave one out cross validation (LOOCV).* Nesse modelo o erro obtido não carrega viés em comparação ao erro de previsão verdadeiro no entanto tem uma variância elevada devido ao seu processo de validação isso gera um custo computacional proporcional ao número de dados [17].

4. METODOLOGIA

4.1 Coleta do soro sanguíneo.

As amostras foram obtidas do -Laboratório de Biologia Molecular FAMEZ (Faculdade de Veterinária e Zootecnia)-. Foram coletadas 80 amostras de soro sanguíneo bovino. Os soros foram armazenados em microtubos a -20°C e foram separados em dois grupos: 40 animais infectados com *B. abortus* e 40 animais não infectados (grupo controle).

4.2 Obtenção de Espectros FTIR.

Os espectros no infravermelho médio foram obtidos pelo espectrômetro de Infravermelho com transformada de Fourier (Agilent Cary 630 FTIR) com um acessório de reflectância total

atenuada. Cada amostra de soro foi medida no intervalo de 2000 cm^{-1} a 900 cm^{-1} com uma resolução de 4 cm^{-1} e doze scans utilizando água como fundo.

4.3 Análise de dados.

O intervalo analisado foi de 1800 cm^{-1} a 900 cm^{-1}, esse intervalo foi escolhido pois é nessa região que apareceram bandas relevantes para a discriminação das classes. Foi realizado um pré-processamento dos dados das medidas dos espectros FTIR de todas as amostras fazendo uma normalização utilizando o *standart normal variate* (SNV) [20] e então os dados foram tratados utilizando a PCA. Após o tratamento os algoritmos de aprendizagem de máquina foram treinados e validados e o com melhor desempenho foi selecionado como classificador. A métrica de avaliação utilizada para escolher o melhor classificador foi a maior acurácia pelo menor número de PCs possíveis. O baixo número de PCs utilizados é um bom parâmetro de eficiência para o modelo preditor pois quanto mais PCs forem utilizado maior a chance de o modelo entrar em sobreajuste [17].

5. RESULTADOS E DISCUSSÃO

As médias dos espectros das amostras estão dispostas na Figura 4. Os espectros vão de 2000 cm^{-1} à 900 cm^{-1} e pode-se extrair de antemão algumas informações químicas sobre as amostras. No intervalo de 1710 cm^{-1} a 1594 cm^{-1} temos uma banda fina e forte com um pico na faixa de 1638 cm^{-1} podendo indicar um estiramento C=C ou um estiramento C=O esses estiramentos podem representar amida I [28].

A seguir temos outra banda que vai de 1594 cm^{-1} a 1480 cm^{-1} forte e fina com um pico em 1543 cm^{-1} podendo indicar um estiramento N-O ou um estiramento N-H esses modos vibracionais podem ser associados a albumina ou IgG4. Na região entre 1480 cm^{-1} e 1430 cm^{-1} podemos ver uma banda fina e fraca com um pico em 1457 cm^{-1} indicando um dobramento C-H associado a apolipoproteína A [28].

Entre o intervalo de 1430 cm^{-1} a 1370 cm^{-1} tem uma banda fina e média com um pico em 1400 cm^{-1} indicando um dobramento O-H, esse dobramento é bem-aceito como um indicador de aminoácidos; proteínas. De 1265 cm^{-1} a 1200 cm^{-1} com um pico em 1245 temos uma banda fraca e fina podendo indicar um estiramento P-O associado a fosfolipídios [28].

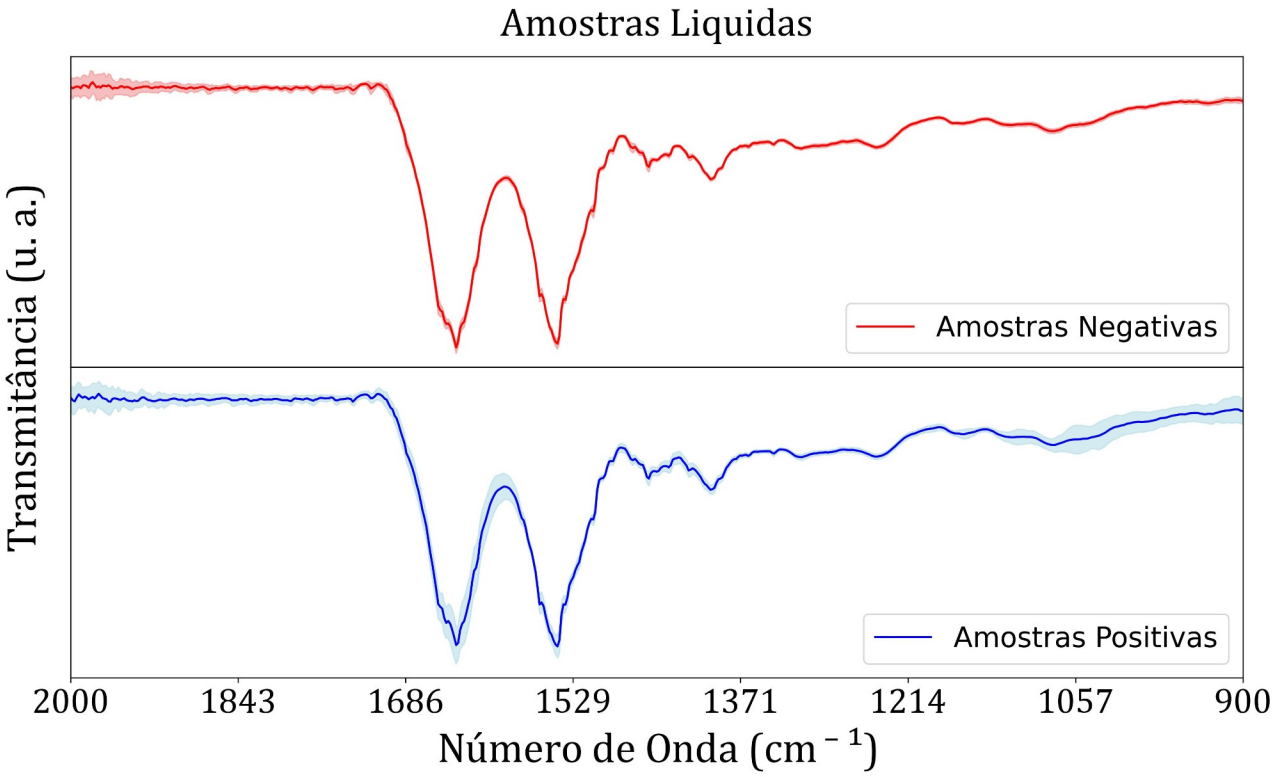

Figura 3: Espectros no Infravermelho médio por Transformada de Fourier para amostras positivas e negativas.

Fonte: Próprio Autor.

5.1. Análise de PCA

Na Figura 4 está disposto o scores da PCA para o espectro FTIR para o intervalo de 1800 cm^{-1} a 900 cm^{-1} projetados em relação aos PC1 e PC2 (parte direita) e suas respectivas cargas (parte esquerda), pode-se observar que existe uma concentração dos scores, porém não é suficiente para classificar as classes indicando que apenas os dois primeiros componentes principais não são suficientes para separar as amostras por grupos.

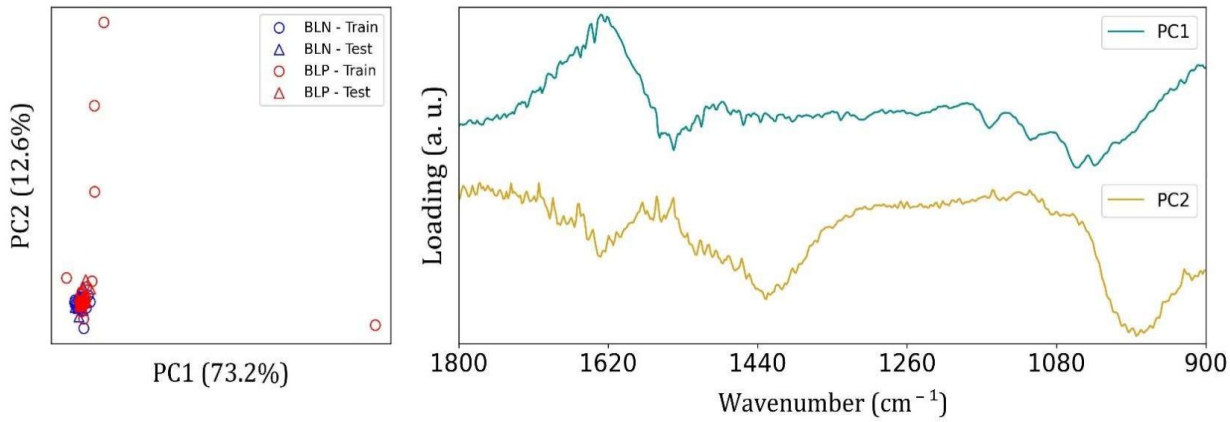

Figura 4: Na esquerda os scores das amostras negativas de treino (círculos azuis), amostras negativas de teste (triângulos azuis), amostras positivas de treino (círculos vermelhas) e amostras positivas de testes (triângulos vermelhos). Onde BLN são amostras não infectadas com *Brucella* e BLP são amostras infectadas com *Brucella*. Na direita os seus respectivos *Loadings* no intervalo de 1800 cm^{-1} a 900 cm^{-1} para os PC1 e PC2.
Fonte: Próprio Autor.

As cargas na Figura 4 mostram a variação dos dados para os primeiros 2 componentes principais. As principais variações dos dados foram observadas em torno de 1640 cm^{-1}, 1550 cm^{-1}, 1435 cm^{-1} e 985 cm^{-1} atribuídas normalmente a bandas vibracionais de amida I e II e lipídios possivelmente relacionadas com a presença de IgG2, IgG3 e IgG4 [28].

5.2. Aprendizagem de Máquina

Seguindo as métricas o classificador escolhido foi o SVM que atingiu uma acurácia de 96,4% para amostras de brucelose negativas e 92,9% para amostras de brucelose positivas. Os parâmetros do SVM foram C=100, com o kernel RBF e gamma=0,0001.

Na parte esquerda da Figura 5 está disposta a matriz de confusão para o método de aprendizagem de máquina SVM construído a partir da validação cruzada *leave one out cross validation.* O método utilizado para a caracterização mostrou um erro de 7,1% para amostras negativas e 3,6% para amostras positivas. Na parte direita da Figura 5 está disposta a matriz de confusão para a validação externa para o mesmo classificador, pode-se notar que a acurácia desse método é de 91,7%.

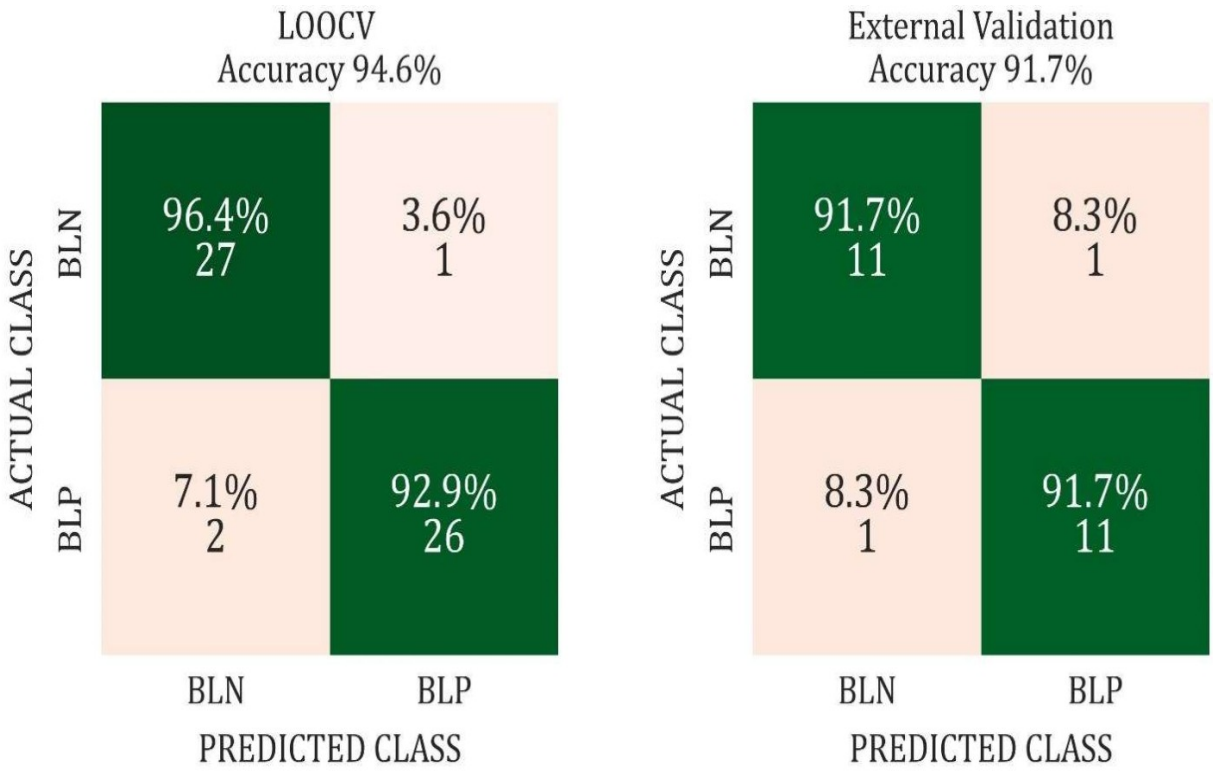

Figura 5: A esquerda encontra-se a matriz de confusão para a validação cruzada do SVM. A direita encontra-se a matriz de confusão para a validação externa do método SVM de aprendizagem de máquina.
Fonte: Próprio Autor.

Portanto, o presente trabalho obteve êxito em atingir objetivo de validar o potencial da espectroscopia no infravermelho por transformada de Fourier com o auxílio da PCA e aprendizagem de máquina como uma ferramenta para diagnosticar brucelose bovina. Utilizando a PCA em conjunto com o algoritmo de aprendizagem de máquina SVM foi possível gerar um classificador que contém uma acurácia acima de 90% o que é um resultado observado em outros testes atualmente utilizados [12],[13].

6. CONCLUSÃO

A espectroscopia FTIR em conjunto com a análise multivariada e aprendizagem de máquina mostrou eficiência para a discriminação de amostras de soro sanguíneo bovino. A média dos espectros das amostras apresentaram bandas associadas a componentes moleculares encontrados no sangue bovino. E usando a PCA associada ao algoritmo de aprendizagem de máquina SVM, foi possível gerar um bom classificador para diferenciar entre amostras infectadas e não infectadas com *Brucella Abortus*, esse modelo preditor mostrou uma acurácia de 94,6% para a validação cruzada e 91,7% para a validação externa. Tais resultados apresentam relevância para a bovinocultura do Brasil, tendo em vista que essa técnica tem um custo mais acessível e a preparação das amostras é mais simples em comparação a outras técnicas de diagnósticos utilizadas atualmente.

REFERÊNCIAS

[1] Bercovich Z., Maintenance of Brucella abortus-free herds: a review with emphasis on the epidemiology and the problems in diagnosing brucellosis in areas of low prevalence, Vet Q. 1998 Jul;20(3):81-8. doi: 10.1080/01652176.1998.9694845. PMID: 9684294.

[2] Barbuddhe, Sukhadeo & Vergis, Jess & Rawool, Deepak. (2020), Immunodetection of bacteria causing brucellosis. 10.1016/bs.mim.2019.11.003.

[3] PAIVA, Donald. Introdução à Espectroscopia 4ª edição. São Paulo: Cengage2010.

[4] Beattie JR, Esmonde-White FWL, Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra, Applied Spectroscopy. 2021;75(4):361-375. doi:10.1177/0003702820987847

[5] Jolliffe IT, Cadima J., Principal component analysis: a review and recent developments, Philos Trans A Math Phys Eng Sci. 2016 Apr 13;374(2065):20150202. doi:10.1098/rsta.2015.0202.

[6] Zhang Z., Introduction to machine learning: k-nearest neighbors, Ann Transl Med. 2016 Jun;4(11):218. doi: 10.21037/atm.2016.03.37.

[7] Belousov, Anton & Verzakov, S. & Von Frese, Juergen. (2002), Applicational aspects of support vector machines, Journal of Chemometrics. 16. 482 - 489. 10.1002/cem.744.

[8] Fichou, Dimitri & Ristivojevic, Petar & Morlock, Gertrud. (2016), Proof-of-Principle of rTLC, an Open-Source Software Developed for Image Evaluation and Multivariate Analysis of Planar Chromatograms, Analytical Chemistry. 88. 10.1021/acs.analchem.6b04017.

[9] Kiyoshi Yamamoto, Hatsuo Ishida, Optical theory applied to infrared spectroscopy, Vibrational Spectroscopy. 2031(94)00022-9. https://doi.org/10.1016/0924.

[10] M. I. Jordan T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science . 349, 255-260(2015). DOI:10.1126/science.aaa8415

[11] Bartlett, M. S., “Multivariate Analysis.” vol. 9, no. 2, 1947, pp. 176–97, Journal of the Royal Statistical Society. Accessed 28 July 2023. https://doi.org/10.2307/2984113.

[12] Meirelles-Bartoli, R. B., & Mathias, L. A.. (2010), ESTUDO COMPARATIVO ENTRE OS TESTES ADOTADOS PELO PNCEBT PARA O DIAGNÓSTICO SOROLÓGICO DA BRUCELOSE EM BOVINOS. Arquivos Do Instituto Biológico, 77(1), 11–17. https://doi.org/10.1590/1808-1657v77p0112010.

[13] Mathias, L. A., Meirelles, R. B., & Buchala, F. G.. (2007), Estabilidade do antígeno de célula total de Brucella abortus para uso no diagnóstico sorológico da brucelose bovina pela reação de fixação de complemento, Pesquisa Veterinária Brasileira, 27(1), 18–22. https://doi.org/10.1590/S0100-736X2007000100004

[14] Mathias, L. A., Corbellini, L. G., Maia, L., Nascimento, K. F., Paulin, L. M. S., Samartino, L. E., Serqueira, M. A., Soares Filho, P. M., & Souza, M. M. A. De .. (2010), Validação interlaboratorial do teste de polarização fluorescente para o diagnóstico sorológico da brucelose bovina, Ciência Rural, 40(10), 2135–2140. https://doi.org/10.1590/S0103-84782010005000161.

[15] Silva Júnior FF, Megid J, Nozaki CN, Pinto JPAN, Avaliação do teste do anel em leite na vigilância epidemiológica da brucelose bovina em rebanhos e em laticínios, Arq Bras Med Vet Zootec. 2007Apr;59(2):295–300. https://doi.org/10.1590/S0102-09352007000200004.

[16] Tutuncu, K. (2021), A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy, European Journal of Science and Technology. https://doi.org/10.31590/ejosat.882749

[17] Daniel Berrar, Cross-Validation, Editor(s): Shoba Ranganathan, Michael Gribskov, Kenta Nakai, Christian Schönbach, Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 2019. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

[18] Cortes, C., Vapnik, V., Support-vector networks, *Mach Learn* 20, 273–297 (1995). https://doi.org/10.1007/BF00994018.

[19] Lorena, Ana & Cognição, Computação & Carvalho, Andre & Computação, Departamento & Computação, Instituto & USP. (2007), Uma Introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada. 43-67. 14. 10.22456/2175-2745.5690.

[20] Rinnan, Åsmund & Berg, Frans & Engelsen, Søren. (2009), Review of the Most Common pre-Processing Techniques for Near-Infrared Spectra, TrAC Trends in Analytical Chemistry. 28. 1201-1222. 10.1016/j.trac.2009.07.007.

[21] Bruno Silva de Rezende, Thiago Franca, Maykko Antônyo Bravo de Paula, Herbert Patric Kellermann Cleveland, Cícero Cena, Carlos Alberto 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, Volume 247, 2023, 112781, ISSN 1011-1344. https://doi.org/10.1016/j.jphotobiol.2023.112781. (https://www.sciencedirect.com/science/article/pii/S1011134423001355)

[22] Akikazu Sakudo, Yoshikazu Suganuma, Rina Sakima, Kazuyoshi Ikuta, Diagnosis of HIV-1 infection by near-infrared spectroscopy: Analysis using molecular clones of various HIV-1 subtypes, Clinica Chimica Acta, Volume 413, Issues 3–4, 2012, Pages 467-472, ISSN 0009-8981. https://doi.org/10.1016/j.cca.2011.10.035. (https://www.sciencedirect.com/science/article/pii/S0009898111005985)

[23] Isabelle Ferreira, Juliana Ferreira-Strixino, Maiara L. Castilho, Claudia B.L. Campos, Claudio Tellez, Leandro Raniero, Characterization of Paracoccidioides brasiliensis by FT-IR spectroscopy and nanotechnology, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Volume 152, 2016, Pages 397-403, ISSN 1386-1425, https://doi.org/10.1016/j.saa.2015.07.061. (https://www.sciencedirect.com/science/article/pii/S1386142515301104)

[24] Jingrui Dou, Wubulitalifu Dawuti, Xiangxiang Zheng, Yousen Zhu, Renyong Lin, Guodong Lü, Yujiang Zhang, Rapid discrimination of Brucellosis in sheep using serum Fourier transform infrared spectroscopy combined with PCA-LDA algorithm, Photodiagnosis and Photodynamic Therapy, Volume 42, 2023, 103567, ISSN 1572-1000, https://doi.org/10.1016/j.pdpdt.2023.103567. (https://www.sciencedirect.com/science/article/pii/S1572100023002946)

[25] STUART, Barbara, Infrared Spectroscopy : Fundamentals and Applications. Chichester, West Sussex, England ; Hoboken, NJ :J. Wiley, 2004.

[26] Haykin, S. Redes Neurais: Uma Fundação Abrangente. New Jersey: Salão Prentice, 1999

[27] HAIR, Joseph F., Analise Multivariada de Dados. 6th Edition. São Paulo: Bookman Companhia Editora, 2009.

[28] Gabriela Pacher, Thiago Franca, Miller Lacerda, Natália O. Alves, Eliane M. Piranda, Carla Arruda, and Cícero Cena, Diagnosis of Cutaneous Leishmaniasis Using FTIR Spectroscopy and Machine Learning: An Animal Model Study ACS Infectious Diseases, Article ASAP, DOI: 10.1021/acsinfecdis.3c00430