



DETECÇÃO DE MASTITE EM LEITE BOVINO POR ESPECTROSCOPIA FTIR E APRENDIZADO DE MÁQUINA UTILIZANDO CONTAGEM DE CÉLULAS SOMÁTICAS

ROMICY DERMONDES SOUZA; ROSANGELA SILVEIRA BARBOSA; SÉRGIO AUGUSTO DE ALBUQUERQUE FERNANDES

RESUMO

A qualidade do leite é um fator crítico para a saúde pública e para a indústria de laticínios. A Contagem de Células Somáticas (CCS) é um indicador-chave da saúde do úbere e da qualidade do leite, com valores elevados estando associados à mastite. Técnicas espectroscópicas, como a Espectroscopia no Infravermelho com Transformada de Fourier (FTIR), combinadas com algoritmos de aprendizado de máquina, surgem como alternativas promissoras para uma detecção mais rápida e confiável. Este estudo teve como objetivo classificar amostras de leite com base na CCS (acima ou abaixo de 500×10^3 células/mL) utilizando espectroscopia FTIR combinada com o algoritmo de aprendizado de máquina Random Forest. Foram analisadas 658 amostras de leite de vacas Jersey. A CCS foi determinada por citometria de fluxo, e os espectros FTIR foram coletados na região do infravermelho médio. O algoritmo Boruta foi utilizado para seleção das bandas espectrais mais discriminativas. O conjunto de dados foi dividido em treinamento (70%) e teste (30%). O modelo de classificação Random Forest foi configurado com 500 árvores, e seu desempenho foi avaliado por meio de matriz de confusão e métricas derivadas. O algoritmo Boruta selecionou 29 números de onda discriminativos, localizados principalmente nas regiões associadas a lipídios ($1724\text{--}1774\text{ cm}^{-1}$) e lactose ($1018\text{--}1126\text{ cm}^{-1}$). O modelo Random Forest alcançou uma acurácia balanceada de 0,797. A especificidade foi alta (0,908), indicando boa capacidade de identificar amostras com baixa CCS. No entanto, a sensibilidade foi moderada (0,623), revelando uma dificuldade em detectar todas as amostras com CCS elevada. A precisão e o F1-Score foram de 0,814 e 0,706, respectivamente. A combinação FTIR-Random Forest mostrou-se um método viável e não destrutivo para a triagem de leite com base na CCS, com forte desempenho na identificação de amostras de baixa CCS. A baixa sensibilidade, contudo, aponta a necessidade de otimizações futuras, como o balanceamento de dados e ajuste de parâmetros, para melhorar a detecção de mastite subclínica.

Palavras-chave: Aprendizado de Máquina; Saúde Animal; Monitoramento de Qualidade.

1 INTRODUÇÃO

A qualidade do leite é fundamental tanto para a saúde pública quanto para a viabilidade econômica da produção de laticínios, pois afeta diretamente o rendimento industrial, a aceitação do produto pelo consumidor e a segurança alimentar (Pegolo et al., 2021). A contagem de

células somáticas (CCS) é amplamente reconhecida como um dos principais indicadores de saúde do úbere e qualidade do leite. Um aumento na CCS geralmente reflete a presença de mastite (inflamação da glândula mamária), levando a alterações negativas na composição do leite, como redução de lactose, proteína e caseína, além de piora nas propriedades tecnológicas e menor rendimento na produção de derivados lácteos (Podhorecká et al., 2021).

Atualmente, a detecção de mastite em vacas é realizada principalmente por métodos indiretos e diretos. Os métodos mais comuns incluem a Contagem de Células Somáticas (CCS) no leite, que é amplamente utilizada como indicador de inflamação do úbere, e o Teste de Mastite da Califórnia (CMT), que é rápido, barato e fácil de aplicar na fazenda (Stanek; Żółkiewski; Januś, 2024). As principais limitações dos métodos tradicionais são a baixa sensibilidade para detecção de mastite subclínica, a possibilidade de resultados falso-positivos ou falso-negativos (Tommasoni et al., 2023). Portanto, métodos mais rápidos e confiáveis, como métodos espectrométricos, como a espectroscopia no infravermelho com transformada de Fourier (FTIR) associado a métodos de aprendizado de máquina, podem ser uma alternativa viável para a detecção de mastite de vacas.

O FTIR é uma técnica rápida e de baixo custo que permite analisar a composição química do leite, detectando alterações Como adulterações, variações nos teores de lactose, minerais, proteínas e outros componentes (CHU et al., 2024). Quando métodos de aprendizado de máquina, como o random forest, são combinados ao FTIR, a análise dos dados espectrais se torna ainda mais precisa e robusta. O random forest consegue classificar e quantificar adulterações, identificar padrões complexos e melhorar a acurácia da detecção de alterações no leite, superando métodos estatísticos tradicionais (Bisutti et al., 2024; Duliński et al., 2025). Assim, a integração de FTIR com aprendizado de máquina oferece uma solução eficiente, rápida e confiável para monitorar a qualidade e autenticidade do leite em diferentes contextos industriais.

Métodos de aprendizado de máquina tem sido amplamente empregados para análise de qualidade do leite. Mazer (2021) empregou redes neurais para detectar adulterantes no leite, relatando 87,9% de acurácia com o modelo Patternnet e 76,4% com o perceptron multicamadas. Rezende et al. (2022) descreveram, de forma qualitativa, alta precisão na detecção de mastite subclínica a partir de imagens termográficas. Apesar disso, nenhum estudos avaliou a classificação por random forest ou mensurou a identificação de amostras com elevada contagem de células somáticas (CCS). Portanto, o presente estudo tem como objetivo a classificação de amostras de leite com elevada contagem de células somáticas (CCS) utilizando dados de espectroscopia no infravermelho por transformada de Fourier (FTIR) e o algoritmo de aprendizado de máquina Random forest.

2 MATERIAL E MÉTODOS

2.1 Coleta a análise das amostras

As amostras de leite foram coletadas na Estação Experimental Terras Baixas (ETB/Embrapa Clima Temperado), em Capão do Leão (RS), a partir de vacas da raça Jersey em lactação, onde foram coletadas ao total 658 amostras de leite. A coleta seguiu o procedimento de higienização dos tetos com solução de iodo, secagem com papel toalha descartável e descarte dos três primeiros jatos. O leite foi acondicionado em frascos estéreis, mantido entre 0–10 °C em caixas com gelo reciclável e transportado ao Laboratório de Análise do Leite (LabLeite) em até 24 h. No laboratório, as amostras foram armazenadas a 0–10 °C e posteriormente analisadas quanto à contagem de células somáticas (CCS) pelo equipamento SomaScope pelo método citométrico em fluxo segundo ISO 13366-2. Os dados também foram analisados por espectroscopia no FTIR em espectrômetro LactoScope FTIR, operando na

região do infravermelho médio ($1000\text{--}3000\text{ cm}^{-1}$), com resolução de 4 cm^{-1} . Os espectros foram registrados no modo de absorbância. Os dados espectrais obtidos foram pre-tratados por correção do baseline usando o método dos quadrados mínimos assimétricos e suavização de Savitzky-Golay antes das análises.

22 Seleção e análise dos espectros

Para reduzir a dimensionalidade e identificar os números de onda mais informativos para a discriminação entre os grupos, empregou-se o algoritmo Boruta, um método baseado em Random forest que compara a importância de cada variável original com a importância de *shadow features*. Antes da execução do modelo, os dados foram processados por meio de normalização centrada na média e escalonamento min-max. O algoritmo foi executado com as seguintes configurações: número máximo de iterações igual a 100, e critério de decisão baseado na significância estatística ao nível de 0,01. Variáveis classificadas como “confirmadas” foram mantidas para as análises de classificação *random forest*. Para a realização da classificação *random forest*, as amostras foram divididas em dois grupos, “Acima de 500” com $n = 430$ e “Abaixo de 500” com $n = 228$, onde “500” representa 500×10^3 células somáticas/mL, limite estabelecido pela Instrução Normativa nº 76/2018 do MAPA (MAPA, 2018). Para as análises, o conjunto de dados foi dividido aleatoriamente em 70% para treinamento e 30% para teste. O modelo foi configurado com 500 árvores ($\text{ntree} = 500$). As métricas foram calculadas a partir da matriz de confusão obtida. Todas as análises foram conduzidas na linguagem R (versão 4.4.3), utilizando os pacotes *boruta* para seleção de variáveis, *randomForest* para modelagem e *readr* para leitura dos dados.

3 RESULTADOS E DISCUSSÃO

A figura 1 mostra os espectros de FTIR das amostras de leite, tanto abaixo como acima de 500×10^3 células somáticas/mL, onde é notada pequenas diferenças entre as absorbâncias, mas os principais picos e características espectrais dos espectros de leite. Os picos em 2924, 2854 e 1747 cm^{-1} são característicos de cadeias de ácidos graxos e grupos éster dos triglicerídeos, associados aos lipídios (Karaman et al., 2021; Toledo-Alvarado et al., 2021). Picos em 1663 (Amida I), 1551 (Amida II) e 1242 cm^{-1} (Amida III) refletem a estrutura secundária das proteínas do leite, como caseínas e proteínas do soro (Cirak; Icyer; Durak, 2018; Conceição et al., 2018). Picos em 1157, 1076 e 1041 cm^{-1} estão associados à lactose (C–O) e fosfatos (P–O), importantes para a composição mineral do leite (Conceição et al., 2018; Toledo-Alvarado et al., 2021).

Baseado no espectros coletados, foi realizado as análise do boruta para seleção dos números de onda mais discriminantes, onde os selecionados foram: 2615, 1774, 1771, 1767, 1763, 1724, 1126, 1099, 1095, 1092, 1088, 1084, 1080, 1076, 1072, 1068, 1065, 1061, 1057, 1053, 1049, 1045, 1041, 1038, 1034, 1030, 1026, 1022, 1018 cm^{-1} . Os números de onda na faixa de $1724\text{--}1774\text{ cm}^{-1}$ é uma região típica do estiramento C=O de ésteres, indicando a presença de triglicerídeos, principais componentes dos lipídios do leite (Karaman et al., 2021; Toledo-Alvarado et al., 2021). A região $1126\text{--}1018\text{ cm}^{-1}$ está associado ao estiramento C–O e C–C, característico da lactose (Conceição et al., 2018; Toledo-Alvarado et al., 2021). Já número de onda selecionado em 2615 cm^{-1} , é incomum para espectros de leite, mas pode estar relacionado

a sobreposição ou bandas combinadas. Resultados similares foram reportados por Freitas et al. (2021), que também observaram bandas representativas de lipídios e lactose na análise espectral pelo boruta.

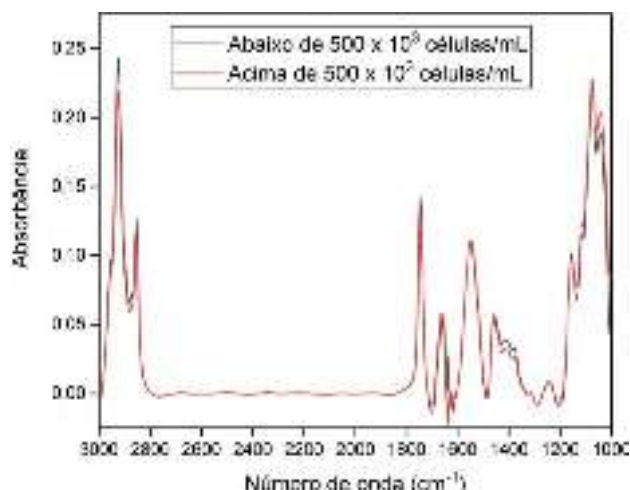


Figura 1. Espectros no FTIR das amostras de leite bovino

A partir das variáveis selecionados pelo boruta, foi conduzida uma análise de classificação *random forest* cujos resultados da matriz de confusão são mostrados na Figura 2. As amostras foram separadas em dois grupos, as amostras que continha um CCS acima de 500×10^3 células somáticas/mL, denominada “Acima de 500”, e as amostras que tinha CCS abaixo de 500×10^3 células somáticas/mL, denominada “Abaixo de 500”. Os resultados obtidos através da matriz de confusão demonstraram que, de um total de 197 amostras analisadas, 157 (79,7%) foram classificadas corretamente pelo modelo. Para amostras com CCS inferior a 500×10^3 células/mL, o modelo apresentou elevada precisão, classificando corretamente 109 amostras (55,3% do total), enquanto apenas 11 amostras (5,6%) foram erroneamente classificadas como pertencentes ao grupo com CCS superior a 500×10^3 células/mL. Em contrapartida, para amostras com CCS superior ao limiar estabelecido, o modelo identificou corretamente 48 amostras (24,4%), porém falhou em reconhecer 29 amostras (14,7%) que pertenciam a este grupo.

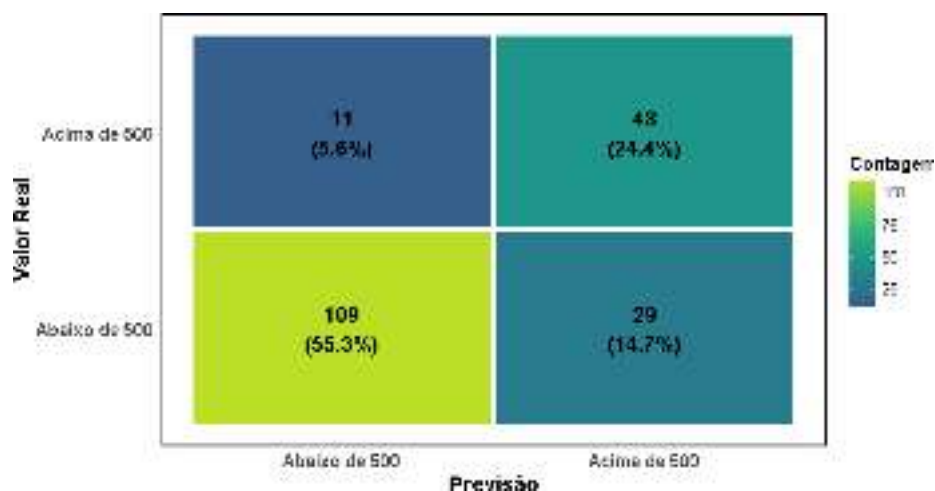


Figura 2. Matriz de confusão resultante da aplicação do modelo Random Forest utilizando os números de onda selecionados pelo algoritmo Boruta

As métricas de desempenho do método random forest calculadas a partir da matriz de confusão são mostradas na Tabela 1. A alta especificidade (0,908) demonstra que o modelo é particularmente eficaz em identificar corretamente amostras com baixa CCS, o que é relevante para processos de controle de qualidade onde falsos positivos podem resultar em descarte desnecessário do leite (Fusco et al., 2020). Os resultados são consistentes com dados encontradas na literatura para métodos de aprendizado de máquina que utilizaram sistemas automatizados de ordenha ou registros de rotina da fazenda alcançam especificidade acima de 0,9 (Fan et al., 2023; Pakrashi et al., 2023). Métodos de redes neurais relatam especificidades na faixa de 80–85%, com alguma queda na validação externa devido à variabilidade específica da fazenda (Naqvi et al., 2022). Apesar dos bons resultado de especificidade, o valor baixo de sensibilidade (0,623) representa uma limitação importante do modelo atual, pois indica que aproximadamente 38% das amostras com CCS elevada não são identificadas pelo sistema. Em contextos de monitoramento de saúde animal e qualidade do leite, falsos negativos podem ter implicações significativas, potencialmente permitindo que amostras de baixa qualidade passem despercebidas pelo sistema de controle (Chicco; Jurman, 2023). Esta limitação sugere a necessidade de refinamentos adicionais no modelo, possivelmente através da incorporação de características espectrais adicionais ou otimização dos parâmetros do algoritmo.

A precisão de 0,814 indica que, quando o modelo classifica uma amostra como tendo CCS elevada, esta predição é confiável em aproximadamente 81% dos casos. Este resultado é promissor para aplicações práticas na indústria láctea, onde a identificação precisa de amostras com alta CCS é fundamental para o controle de qualidade e para a detecção precoce de possíveis quadros de mastite subclínica em rebanhos leiteiros. O F1-Score de 0,706, representando a média harmônica entre precisão e recall, confirma que o modelo apresenta desempenho satisfatório, mas com espaço para aprimoramentos, especialmente no que diz respeito à sensibilidade. A acurácia balanceada de 0,797 indica que, apesar do desbalanceamento observado entre as classes, o modelo mantém bom desempenho geral. Os resultados encontrados são inferiores aos métodos apresentados. Estudos anteriores utilizando deep learning e imagens térmicas, por exemplo, relatam F1-Scores entre 0,87 e 0,91, acurácia geral entre 0,86 e 0,89, e acurácia balanceada próxima de 0,86. Modelos de machine learning supervisionados, como SVM e ensemble, também apresentam F1-Scores médios entre 0,76 e 0,78, com acurácia variando de 0,86 a 0,89. Os resultados apresentados no presente estudo são menores que os apresentados pela literatura, apesar disso, a combinação de FTIR e aprendizado de máquina para a detecção de mastite em vacas leiteiras se apresenta como método promissor.

Tabela 1. Métricas de desempenho calculado a partir da matriz de confusão obtido pelo random forest.

Métrica	Valor
Acurácia balanceada	0,797
Precisão	0,814
Sensibilidade	0,623
Especificidade	0,908
F1-Score	0,706

4 CONCLUSÃO

A espectroscopia FTIR associada ao algoritmo random forest é uma ferramenta viável para a classificação de amostras de leite com base na contagem de células somáticas. O modelo desenvolvido mostrou-se robusto para identificar amostras com baixa CCS, sendo vantajoso

para evitar falsos positivos e o descarte desnecessário de leite. No entanto, a baixa sensibilidade indica uma limitação na detecção de amostras com CCS elevada, o que pode permitir que amostras de baixa qualidade passem despercebidas. Para trabalhos futuros, recomenda-se a otimização dos parâmetros do algoritmo, o balanceamento do conjunto de dados e a incorporação de variáveis adicionais para aumentar a sensibilidade do modelo. Apesar dos resultados serem inferiores a alguns métodos da literatura, a combinação FTIR-random forest se mostrou um método promissor, rápido e não destrutivo para a triagem da qualidade do leite.

REFERÊNCIAS

- BISUTTI, V.; MOTA, L. F. M.; GIANNUZZI, D.; TOSCANO, A.; AMALFITANO, N.; SCHIAVON, S.; PEGOLO, S.; CECCHINATO, A. Infrared spectroscopy coupled with machine learning algorithms for predicting the detailed milk mineral profile in dairy cattle. **Food Chemistry**, v. 461, p. 140800, 2024.
- CHICCO, D.; JURMAN, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. **BioData Mining**, v. 16, n. 1, p. 4, 2023.
- CHU, C. et al. Rapid detection and quantification of melamine, urea, sucrose, water, and milk powder adulteration in pasteurized milk using Fourier transform infrared (FTIR) spectroscopy coupled with modern statistical machine learning algorithms. **Heliyon**, v. 10, n. 12, p. e32720, 2024.
- CIRAK, O.; ICYER, N. C.; DURAK, M. Z. Rapid detection of adulteration of milks from different species using Fourier Transform Infrared Spectroscopy (FTIR). **Journal of Dairy Research**, v. 85, n. 2, p. 222–225, 2018.
- CONCEIÇÃO, D.; GONÇALVES, BEN-HUR; DA HORA, F.; FALEIRO, A.; SANTOS, L.; FERRÃO, S. Use of FTIR-ATR Spectroscopy Combined with Multivariate Analysis as a Screening Tool to Identify Adulterants in Raw Milk. **Journal of the Brazilian Chemical Society**, 2018.
- DULIŃSKI, R.; GANCARZ, M.; SHAKHOVSKA, N.; BYCZYŃSKI, L.. The Application of Fourier Transform Infrared Spectroscopy and Chemometrics in Identifying Signatures for Sheep's Milk Authentication. **Processes**, v. 13, n. 2, p. 518, 2025.
- FAN, X.; WATTERS, R. D.; NYDAM, D. V.; VIRKLER, P. D.; WIELAND, M.; REED, K. F. Multivariable time series classification for clinical mastitis detection and prediction in automated milking systems. **Journal of Dairy Science**, v. 106, n. 5, p. 3448–3464, 2023. ISSN: 00220302.
- FREITAS, A. G. M.; MINHO, L. A. C.; MAGALHÃES, B. E. A.; SANTOS, W. N. L.; SANTOS, L. S.; FERNANDES, S. A. A. Infrared spectroscopy combined with random forest to determine tylosin residues in powdered milk. **Food Chemistry**, v. 365, p. 130477, 2021.
- FUSCO, V.; CHIEFFI, D.; FANELLI, F.; LOGRIECO, A. F.; CHO, G.-S.; KABISCH, J.; BÖHNLEIN, C.; FRANZ, C. M. A. P. Microbial quality and safety of milk and milk products in the 21st century. **Comprehensive Reviews in Food Science and Food Safety**, v. 19, n. 4, p. 2013–2049, 2020.

KARAMAN, A. D.; AYKAS, D. P.; RODRIGUEZ-SAONA, R.; ALVAREZ, V. Determination of the Compositional Differences in Organic and Conventional Milk using FTIR Spectroscopy. **Turkish Journal of Agriculture - Food Science and Technology**, v. 9, n. 3, p. 452–459, 2021.

MAPA. INSTRUÇÃO NORMATIVA Nº 76, DE 26 DE NOVEMBRO DE 2018. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750137/do1-2018-11-30-instrucao-normativa-n-76-de-26-de-novembro-de-2018-52749894 IN 76.

MAZER, R. H.; ULSON, J. A. C.; MEDEIROS, M. I. M. Identificação de adulterantes no leite bovino por meio de redes neurais artificiais e técnica cromática. **Revista Científica Multidisciplinar Núcleo do Conhecimento**, p. 53–68, 2021.

NAQVI, S. A.; KING, M. T. M.; MATSON, R. D.; DEVRIES, T. J.; DEARDON, R.; BARKEMA, H. W. Mastitis detection with recurrent neural networks in farms using automated milking systems. **Computers and Electronics in Agriculture**, v. 192, p. 106618, 2022.

PAKRASHI, A.; RYAN, C.; GUÉRET, C.; BERRY, D. P.; CORCORAN, M.; KEANE, M. T.; MAC NAMEE, B. Early detection of subclinical mastitis in lactating dairy cows using cow-level features. **Journal of Dairy Science**, v. 106, n. 7, p. 4978–4990, 2023.

PEGOLO, S. et al. Associations between differential somatic cell count and milk yield, quality, and technological characteristics in Holstein cows. **Journal of Dairy Science**, v. 104, n. 4, p. 4822–4836, 2021. ISSN: 00220302.

PODHORECKÁ, K.; BORKOVÁ, M.; ŠULC, M.; SEYDLOVÁ, R.; DRAGOUNOVÁ, H.; ŠVEJCAROVÁ, M.; PEROUTKOVÁ, J.; ELICH, O. Somatic Cell Count in Goat Milk: An Indirect Quality Indicator. **Foods**, v. 10, n. 5, p. 1046, 2021.

REZENDE, E. S. J.; MOURA, D. J.; PEREIRA, J. L. A. R.; FERRAZ, M. A. J.; FERRAZ, G. A. J.; JUNIOR, E. L. M.; COSTA, R. P. Modelagem, classificação e pré-validação de diagnóstico de mastite subclínica em função da temperatura do úbere. **Brazilian Journal of Development**, v. 8, n. 3, p. 15543–15552, 2022.

STANEK, P.; ŻÓLKIEWSKI, P.; JANUŚ, E.. A Review on Mastitis in Dairy Cows Research: Current Status and Future Perspectives. **Agriculture**, v. 14, n. 8, p. 1292, 2024.

TOLEDO-ALVARADO, H.; PÉREZ-CABAL, M. A.; TEMPELMAN, R. J.; ECCHINATO, A.; BITTANTE, G.; DE LOS CAMPOS, G.; VAZQUEZ, A. I. Association between days open and milk spectral data in dairy cows. **Journal of Dairy Science**, v. 104, n. 3, p. 3665–3675, 2021.

TOMMASONI, C.; FIORE, E.; LISUZZO, A.; GIANESELLA, M. Mastitis in Dairy Cattle: On-Farm Diagnostics and Future Perspectives. **Animals**, v. 13, n. 15, p. 2538, 2023.