



UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA – UESB
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA E CIÊNCIA DE ALIMENTOS
Área de Concentração: Ciências de Alimentos



**MÉTODOS QUIMIOMÉTRICOS ASSOCIADO AO INFRAVERMELHO MÉDIO
PARA DETECÇÃO E QUANTIFICAÇÃO DE CIPROFLOXACINA E CÉLULAS
SOMÁTICAS EM LEITE LÍQUIDO**

Autor: Rômicy Dermondes Souza

Orientador: Prof. Dr. Sérgio A. de A. Fernandes

ITAPETINGA

BAHIA – BRASIL

FEVEREIRO DE 2026

RÔMICY DERMONDES SOUZA

**MÉTODOS QUIMIOMÉTRICOS ASSOCIADO AO INFRAVERMELHO MÉDIO
PARA DETECÇÃO E QUANTIFICAÇÃO DE CIPROFLOXACINA E CÉLULAS
SOMÁTICAS EM LEITE LÍQUIDO**

Tese apresentada à Universidade Estadual
do Sudoeste da Bahia como parte
integrante das exigências do programa de
Pós-Graduação em Engenharia e Ciência
de Alimentos para obtenção do Título de
Doutor em Engenharia e Ciência de
Alimentos.

ITAPETINGA

BAHIA – BRASIL

FEVEREIRO DE 2026

637.127
S718m

Souza, Rômicy Dermondes.

Espectroscopia MIR associada a métodos quimiométricos para detecção de ciprofloxacina e contagem de células somáticas em leite *in natura*. /

Rômicy Dermondes Souza. – Itapetinga-BA: UESB, 2026.

98f.

Tese apresentada à Universidade Estadual do Sudoeste da Bahia como parte integrante das exigências do programa de Pós-Graduação em Engenharia e Ciência de Alimentos para obtenção do Título de Doutor em Engenharia e Ciência de Alimentos. Área de Concentração: Ciências de Alimentos. Sob a orientação do Prof. D. Sc. Sérgio A. de A. Fernandes.

1. Leite – Análise – Resíduo de antibiótico. 2. Leite – Espectroscopia. 3. Leite – Quimiometria. 4. Leite – Análise – Células somáticas. I. Universidade Estadual do Sudoeste da Bahia - Programa de Pós-Graduação de Doutorado em Engenharia e Ciências de Alimentos, *Campus* de Itapetinga. II. Fernandes, Sérgio A. de A. III. Título.

CDD (21): 637.127

Catálogo na Fonte:

Adalice Gustavo da Silva – CRB 535-5ª Região
Bibliotecária – UESB – Campus de Itapetinga-BA

Índice Sistemático para desdobramentos por Assunto:

1. Produtos lácteos – Controle de qualidade
2. Leite – Análise – Aprendizado de máquina

DECLARAÇÃO DE APROVAÇÃO

Título: “ESPECTROSCOPIA MIR ASSOCIADA A MÉTODOS QUIMIOMÉTRICOS PARA DETECÇÃO DE CIPROFLOXACINA E CONTAGEM DE CÉLULAS SOMÁTICAS EM LEITE IN NATURA.

”

Autor (a): ROMICY DERMONDES SOUZA

Orientador (a): Prof. Dr. Sérgio Augusto de Albuquerque Fernandes

Aprovado como parte das exigências para obtenção do Título de ***DOUTOR EM ENGENHARIA E CIÊNCIA DE ALIMENTOS, ÁREA DE CONCENTRAÇÃO: CIÊNCIA DE ALIMENTOS***, pela Banca Examinadora.

Documento assinado digitalmente
 **HILTON COSTA LOUZEIRO**
Data: 27/02/2026 11:32:53-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Hilton Costa Louzeiro
(UFMA)

Documento assinado digitalmente
 **MARIA ALVES FONTENELE**
Data: 27/02/2026 11:37:36-0300
Verifique em <https://validar.iti.gov.br>

Prof.^a Dr.^a Maria Alves Fontenele
(UFMA)

FABIANO
BARRETO:000579
86045
Assinado de forma digital por
FABIANO
BARRETO:00057986045
Dados: 2026.02.27 11:50:28
-03'00'

Prof. Dr. Fabiano Barreto
(MAPA)

Documento assinado digitalmente
 **WELLINGTON DA SILVA LYRA**
Data: 27/02/2026 11:55:08-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Wellington da Silva Lyra
(UFPB)

Documento assinado digitalmente
 **SERGIO AUGUSTO DE ALBUQUERQUE FERNANDES**
Data: 02/03/2026 12:11:49-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Sérgio Augusto de Albuquerque Fernandes
Orientador – UESB
Presidente da Banca

Itapetinga-BA, 27 de fevereiro de 2026.

AGRADECIMENTOS

Aos meus pais, Cícera de Jesus Dermondes Souza e Francisco Mizael Vieira de Souza, por todo o amor, dedicação e esforço ao longo da minha vida. Sou profundamente grato pelos ensinamentos, pelos valores transmitidos e pelo apoio incondicional, que foram fundamentais para que eu chegasse até aqui.

À minha companheira, Carmen Ibarra Arce, pelo amor, companheirismo, paciência e incentivo constante durante toda essa jornada. Obrigado por estar ao meu lado nos momentos mais desafiadores, pelos cafunés para aliviar o cansaço e por me dar forças para continuar. Sua companhia deixou todo esse esforço “estatisticamente” significativo.

Ao meu orientador Sergio Fernandes, pela confiança depositada em meu trabalho, pela disponibilidade e pelas valiosas contribuições acadêmicas, que foram essenciais para o desenvolvimento desta pesquisa.

Aos meus coorientadores, Rosangela Barbosa e Sergio Castro, pelo apoio, pelas sugestões pertinentes, pelo tempo dedicado e pelas importantes contribuições que enriqueceram este estudo.

À toda a equipe do NEAL pela ajuda prestada e troca de experiências.

A todas as pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho, seja por meio de apoio acadêmico, incentivo, amizade ou palavras de encorajamento, deixo aqui o meu sincero agradecimento.

Aos integrantes da banca examinadora pelas contribuições e disponibilidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelas contribuições financeiras ao longo desses quatro anos.

À Universidade Estadual do Sudoeste da Bahia, por ter me possibilitado desenvolver este trabalho.

Muito obrigado!

SUMÁRIO

LISTA DE ABREVIACÕES	VI
LISTA DE TABELAS	VIII
LISTA DE FIGURAS	IX
RESUMO.....	X
ABSTRACT.....	XI
I – REFERENCIAL TEÓRICO.....	1
1 INTRODUÇÃO.....	1
2 REVISÃO DE LITERATURA.....	3
2.1 <i>Produção e consumo de leite</i>	3
2.2 <i>Composição físico-química</i>	4
2.3 <i>Parâmetros de qualidade e segurança</i>	5
2.3.1 Células somáticas	6
2.3.2 Resíduos de antibióticos	7
2.3.3 Ciprofloxacina	8
2.4 <i>Espectroscopia no infravermelho médio</i>	10
2.5 <i>Quimiometria</i>	12
2.5.1 Conceitos fundamentais	12
2.5.2 Análise componentes principais	13
2.5.3 Mínimos quadrados parciais	15
2.6 <i>Aprendizado de máquina</i>	16
2.6.1 Random Forest	17
2.6.2 Algoritmo Boruta	18
3 REFERÊNCIAS.....	19
II – OBJETIVOS GERAIS	33
3 OBJETIVO GERAL.....	33
4 OBJETIVOS ESPECÍFICOS.....	33
III – CAPÍTULO I.....	34
ARTIGO 1 - DETECTION AND CLASSIFICATION OF CIPROFLOXACIN RESIDUES IN BOVINE MILK BY FTIR SPECTROSCOPY AND MULTIVARIATE ANALYSIS.....	34
ABSTRACT.....	36
INTRODUCTION.....	37
MATERIALS AND METHODS.....	38
Material collection	38
Experimental design	39
FTIR analysis	39
Selection of spectral bands	40

Predictive analytics	40
Classification analysis	41
RESULTS AND DISCUSSION.....	42
Milk composition	42
Spectra in the mid-infrared	42
Variable selection	44
Predictive models.....	46
Classification models	48
CONCLUSION.....	49
AUTHOR CONTRIBUTIONS.....	49
ACKNOWLEDGMENT.....	50
ABBREVIATIONS.....	50
REFERENCES.....	50
IV – CAPÍTULO II.....	58
ARTIGO 2 - EVALUATION OF LINEAR AND NONLINEAR MODELS FOR PREDICTION AND CLASSIFICATION OF SOMATIC CELL COUNT IN BOVINE MILK USING FTIR SPECTROSCOPY	58
ABSTRACT.....	61
INTRODUCTION.....	62
MATERIALS AND METHODS.....	63
Collection of samples	63
Analysis of samples	64
Selection of variables.....	65
Predictive analysis	66
Analysis of classification	66
RESULTS AND DISCUSSION.....	67
Composition of milk	67
Analysis of the spectra	69
Selection of variables.....	70
Predictive methods.....	72
Discriminatory methods.....	73
CONCLUSION.....	75
AUTHOR CONTRIBUTIONS.....	76
ACKNOWLEDGMENT.....	76
ABBREVIATIONS.....	77
REFERENCES.....	77

LISTA DE ABREVIACÕES

ANVISA	– Agência Nacional de Vigilância Sanitária
ATR	– Attenuated Total Reflectance (Reflexão Total Atenuada)
CAPES	– Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CCS	– Contagem de Células Somáticas
CV	– Coeficiente de Variação (Coefficient of Variation)
DRfA	– Dose de Referência Aguda
Embrapa	– Empresa Brasileira de Pesquisa Agropecuária
ETB	– Estação Experimental Terras Baixas
FAO	– Food and Agriculture Organization (Organização das Nações Unidas para Alimentação e Agricultura)
FTIR	– Fourier Transform Infrared Spectroscopy (Espectroscopia no Infravermelho por Transformada de Fourier)
HPLC	– High Performance Liquid Chromatography (Cromatografia Líquida de Alta Eficiência)
IBGE	– Instituto Brasileiro de Geografia e Estatística
IDA	– Ingestão Diária Aceitável
IFA	– Insumos Farmacêuticos Ativos
LDA	– Linear Discriminant Analysis (Análise Discriminante Linear)
LMR	– Limite Máximo de Resíduos (Maximum Residue Limit – MRL)
MAE	– Mean Absolute Error (Erro Absoluto Médio)
MAPA	– Ministério da Agricultura, Pecuária e Abastecimento
MIR	– Mid-Infrared (Infravermelho Médio)
MSC	– Multiplicative Scatter Correction (Correção Multiplicativa de Espalhamento)
mtry	– Número de variáveis candidatas por nó no Random Forest
NIR	– Near-Infrared Spectroscopy (Espectroscopia no Infravermelho Próximo)
OOB	– Out-of-Bag
PCA	– Principal Component Analysis (Análise de Componentes Principais)
PLS	– Partial Least Squares (Mínimos Quadrados Parciais)
PNCRC	– Plano Nacional de Controle de Resíduos e Contaminantes
RBQL	– Rede Brasileira de Qualidade do Leite
RF	– Random Forest (Floresta Aleatória)
RMSE	– Root Mean Square Error (Raiz do Erro Quadrático Médio)
R ²	– Coeficiente de Determinação
RPD	– Residual Prediction Deviation (Desvio de Predição Residual)

SCC – Somatic Cell Count (Contagem de Células Somáticas)

SD – Standard Deviation (Desvio Padrão)

SNV – Standard Normal Variate (Variável Normal Padrão)

SPE – Solid Phase Extraction (Extração em Fase Sólida)

UE – União Europeia

LISTA DE TABELAS

I – REFERENCIAL TEÓRICO

Tabela 1. Classes de antimicrobianos utilizadas na pecuária, seus usos principais e impactos predominantes	8
Tabela 2. Regiões espectrais do infravermelho e atribuições químicas no leite	12

III – CAPÍTULO I – ARTIGO 1 – Detection and classification of ciprofloxacin residues in bovine milk by FTIR spectroscopy and multivariate analysis

Table 1. Average composition of bovine milk used in the study.....	43
Table 2. Set of wave numbers selected by PCA and Boruta for milk samples containing different concentrations of ciprofloxacin	45
Table 3. Performance metrics of PLS and Random Forest regression models in the quantification of ciprofloxacin in liquid milk	49
Table 4. Performance metrics of LDA and RF models in the classification of milk samples containing ciprofloxacin.....	50

IV – CAPÍTULO II – ARTIGO 2 – Evaluation of linear and nonlinear models for prediction and classification of somatic cell count in bovine milk using FTIR spectroscopy

Table 1. Descriptive statistics of milk components	74
Table 2. Wave numbers selected by the Boruta and PCA approaches associated with somatic cell count.....	77
Table 3. Performance metrics of RF and PLS models for classification of milk samples based on SCC.....	81

LISTA DE FIGURAS

I – REFERENCIAL TEÓRICO

Figura 1. Estrutura química da ciprofloxacina	9
--	---

III – CAPÍTULO I – ARTIGO 1 – Detection and classification of ciprofloxacin residues in bovine milk by FTIR spectroscopy and multivariate analysis

Figure 1. Experimental design for fortifying milk samples with ciprofloxacin	40
Figure 2. Mid-infrared absorption spectrum of ciprofloxacin	44
Figure 3. Mid-infrared spectrum of raw milk with different concentrations of ciprofloxacin (0 – 250 µg/L)	45
Figure 4. Variance explained by principal component of milk samples impregnated with ciprofloxacin at different concentrations	46
Figure 5. Predictive performance of random forest regression and PLS for quantifying ciprofloxacin at concentrations of 0–250 µg/L in liquid milk	48

IV – CAPÍTULO II – ARTIGO 2 – Evaluation of linear and nonlinear models for prediction and classification of somatic cell count in bovine milk using FTIR spectroscopy

Figure 1. Flowchart of the experimental protocol for sample collection and laboratory analysis of milk samples	70
Figure 2. Relationship between the Somatic Cell Count (SCC) and milk components: fat, protein and lactose	73
Figure 3. Spectrum in the mean infrared spectrum of milk samples with SCC values greater and less than 500,000 cells/mL	75
Figure 4. Comparison of the predictive performance of RF and PLS in estimating the concentration of SCC in milk	79

RESUMO

Este trabalho teve como objetivo desenvolver e validar modelos baseados em espectroscopia no infravermelho médio (MIR), associados a métodos quimiométricos e de aprendizado de máquina, para o monitoramento de dois indicadores críticos: a Contagem de Células Somáticas (CCS) e a presença de resíduos do antibiótico ciprofloxacina. A metodologia envolveu a análise de 624 amostras de leite para CCS e amostras enriquecidas com ciprofloxacina (0 a 250 µg/L), utilizando a técnica de MIR. Foram avaliados métodos de seleção de variáveis, como a Análise de Componentes Principais (PCA) e o algoritmo Boruta, além de modelos de regressão (PLS e Random Forest - RF) e classificação (LDA e RF). Os resultados para a quantificação de CCS mostraram um desempenho limitado dos modelos de regressão, com coeficientes de determinação (R^2) de 0,293 para o PLS e 0,255 para o RF, indicando dificuldades na predição quantitativa direta devido à sobreposição de bandas espectrais dos constituintes majoritários do leite. No entanto, na análise qualitativa (classificação com limite de 500.000 células/mL), o modelo LDA apresentou sensibilidade superior (79,5%) em comparação ao RF (45,5%), sendo mais adequado para triagem inicial de mastite. Para a ciprofloxacina, a regressão também apresentou baixo desempenho (R^2 de 0,325 para RF), mas o modelo de classificação RF destacou-se com uma acurácia de 76% e F1-Score de 0,784 na distinção de amostras acima e abaixo do Limite Máximo de Resíduo de 100 µg/L. Conclui-se que, embora a predição quantitativa precisa de CCS e ciprofloxacina em leite líquido via FTIR ainda apresente desafios técnicos devido à complexidade da matriz, a abordagem qualitativa baseada em aprendizado de máquina demonstra viabilidade como ferramenta de triagem rápida, não destrutiva e de baixo custo. Tais modelos podem ser integrados à rotina laboratorial para a identificação preliminar de desconformidades, otimizando o controle de qualidade e reforçando a segurança alimentar no setor lácteo.

Palavras-chave: Leite; Mastite; Ciprofloxacina; FTIR; Quimiometria; Aprendizado de Máquina.

ABSTRACT

This study aimed to develop and validate models based on mid-infrared spectroscopy (MIR), associated with chemometric and machine learning methods, for monitoring two critical indicators: somatic cell count (SCC) and the presence of ciprofloxacin antibiotic residues. The methodology involved the analysis of 624 milk samples for SCC and samples enriched with ciprofloxacin (0 to 250 $\mu\text{g/L}$) using the MIR technique. Variable selection methods such as Principal Component Analysis (PCA) and the Boruta algorithm were evaluated, as well as regression (PLS and Random Forest - RF) and classification (LDA and RF) models. The results for SCC quantification showed limited performance of the regression models, with coefficients of determination (R^2) of 0.293 for PLS and 0.255 for RF, indicating difficulties in direct quantitative prediction due to the overlap of spectral bands of the major constituents of milk. However, in the qualitative analysis (classification with a limit of 500,000 cells/mL), the LDA model showed superior sensitivity (79.5%) compared to RF (45.5%), being more suitable for initial mastitis screening. For ciprofloxacin, regression also showed poor performance (R^2 of 0.325 for RF), but the RF classification model stood out with an accuracy of 76% and an F1-Score of 0.784 in distinguishing samples above and below the Maximum Residue Limit of 100 $\mu\text{g/L}$. It is concluded that, although the accurate quantitative prediction of CCS and ciprofloxacin in liquid milk via FTIR still presents technical challenges due to the complexity of the matrix, the qualitative approach based on machine learning demonstrates feasibility as a rapid, non-destructive, and low-cost screening tool. Such models can be integrated into laboratory routines for the preliminary identification of non-conformities, optimizing quality control and reinforcing food safety in the dairy sector.

Keywords: Milk; Mastitis; Ciprofloxacin; FTIR; Chemometrics; Machine Learning.

I – REFERENCIAL TEÓRICO

1 INTRODUÇÃO

O leite é um alimento de grande importância nutricional e econômica em todo o mundo, com uma composição bioquímica complexa e papel essencial na alimentação humana. Assegurar a qualidade e a segurança do leite é fundamental tanto para a proteção da saúde pública quanto para a sustentabilidade do setor lácteo. No Brasil, essa garantia é regulamentada pelas Instruções Normativas nº 76 e 77, de 2018, do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Essas normas estabelecem limites obrigatórios para indicadores de qualidade, como a Contagem de Células Somáticas (CCS) que reflete a saúde mamária dos animais, e proíbem a presença de resíduos de medicamentos veterinários, especialmente antibióticos, no produto final (MAPA, 2018b, 2018a).

A presença de resíduos de antibióticos no leite, mesmo em concentrações reduzidas, constitui um risco significativo à saúde pública. Esses resíduos podem causar reações alérgicas em pessoas sensíveis e, de forma mais preocupante, favorecer o desenvolvimento de resistência antimicrobiana, um dos principais desafios sanitários globais na atualidade (COSTA et al., 2024). Além disso, uma contagem elevada de células somáticas no leite é um indicador importante de mastite subclínica no rebanho, uma inflamação da glândula mamária que não apresenta sinais clínicos evidentes. Essa condição não apenas prejudica a saúde dos animais, mas também reduz a qualidade do leite, diminui o rendimento nos processos industriais de fabricação de derivados lácteos e encurta o tempo de conservação do produto final (STANEK; ŻÓŁKIEWSKI; JANUŚ, 2024).

Nesse contexto, a detecção rápida e confiável de contaminantes e indicadores de qualidade representa um desafio fundamental para a indústria láctea. Técnicas analíticas tradicionais, como a cromatografia líquida de alta eficiência (HPLC) para identificar antibióticos e a citometria de fluxo para contagem de células somáticas, que, apesar de precisas, apresentam limitações práticas, são processos trabalhosos, demandam tempo, alteram irreversivelmente a amostra e dependem do uso de reagentes químicos, o que eleva os custos operacionais e gera resíduos com impacto ambiental (DENOBILE; DE SOUZA NASCIMENTO, 2004). Essa limitação dos métodos tradicionais evidencia a necessidade de abordagens analíticas alternativas que sejam rápidas, não destrutivas e ambientalmente sustentáveis. A espectroscopia no infravermelho médio (MIR) destaca-se como uma alternativa promissora, pois permite caracterizar de forma abrangente a composição química do leite em poucos segundos, com preparação mínima ou nula da amostra e sem gerar resíduos químicos

(MENDES; DUARTE, 2021). Contudo, os espectros MIR são complexos, com alta dimensionalidade e sobreposição de bandas, o que torna a extração de informações específicas um desafio, que pode ser superado pela utilização de a quimiometria e o aprendizado de máquina, que fornecem métodos estatísticos e algoritmos capazes de processar os dados espectrais, reconhecer padrões sutis e desenvolver modelos preditivos robustos para quantificar ou classificar compostos de interesse no leite (KAROUI; DOWNEY; BLECKER, 2010).

Considerando a necessidade de métodos analíticos avançados para o monitoramento da qualidade do leite, esta tese tem como objetivo desenvolver e validar modelos baseados em espectroscopia no infravermelho médio (MIR), associados a métodos quimiométricos e de aprendizado de máquina, para monitorar dois indicadores críticos de qualidade do leite: a presença de resíduos de antibióticos e a contagem de células somáticas. O estudo compara diferentes abordagens de modelagem, desde técnicas clássicas de regressão multivariada, como os Mínimos Quadrados Parciais (PLS), até algoritmos mais recentes de aprendizado de máquina, como o Random Forest (RF), visando identificar as estratégias mais eficientes para a detecção e quantificação desses parâmetros de forma rápida, não destrutiva e sustentável. Além disso, serão avaliados métodos de seleção de variáveis, como o algoritmo Boruta, com o objetivo de aprimorar o desempenho dos modelos e identificar as regiões do espectro MIR mais informativas para a predição dos analitos de interesse. Espera-se que os resultados desta pesquisa contribuam para o desenvolvimento de ferramentas analíticas mais eficientes, rápidas e sustentáveis, com potencial para aplicação em larga escala no controle de qualidade do leite, reforçando assim a segurança alimentar e a proteção da saúde pública.

2 REVISÃO DE LITERATURA

2.1 Produção e consumo de leite

A produção de leite é uma das atividades mais importantes do setor agropecuário mundial, sendo essencial para garantir o acesso a alimentos nutritivos, além de contribuir para o desenvolvimento econômico e social de muitos países. Toda a cadeia do leite, que vai da ordenha nas fazendas até o consumidor final, forma um sistema complexo e interligado que envolve milhões de pessoas, produtores rurais, técnicos, trabalhadores da indústria, comerciantes, e representa uma parcela relevante da economia global. Em 2024, a produção mundial de leite chegou a 906 milhões de toneladas, o que representa um crescimento de 1,8% em comparação com 2023. Esse aumento foi impulsionado, principalmente, pelo maior consumo de leite e derivados em países da Ásia e da África (FAO, 2024).

A Índia, os Estados Unidos, a China, o Paquistão e o Brasil são os cinco maiores produtores de leite do mundo (FAO, 2024). O Brasil se destaca nesse cenário global, sendo o quarto maior produtor de leite do mundo e o principal produtor da América Latina. A produção nacional foi estimada em 35,8 bilhões de litros em 2024, com expectativa de chegar a 36,5 bilhões de litros em 2025, um crescimento de cerca de 2%. Essa atividade é fundamental para a economia do campo brasileiro, gerando renda e emprego, principalmente em Minas Gerais, Paraná e Rio Grande do Sul, os três estados que respondem pela maior parte da produção nacional (IBGE, 2025).

Quanto ao consumo, o consumo de leite no Brasil varia dependendo da idade, da região do país e da situação socioeconômica das pessoas. Entre adultos e idosos, cerca de 74 em cada 100 consomem leite, mas só 44 em cada 100 o fazem com frequência ou seja, pelo menos cinco vezes por semana. Esse hábito é mais comum entre mulheres, pessoas mais velhas e aquelas que têm uma alimentação de melhor qualidade (LUZ et al., 2021). Entre adolescentes, o consumo inadequado de leite é alto, onde quase 59 em cada 100 adolescentes não bebem leite todos os dias. Esse padrão é mais comum no Nordeste do Brasil e entre jovens cujas mães têm menor nível de escolaridade e que praticam menos atividade física (SILVA et al., 2020). Dados nacionais coletados entre 2008 e 2009 mostram que apenas 43% da população brasileira consumia produtos lácteos de forma regular. Entre esses produtos, o leite foi o mais consumido, presente na dieta de 21,1% das pessoas. A quantidade média diária de laticínios ingerida por pessoa foi de 142 gramas. Esse consumo foi mais frequente em áreas urbanas e entre dois grupos etários, jovens e idosos (MURPHY et al., 2016; POSSA et al., 2017).

Diante desta produção e consumo de leite e seus derivados no Brasil, a segurança alimentar e qualidade do leite tornaram-se aspectos cruciais que envolvem toda a cadeia produtiva, desde a ordenha até o consumidor final. Organizações internacionais, como a *Codex Alimentarius* e Comitê regulatório da UE, definem parâmetros que servem de referência para muitos países, incluindo limites para bactérias, contagem de células somáticas, resíduos químicos e práticas de higiene na produção e processamento do leite (FAO, 2004, 2023; UE, 2017). No Brasil, a legislação é regida principalmente pelas Instruções Normativas 76 e 77 do Ministério da Agricultura, Pecuária e Abastecimento (MAPA), que determinam padrões para contagem bacteriana total, SCC, teor de gordura, sólidos totais, lactose e proteína, além de critérios para rotulagem e rastreabilidade (MAPA, 2018a, 2018b). No contexto da exigência por qualidade e segurança, destaca-se a preocupação com a presença de resíduos de substâncias químicas, especialmente os antibióticos, amplamente empregados no setor leiteiro.

2.2 Composição físico-química

O leite de vaca é um fluido bioquímico complexo e altamente nutritivo, composto por centenas de substâncias, como proteínas, gorduras, carboidratos, minerais e vitaminas. Suas proteínas representam entre 2,6% e 5,3% da composição total e incluem frações importantes como a β -caseína, a κ -caseína e as proteínas do soro, entre elas, imunoglobulinas, α -lactoalbumina, β -lactoglobulina, lactoferrina e transferrina. Essas proteínas têm alto valor biológico porque contêm todos os aminoácidos essenciais. A gordura do leite corresponde a aproximadamente 4,5% de sua composição e é formada principalmente por ácidos graxos saturados, que podem representar até 65% desse total. Além desses ácidos, também estão presentes glicerídeos e lipídios complexos (RODRÍGUEZ-BERMÚDEZ et al., 2023; ZOU et al., 2024). O principal carboidrato é a lactose, fonte importante de energia e única fonte de galactose para o organismo humano. O leite contém minerais como cálcio, potássio, fósforo e magnésio em formas biologicamente disponíveis, ou seja, facilmente absorvidas pelo organismo, sendo essenciais para o crescimento e o desenvolvimento corporal. Além disso, fornece vitaminas lipossolúveis (A, D, E e K) e hidrossolúveis (do complexo B e vitamina C), o que contribui para seu alto valor nutricional (FOROUTAN et al., 2019; KUDRINA; KOGEVNIKOVA; KHUDYAKOVA, 2025).

A composição do leite pode variar conforme a raça bovina, por exemplo Holstein e Jersey, a região geográfica e o tipo de alimentação utilizada, como pastagem ou ração concentrada. Essas variações influenciam componentes específicos, como a concentração de proteínas bioativas, entre elas, a lactoferrina, e o perfil lipídico (ZOU et al., 2024). Fatores

genéticos também influenciam a composição do leite em termos de gordura, proteína e sólidos totais com herdabilidades moderadas a altas para esses traços. Além disso, variações diurnas na composição do leite foram observadas, com diferenças em metabólitos menores e componentes imunológicos entre o leite coletado durante o dia e à noite (TENG et al., 2021; WODAJO TIRFIE, 2023).

O leite bovino possui diversas propriedades físico-químicas que determinam sua qualidade e desempenho na indústria de alimentos. Entre elas estão a composição, bem como características como pH, densidade, viscosidade, tensão superficial e estabilidade térmica. Esses fatores influenciam diretamente o comportamento do leite durante os processos tecnológicos, como pasteurização, homogeneização e fermentação (CHEN et al., 2024; MCCARTHY, 2022). A estrutura das micelas de caseína é essencial para a coagulação do leite, etapa fundamental na produção de queijos. Por sua vez, as propriedades térmicas do leite definem sua resistência a tratamentos térmicos, como a pasteurização e a ultrafiltração (GAI et al., 2021; NUNES; TAVARES, 2019). Variantes genéticas das proteínas do leite afetam diretamente propriedades tecnológicas importantes, como a coagulação enzimática, a estabilidade ao calor e a capacidade de formação de gel. Essas características influenciam o rendimento e a qualidade dos produtos lácteos derivados, como queijos e iogurtes (GAI et al., 2021). Tecnologias emergentes, como a alta pressão hidrostática e o ultrassom, vêm sendo investigadas por sua capacidade de preservar os componentes bioativos do leite, ao mesmo tempo em que alteram suas propriedades funcionais, como solubilidade, capacidade emulsificante e formação de espuma (CHEN et al., 2024; NUNES; TAVARES, 2019). Além disso, fatores ambientais e práticas de manejo animal podem modificar as características físico-químicas do leite, afetando sua adequação para a produção de diferentes derivados lácteos (BERRUGA et al., 2021).

2.3 Parâmetros de qualidade e segurança

A qualidade e a segurança do leite dependem de uma combinação de parâmetros químicos, microbiológicos e físicos, que afetam tanto a saúde do consumidor quanto o valor comercial do produto. Entre os parâmetros químicos essenciais estão o teor de gordura, proteína, lactose e a ausência de substâncias inibitórias. A qualidade microbiológica é avaliada com base na contagem de bactérias deteriorantes, como *Pseudomonas* spp., patógenos como *Salmonella* e *Staphylococcus* e células somáticas, que servem como indicadores da saúde da glândula mamária do animal (BACIGALE et al., 2023; FUSCO et al., 2020). Preocupações emergentes incluem a presença de contaminantes químicos no leite, como micotoxinas,

resíduos de antibióticos, metais pesados e toxinas de origem vegetal, que podem representar riscos à saúde humana com o consumo prolongado (FUSCO et al., 2020; PUGA-TORRES et al., 2022). Como visto, são vastos os parâmetros associados à qualidade do leite, todavia, a presente tese focará apenas em realizar a fundamentação teórica de dois parâmetros, células somáticas e resíduos de antibióticos, mais especificamente a ciprofloxacina.

2.3.1 Células somáticas

As células somáticas presentes no leite constituem uma mistura de células epiteliais e células do sistema imune, como neutrófilos, macrófagos e linfócitos, liberadas naturalmente durante a ordenha. Essas células desempenham um papel essencial na defesa da glândula mamária contra infecções e na preservação da integridade do tecido mamário (ALHUSSIEN; DANG, 2018). A contagem de células somáticas (CCS) no leite é amplamente empregada como indicador da saúde da glândula mamária e da qualidade do leite, pois valores elevados geralmente sinalizam inflamação ou infecção, como a mastite (KASKOUS; FARSCHTSCHI; PFAFFL, 2022). A mastite pode se apresentar nas formas clínica ou subclínica. A forma subclínica é caracterizada por uma contagem elevada de CCS na ausência de sinais visíveis da doença, mas ainda assim resulta em redução da produção de leite e alterações em sua composição.

A alta CCS afeta negativamente a composição do leite ao reduzir os teores de caseína, lactose e sólidos totais, ao mesmo tempo que aumenta as proteínas do soro, os ácidos graxos livres e altera o equilíbrio mineral. Essas mudanças resultam em menor teor proteico e reduzem o rendimento industrial do leite (CINAR et al., 2015; ERMETIN; KUL; OKUYUCU, 2024). O leite com alto teor de SCC também apresenta aumento da atividade enzimática, causando lipólise e proteólise mais rápidas, o que deteriora a qualidade do leite durante o armazenamento, resultando em sabores desagradáveis como ranço e amargor que reduzem a vida útil (MA et al., 2000). Industrialmente, leite com alto SCC causa tempos de coagulação mais longos, redução na produção de queijo, tempos de batimento mais longos da manteiga, alteração da estabilidade térmica dos pós e, no geral, pior qualidade do produto (AULDIST, 2011). Portanto, manter um baixo SCC é fundamental para preservar a qualidade composicional do leite, a eficiência do processamento industrial e a vida útil do leite.

Os padrões brasileiros atuais para a CCS no leite são estabelecidos pelas Instruções Normativas nº 76 e nº 77 de 2018 (MAPA, 2018a, 2018b), que definem requisitos rigorosos e uniformes em todo o território nacional, com o objetivo de assegurar a saúde do rebanho e a qualidade do produto final. Para o leite cru refrigerado, o limite máximo permitido é de 500.000

células/mL, calculado como média geométrica trimestral. Esse valor deve ser respeitado por todos os produtores que entregam leite à indústria, independentemente da região do país. No caso do leite Tipo A, a legislação impõe um critério ainda mais restritivo, fixando o limite máximo de CCS em 400.000 células/mL. O descumprimento desses parâmetros por três meses consecutivos obriga o estabelecimento processador a suspender a coleta de leite na propriedade. Para retomar o fornecimento, o produtor deve implementar medidas de controle sanitário e ajustes no manejo da ordenha e do rebanho, conforme exigido pela legislação.

2.3.2 Resíduos de antibióticos

A presença de resíduos de antibióticos no leite representa um dos principais desafios para a segurança alimentar e a saúde pública em escala global, sendo consequência direta do uso frequente de medicamentos veterinários na produção leiteira. Esses resíduos são definidos como compostos farmacologicamente ativos, incluindo seus metabólitos e produtos de degradação, que permanecem no leite após a administração do fármaco ao animal. Estima-se que, anualmente, mais de 60 mil toneladas de antibióticos sejam utilizadas na pecuária global, com finalidades terapêuticas e profiláticas e, em alguns casos, como promotores de crescimento (SACHI et al., 2019). O uso indiscriminado desses fármacos, combinado ao não respeito do período de carência, leva à contaminação do leite que chega às indústrias e ao consumidor final (COSTA et al., 2024).

A ingestão de leite contaminado com resíduos de antibióticos representa um risco significativo à saúde humana. O principal impacto é a seleção de bactérias resistentes a esses medicamentos, um processo que contribui diretamente para a resistência antimicrobiana, um problema global que torna infecções comuns cada vez mais difíceis de tratar. Além disso, a exposição prolongada a baixas concentrações de antibióticos pode causar reações de hipersensibilidade e alergias agudas e também alterar o equilíbrio da microbiota intestinal humana (PIRES et al., 2024). Do ponto de vista industrial, a presença dessas substâncias afeta negativamente a produção de derivados lácteos fermentados, como queijos e iogurtes, pois os resíduos de antibióticos inibem o crescimento das culturas *starters*, levando a falhas no processo de acidificação e causando prejuízos econômicos ao setor de laticínios (ALENEZI et al., 2024). A Tabela 1 apresenta as principais classes de antibióticos detectadas em leite e seus respectivos impactos e contextos de uso.

Tabela 1. Classes de antimicrobianos utilizadas na pecuária, seus usos principais e impactos predominantes

Classe	Principal uso na pecuária	Impacto Predominante
Beta-lactâmicos	Tratamento de mastite e terapia de vaca seca	Reações alérgicas e inibição de fermentação
Tetraciclinas	Amplo espectro (terapêutico e profilático)	Seleção de resistência bacteriana
Sulfonamidas	Controle de infecções sistêmicas	Riscos toxicológicos e persistência
Quinolonas	Tratamento de infecções graves	Desenvolvimento de multirresistência

Informações extraídas de SACHI et al. (2019) e COSTA et al. (2024)

No Brasil, o controle de resíduos de antibióticos no leite é regulamentado pelo Ministério da Agricultura e Pecuária (MAPA) e pela Agência Nacional de Vigilância Sanitária (ANVISA). A Instrução Normativa nº 77, de 2018, determina que as indústrias de laticínios devem adotar programas de autocontrole, incluindo testes de triagem para detectar resíduos de antibióticos em cada tanque de recepção de leite, com ênfase nos grupos dos beta-lactâmicos e das tetraciclinas (MAPA, 2018a). Complementarmente, a ANVISA estabelece, por meio da Instrução Normativa nº 162, de 2022, os Limites Máximos de Resíduos (LMR) permitidos para cada princípio ativo. Esses limites servem como referência técnica para o Plano Nacional de Controle de Resíduos e Contaminantes (PNCRC) (BRASIL, 2017).

A eficácia do monitoramento desses contaminantes depende da aplicação de métodos analíticos precisos e sensíveis. Embora testes rápidos de triagem sejam amplamente utilizados nas plataformas de recepção de leite por sua baixa custo e rapidez, métodos confirmatórios e quantitativos, como a cromatografia líquida de alta eficiência (HPLC), são essenciais para assegurar a conformidade com os LMR estabelecidos, sendo estes métodos demorados e custosos (COSTA et al., 2024). Nesse contexto, a integração de técnicas espectroscópicas, como a espectroscopia no infravermelho médio combinada a métodos quimiométricas, surge como uma alternativa promissora para a detecção simultânea e rápida de múltiplos resíduos de antibióticos. Essa abordagem permite otimizar o controle de qualidade e assegurar a integridade do leite líquido destinado ao consumo humano.

2.3.3 Ciprofloxacina

A ciprofloxacina é uma fluoroquinolona amplamente utilizada devido ao seu amplo espectro de ação contra bactérias Gram-negativas e Gram-positivas. Seu mecanismo de ação consiste na inibição das enzimas DNA girase e topoisomerase IV, fundamentais para a replicação do material genético bacteriano (DOS SANTOS et al., 2026). Embora seu uso direto em medicina veterinária seja restrito em diversos países, a ciprofloxacina desempenha um papel relevante na segurança do leite, pois é o principal metabólito ativo da enrofloxacin, um antimicrobiano de uso exclusivamente veterinário (MAHMOOD et al., 2016). Em bovinos tratados com enrofloxacin, a biotransformação hepática leva à excreção de ciprofloxacina no leite, podendo essa substância corresponder a até 90% do resíduo total detectado. Essa característica torna a monitoração da ciprofloxacina essencial para avaliar a exposição do consumidor a resíduos antimicrobianos (COSTA et al., 2024). O Quadro 1 apresenta, de forma resumida, as principais propriedades físico-químicas e os parâmetros regulatórios da ciprofloxacina no contexto da produção leiteira e na Figura 1 é mostrado sua estrutura química.

Quadro 1. Características e parâmetros legislativos da ciprofloxacina

Parâmetro	Descrição
Classe Química	Fluoroquinolona
Origem no Leite	Metabolismo da Enrofloxacin (até 90% do resíduo)
Mecanismo de Ação	Inibição da DNA girase e Topoisomerase IV
LMR (Brasil)	100 µg/L (Soma de Enrofloxacin + Ciprofloxacina)
Impacto Principal	Resistência Antimicrobiana e Toxicidade Orgânica
Informações extraídas de ANVISA (2022), MAHMOOD et al. (2016) e COSTA et al. (2024)	

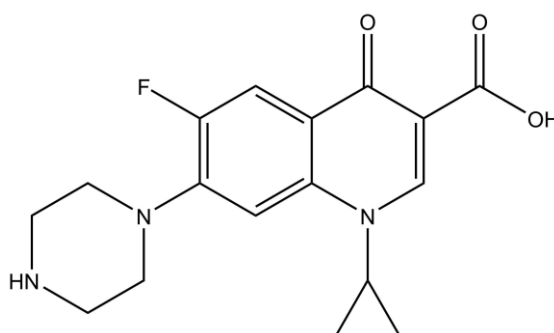


Figura 1. Estrutura química da ciprofloxacina Fonte: Autor

No âmbito regulatório brasileiro, a Agência Nacional de Vigilância Sanitária (ANVISA), por meio da Instrução Normativa nº 162, de 2022, determina que o LMR para fluoroquinolonas em leite deve ser expresso como a soma das concentrações de enrofloxacin

e sua principal metabólito ativa, a ciprofloxacina. O LMR estabelecido para leite bovino é de 100 µg/L, valor referente à soma das concentrações de enrofloxacin e ciprofloxacina (ANVISA, 2022).

A detecção eficaz de ciprofloxacina em matrizes complexas, como o leite, requer métodos analíticos que combinem sensibilidade e rapidez. A cromatografia, embora seja um método de referência para a detecção de ciprofloxacina, apresenta limitações relevantes. Entre elas destacam-se o alto custo operacional, associado ao consumo de solventes e colunas, o tempo prolongado de análise e a complexidade na preparação da amostra, que frequentemente requer etapas de extração em fase sólida (SPE). Essas etapas não apenas geram resíduos químicos, mas também podem introduzir erros analíticos. Adicionalmente, o método permanece suscetível a interferências da matriz do leite, que podem comprometer a precisão dos resultados (APARNATHI et al., 2023; DI ROCCO et al., 2020; WILKIE et al., 2025). Além disso, o método permanece suscetível a interferências matriciais que comprometem a precisão. Como alternativa promissora, a espectroscopia no infravermelho por transformada de Fourier (FTIR), combinada com ferramentas quimiométricas, tem se destacado por permitir uma análise rápida, de baixo custo e com preparação mínima da amostra. Essa abordagem emprega modelos estatísticos para identificar a assinatura espectral da ciprofloxacina mesmo em matrizes complexas, dispensando etapas de separação física do analito. Embora a sensibilidade do FTIR-quimiometria possa ser inferior à da LC-MS/MS na análise de traços, como demonstra alguns estudos (BIANCOLILLO et al., 2020; CEBI et al., 2023; NAZRI, 2025), sua eficiência, sustentabilidade e custo-benefício a qualificam como uma alternativa superior para monitoramento de rotina e triagem em larga escala, superando as principais limitações práticas dos métodos cromatográficos.

2.4 Espectroscopia no infravermelho médio

A espectroscopia no infravermelho médio (MIR) baseia-se na interação da radiação eletromagnética com a matéria na região espectral entre 4000 e 400 cm⁻¹, o que corresponde a comprimentos de onda de 2,5 a 25 µm (MENDES; DUARTE, 2021). O princípio físico dessa técnica reside na absorção de fótons cuja energia corresponde às transições entre níveis vibracionais de moléculas poliatômicas. Para que uma molécula absorva radiação no infravermelho médio, sua vibração deve provocar uma variação líquida no momento de dipolo elétrico. Essa condição torna a técnica particularmente sensível a grupos funcionais polares, comuns em matrizes biológicas complexas como o leite. Os modos vibracionais fundamentais

são classificados em estiramentos, que envolvem mudanças no comprimento das ligações, e deformações angulares, que alteram os ângulos entre as ligações químicas (LARKIN, 2018).

O espectro MIR é dividido em duas regiões principais: a região dos grupos funcionais ($4000\text{--}1500\text{ cm}^{-1}$) e a região da impressão digital ($1500\text{--}400\text{ cm}^{-1}$). Na primeira, as bandas de absorção correspondem a vibrações características de grupos químicos específicos, como as ligações O–H da água, C–H dos lipídios e as bandas Amida I e Amida II das proteínas (DE FREITAS et al., 2021; RACHAH et al., 2021). Já a região da impressão digital apresenta padrões espectrais complexos e únicos, originados de vibrações combinadas do esqueleto molecular, conferindo à técnica a capacidade de identificar e diferenciar compostos (MENDES; DUARTE, 2021). A quantificação de analitos baseia-se na Lei de Beer-Lambert, que estabelece uma relação linear entre a absorbância e a concentração da substância, fundamentando o emprego da espectroscopia para análises quantitativas em sistemas contendo multicomponentes, como o leite (LARKIN, 2018; LOIANNO, 2023).

Os equipamentos atuais para análise MIR utilizam predominantemente a espectroscopia FTIR. Diferentemente dos espectrômetros dispersivos tradicionais, que avaliam as frequências da radiação de forma sequencial, o FTIR emprega um interferômetro para registrar simultaneamente todas as frequências presentes na amostra. O sinal obtido contém a informação espectral completa e é convertido, por meio da transformada de Fourier, em um espectro de absorção interpretável, permitindo maior eficiência na coleta de dados e melhor relação sinal-ruído em comparação com métodos convencionais (LARKIN, 2018). Essa abordagem apresenta vantagens fundamentais, como a coleta simultânea de todas as frequências espectrais, que resulta em uma relação sinal-ruído superior, e a calibração altamente estável do sistema óptico, que garante precisão nas medições de frequência. Essas características possibilitam análises rápidas e com elevada reprodutibilidade (NANDIYANTO; RAGADHITA; FIANDINI, 2022). Na análise de líquidos como o leite, é comum utilizar o modo de reflexão total atenuada (ATR). Nesse método, a amostra entra em contato direto com um cristal de alto índice de refração. Esse procedimento dispensa etapas complexas de preparação da amostra, como diluições ou secagem, preservando a integridade da matriz original (MENDES; DUARTE, 2021).

Na análise de leite fluido, o MIR permite quantificar simultaneamente os constituintes fundamentais. O perfil espectral característico do leite é definido por três regiões de absorção principais, sendo elas a banda atribuída aos triacilgliceróis (vibração de estiramento do grupo carbonila C=O, próximo a 1745 cm^{-1}), a banda da Amida I das proteínas (aproximadamente 1650 cm^{-1}) e as vibrações dos grupos C-O da lactose (na faixa de $1200\text{ a }900\text{ cm}^{-1}$). Essas assinaturas espectrais constituem a base para a caracterização química rápida e não destrutiva

da matriz láctea (DULÍNSKI et al., 2025; KARAMAN et al., 2021; TSENKOVA et al., 2001). A Tabela 2 resume as principais regiões espectrais e as vibrações moleculares associadas aos constituintes do leite.

Tabela 2. Regiões espectrais do infravermelho e atribuições químicas no leite

Região Espectral	Tipo de Vibração	Atribuição Química
4000 – 3000 cm ⁻¹	Estiramento O-H e N-H	Água e Proteínas (Amida A)
3000 – 2800 cm ⁻¹	Estiramento C-H (CH ₂ e CH ₃)	Cadeias acíclicas de Lipídios
1750 – 1700 cm ⁻¹	Estiramento C=O (Éster)	Triacilgliceróis (Gordura)
1700 – 1600 cm ⁻¹	Estiramento C=O (Amida I)	Estrutura secundária de Proteínas
1550 – 1500 cm ⁻¹	Deformação N-H (Amida II)	Proteínas Totais
1200 – 900 cm ⁻¹	Estiramento C-O e C-C	Lactose e Carboidratos

Informações extraídas de DULÍNSKI et al. (2025), KARAMAN et al. (2021) e TSENKOVA et al. (2001)

Métodos para melhorar a efetividade do MIR para aplicações mais precisas tem sido estudados, sendo uma delas sua combinação com métodos quimiométricas de calibração multivariada. Essa abordagem permite distinguir bandas espectrais sobrepostas e quantificar compostos presentes em concentrações reduzidas, como resíduos de antibióticos e marcadores associados à CCS. Como resultado, a espectroscopia MIR converte-se em uma ferramenta analítica robusta, não destrutiva e de alto desempenho, capaz de fornecer resultados rápidos e confiáveis.

2.5 Quimiometria

2.5.1 Conceitos fundamentais

A quimiometria consolidou-se como uma disciplina fundamental na química analítica moderna, sendo definida como a aplicação de métodos matemáticos e estatísticos para extrair o máximo de informação relevante de sistemas químicos complexos (YAN, 2025). Em matrizes biológicas como o leite, os sinais analíticos de interesse costumam estar sobrepostos aos sinais provenientes dos constituintes majoritários, como água, gordura e proteínas. Nesses casos, a quimiometria possibilita a extração e resolução dessas informações por meio de modelos multivariados (MENDES; DUARTE, 2021). Ao contrário da análise univariada tradicional, que examina uma única variável de cada vez, os métodos quimiométricos avaliam conjuntamente todas as variáveis obtidas por técnicas espectroscópicas. Essa abordagem permite identificar

relações complexas entre as variáveis, sustentando tanto a análise exploratória quanto a calibração multivariada (DOS SANTOS et al., 2026; SAJI et al., 2024).

A etapa de pré-processamento dos dados espectrais constitui uma fase essencial para o desempenho adequado de modelos quimiométricos. Os espectros originais obtidos na região do infravermelho médio frequentemente apresentam distúrbios físicos e instrumentais que não refletem a composição química real da amostra. Entre esses distúrbios incluem-se o ruído eletrônico, alterações na trajetória da radiação no equipamento, deslocamentos na linha de base e fenômenos de espalhamento da luz. A correção dessas interferências por meio de pré-processamento adequado é necessária para realçar as informações químicas relevantes e garantir a robustez das análises subsequentes (DAYANANDA et al., 2024; YI et al., 2016). O objetivo principal do pré-processamento é eliminar ou reduzir essas variações indesejadas, isolando a variação química de interesse e aumentando a robustez e a capacidade preditiva dos modelos utilizados posteriormente (YAN, 2025).

Dentre as técnicas de pré-processamento mais disseminadas, destacam-se os métodos de correção de espalhamento, como a Variável Normal Padrão (SNV) e a Correção Multiplicativa de Espalhamento (MSC). O SNV atua normalizando cada espectro individualmente subtraindo sua média e dividindo pelo seu desvio padrão (MOKARI; GUO; BOCKLITZ, 2023). Complementarmente, o uso de derivadas espectrais, frequentemente Savitzky-Golay, desempenha um papel fundamental na correção de desvios de linha de base e na resolução de bandas espectrais sobrepostas. Ao realçar variações sutis na inclinação do espectro, as derivadas aumentam a resolução efetiva dos sinais, facilitando a distinção entre componentes químicos com absorções próximas. Associada a esse procedimento, a suavização é aplicada para atenuar o ruído de alta frequência, preservando as características espectrais autênticas e evitando que flutuações aleatórias interfiram no desempenho dos modelos quimiométricos (BUTLER et al., 2018; LEE; LIONG; JEMAIN, 2018), como análise dos componentes principais (PCA) e mínimos quadrados parciais (PLS)

2.5.2 Análise componentes principais

A PCA é um método estatístico multivariado frequentemente utilizado para reduzir a dimensionalidade de conjuntos de dados complexos, mantendo a maior parte da variabilidade presente nos dados originais. O método transforma um conjunto de variáveis que podem apresentar relações entre si em um novo conjunto de variáveis não correlacionadas, denominadas componentes principais. A PCA é amplamente aplicada na análise exploratória de dados, na identificação de padrões, na detecção de observações atípicas e na visualização de

estruturas subjacentes aos dados (SANTOS et al., 2019; WOLD; ESBENSEN; GELADI, 1987).

Do ponto de vista matemático, a PCA é realizada mediante a decomposição em valores singulares ou pelo cálculo dos valores e vetores próprios da matriz de covariância dos dados originais. Os valores próprios obtidos indicam a quantidade de variabilidade dos dados explicada por cada componente principal, enquanto os vetores próprios, também denominados *loadings*, estabelecem as direções no espaço das variáveis originais associadas a essas variabilidades máximas (MEGLEN, 1992). O primeiro componente principal retém a maior fração da variabilidade total dos dados; o segundo componente principal captura a segunda maior fração, e assim sucessivamente para os demais componentes (SANTOS et al., 2019).

Quando combinada à espectroscopia no infravermelho médio (MIR), o pré-processamento dos dados espectrais constitui uma etapa essencial antes da aplicação da PCA. Desvios na linha de base dos espectros, alterações decorrentes do espalhamento da radiação e variações introduzidas pelo equipamento de medição podem mascarar as informações químicas associadas à amostra de interesse (BUTLER et al., 2018; LEE; LIONG; JEMAIN, 2018). Técnicas como a correção de espalhamento multivariado e a variância normal padrão são frequentemente utilizadas para mitigar os efeitos de espalhamento. A normalização dos espectros é uma etapa fundamental que garante a contribuição equilibrada de todas as variáveis na análise. Sem esse procedimento, variáveis que apresentam valores mais altos poderiam influenciar desproporcionalmente os resultados da PCA, comprometendo a representatividade das demais informações espectrais. A normalização ajusta a escala das variáveis, permitindo que cada região do espectro exerça influência proporcional na identificação das estruturas latentes nos dados. (RINNAN; BERG; ENGELSEN, 2009).

A interpretação dos resultados da PCA fundamenta-se na análise dos gráficos de escores e de *loadings*. Os gráficos de escores apresentam a posição de cada amostra no espaço formado pelos componentes principais, possibilitando a identificação de padrões de agrupamento entre as amostras, a observação de tendências sistemáticas e a detecção de outliers. Amostras que apresentam posições próximas no gráfico de escores possuem perfis espectrais semelhantes (JOLLIFFE; CADIMA, 2016). Por outro lado, os gráficos de *loadings* expressam a contribuição de cada variável para a formação de cada componente principal. Regiões com valores elevados nos *loadings* correspondem às porções do espectro que exercem maior influência na diferenciação entre as amostras ou na explicação da variância dos dados, fornecendo informações sobre as alterações químicas (MEGLEN, 1992).

2.5.3 Mínimos quadrados parciais

O PLS é um método estatístico multivariado robusto, frequentemente empregado para modelar relações entre um conjunto de variáveis explicativas (matriz X) e um conjunto de variáveis resposta (matriz Y). Essa abordagem é especialmente útil quando as variáveis preditoras apresentam colinearidade ou quando o número de variáveis excede o número de observações, situações comuns em dados espectroscópicos e outras estruturas multivariadas complexas (WOLD; SJÖSTRÖM; ERIKSSON, 2001). Diferentemente da PCA, cujo objetivo é capturar a maior parte da variabilidade presente nas variáveis preditoras, o método PLS busca identificar componentes que maximizem a covariância entre as variáveis preditoras e as variáveis resposta. Essa característica torna o PLS especialmente adequado para situações em que se pretende desenvolver modelos preditivos ou investigar relações entre conjuntos de variáveis (DE JONG, 1993).

A fundamentação matemática do método PLS baseia-se na extração de componentes latentes tanto para o conjunto de variáveis preditoras (matriz X) quanto para o conjunto de variáveis resposta (matriz Y). Esses componentes são obtidos de modo a capturar a parcela da variabilidade presente em X que apresenta maior relação com Y (MATEOS-APARICIO, 2011).

A elaboração de modelos robustos e confiáveis pelo método de PLS requer pré-processamento dos dados espectrais. Nesse processo, aplicam-se técnicas como a correção do espalhamento da radiação e a suavização dos espectros, geralmente com o filtro de Savitzky-Golay. Esses procedimentos visam atenuar ruídos instrumentais, corrigir desvios na linha de base dos espectros e acentuar as regiões espectrais efetivamente associadas às propriedades químicas ou físicas de interesse, aumentando assim a qualidade preditiva do modelo (SIVAKESAVA; IRUDAYARAJ, 2002). A definição do número apropriado de variáveis latentes é fundamental para prevenir o ajuste excessivo do modelo, o denominado *overfitting*, situação em que este se adapta demasiadamente aos dados de treinamento e comprometendo sua capacidade de generalização. Para selecionar o número ótimo de variáveis latentes, utilizam-se métodos de validação cruzada como o *leave-one-out* ou a *k-fold*. Esses métodos permitem estimar o erro de predição do modelo em dados não utilizados no treinamento, possibilitando a escolha do número de variáveis latentes que minimiza o erro de validação e assegura um equilíbrio entre complexidade do modelo e capacidade preditiva (MEHMOOD et al., 2012).

A validação de modelos construídos pelo método PLS baseia-se em métricas estatísticas que avaliam sua capacidade preditiva e sua estabilidade. O erro médio quadrático de calibração (RMSEC) quantifica o ajuste do modelo aos dados utilizados em seu treinamento, enquanto o

erro médio quadrático de predição (RMSEP) mede o desempenho do modelo ao prever amostras independentes não incluídas na etapa de calibração. O coeficiente de determinação (R^2) expressa a proporção da variabilidade da resposta que é explicada pelo modelo. Para aplicações de quantificação, destaca-se o parâmetro RPD (*Residual Prediction Deviation*), definido como a razão entre o desvio padrão das amostras de referência e o RMSEP. Modelos com valores de RPD superiores a 3 são geralmente considerados adequados para estimativas quantitativas precisas, pois indicam que a variabilidade natural dos dados é suficientemente maior que o erro de predição do modelo (MATEOS-APARICIO, 2011; MEHMOOD et al., 2012).

2.6 Aprendizado de máquina

O aprendizado de máquina é uma subárea da inteligência artificial dedicada ao desenvolvimento de algoritmos que permitem aos sistemas computacionais aprimorar seu desempenho em determinadas tarefas com base na análise de dados, sem a necessidade de programação explícita para cada situação. Em vez de seguir instruções rígidas previamente definidas, esses algoritmos identificam padrões e regularidades nos dados de treinamento, utilizando essas informações para realizar inferências ou tomar decisões em novos conjuntos de dados (SHINDE; SHAH, 2018). No campo da quimiometria, o aprendizado de máquina tem se consolidado como uma ferramenta essencial para a extração de informações relevantes de conjuntos de dados complexos, como os obtidos por meio da espectroscopia MIR, que frequentemente apresentam alta dimensionalidade e colinearidade entre as variáveis (HOUHOU; BOCKLITZ, 2021).

O aprendizado de máquina pode ser classificado como aprendizado supervisionado e o aprendizado não supervisionado. No aprendizado supervisionado, o algoritmo é treinado com um conjunto de dados previamente rotulado, no qual cada observação de entrada está associada a um valor de referência conhecido, no caso a variável resposta. O objetivo é que o modelo aprenda a mapear as entradas para as saídas corretas, generalizando para novos dados não vistos. Dentro do aprendizado supervisionado, destacam-se duas categorias principais: a regressão, utilizada para prever valores contínuos, por exemplo, a concentração de um analito ou o teor de gordura no leite), e a classificação, empregada para atribuir amostras a categorias, por exemplo, classificar amostras de leite como "conforme" ou "não conforme" para a presença de um contaminante (AGIOMAVRITI et al., 2024).

Em contraste com o aprendizado supervisionado, o aprendizado não supervisionado opera com conjuntos de dados não rotulados, isto é, sem informações prévias sobre as categorias

ou valores de referência associados às observações. Seu objetivo é identificar padrões estruturais, agrupamentos naturais ou relações entre as variáveis com base exclusivamente nas características intrínsecas dos dados. As principais aplicações desse paradigma incluem o agrupamento ou *clustering*, que organiza as observações em grupos homogêneos segundo critérios de similaridade e a redução de dimensionalidade, que busca representar os dados em um espaço de menor dimensão preservando a informação mais relevante (AGIOMAVRITI et al., 2024; SHINDE; SHAH, 2018). A PCA, anteriormente discutida, constitui um exemplo clássico de técnica de redução de dimensionalidade não supervisionada.

A relevância do aprendizado de máquina na análise de alimentos, com ênfase no leite, tem aumentado significativamente. Sua capacidade de processar grandes volumes de dados espectrais e identificar relações complexas, inclusive não lineares, torna essa abordagem especialmente adequada para a detecção de fraudes, a identificação de contaminantes e a avaliação de parâmetros de qualidade (FREITAS et al., 2021). A combinação de modelos de aprendizado de máquina com o MIR possibilita o desenvolvimento de sistemas de controle de qualidade que são rápidos, não destrutivos e eficientes, podendo ser implementados em tempo real nas linhas de produção da indústria láctea (DULÍNSKI et al., 2025).

2.6.1 Random Forest

O Random Forest (RF) é um algoritmo de aprendizado de máquina versátil e amplamente utilizado. O RF utiliza a técnica de *bagging*, na qual múltiplas árvores de decisão são construídas de forma independente, cada uma a partir de uma amostra aleatória com reposição extraída do conjunto de dados de treinamento. A previsão final do modelo é obtida pela combinação dos resultados individuais de cada árvore por meio da maioria dos votos em classificação ou da média das previsões em regressão. Esse procedimento reduz a variabilidade das estimativas e aumenta a estabilidade do modelo, conferindo ao RF elevada robustez e capacidade de generalização para novos dados (BREIMAN, 2001; TOUW et al., 2013).

O funcionamento do RF fundamenta-se em dois mecanismos de aleatoriedade que contribuem para sua robustez, a amostragem bootstrap e a seleção aleatória de variáveis preditoras. No primeiro, cada árvore de decisão é construída a partir de uma amostra aleatória extraída com reposição do conjunto de dados original. Consequentemente, algumas observações podem aparecer repetidas vezes nessa amostra, enquanto outras não são incluídas, formando o conjunto *out-of-bag* (OOB). Esse conjunto OOB, composto pelas observações não utilizadas no treinamento de uma determinada árvore, pode ser empregado para estimar internamente o desempenho do modelo, sem a necessidade de uma partição separada de

validação (BREIMAN, 2001). Além disso, em cada nó de cada árvore, apenas um subconjunto aleatório das variáveis preditoras é considerado para a divisão, e não todas as variáveis. Esse parâmetro, conhecido como *mtry*, garante a diversidade das árvores e evita a correlação entre elas, garantindo que o erro de predição do RF seja menor do que o de uma única árvore (PARMAR; KATARIYA; PATEL, 2019).

As características do RF o tornam adequado para a análise de dados espectrais obtidos por MIR, onde é capaz de lidar de forma eficiente com conjuntos de dados de alta dimensionalidade, onde o número de variáveis é muito maior que o número de amostras, uma característica comum em espectroscopia (DE SANTANA; BORGES NETO; POPPI, 2019). Além disso, o RF é capaz de modelar relações não-lineares complexas entre os espectros e as propriedades químicas de interesse, e é bastante robusto a outliers e ruídos nos dados (DA SILVA; OLIVEIRA; VILLA, 2025). A arquitetura do Random Forest, baseada na agregação de múltiplos modelos, confere-lhe menor suscetibilidade ao *overfitting* quando comparado a árvores de decisão individuais. Essa característica resulta em modelos mais estáveis e com maior capacidade de generalização para novas observações.

2.6.2 Algoritmo Boruta

O Algoritmo Boruta é um método estatístico para seleção de variáveis com propriedades robustas, que se distingue por identificar todas as variáveis que apresentam associação com o composto estudado, em contraste com abordagens que retêm apenas um conjunto reduzido de variáveis suficientes para a modelagem preditiva. Ao contrário de métodos que ordenam as variáveis segundo critérios de relevância, como a importância de Gini empregada na técnica de RF, o Boruta emprega uma estratégia estatística rigorosa para separar variáveis importantes de ruído, tornando-o particularmente valioso em dados com alta dimensionalidade, como os encontrados em dados espectrais (KURSA; RUDNICKI, 2010; ZHOU; XIN; LI, 2023).

O Boruta é um método que utiliza diretamente o RF como base para seu funcionamento, atuando como uma camada adicional que organiza e interpreta os resultados do RF. Sua estratégia consiste em comparar a importância de cada variável original do conjunto de dados com a importância de variáveis artificiais geradas aleatoriamente, chamadas de variáveis sombras, para identificar quais variáveis carregam informação real e quais se comportam como ruído (IRANZAD; LIU, 2025). Para cada variável original do conjunto de dados, o Boruta gera uma versão modificada na qual os valores são reorganizados aleatoriamente, processo conhecido como embaralhamento. Essas versões modificadas são denominadas variáveis

sombra e servem como referência para distinguir variáveis informativas de ruído (MANIKANDAN et al., 2024).

O Boruta opera por meio de um processo repetitivo. Em cada ciclo, um modelo RF é treinado com todas as variáveis, tanto as originais quanto as sombra. Após o treinamento, calcula-se a importância de cada variável. Em seguida, uma variável original é classificada como relevante quando sua importância supera, de forma significativa, a maior importância observada entre as variáveis sombra naquela iteração. Esse procedimento se repete até que todas as variáveis sejam classificadas como relevantes ou irrelevantes, ou até um limite máximo de iterações ser atingido (KURSA; JANKOWSKI; RUDNICKI, 2010). Variáveis cuja importância permanece abaixo da média observada nas variáveis sombra são classificadas como irrelevantes e excluídas das iterações seguintes. Já as variáveis que não apresentam evidência estatística clara são mantidas em estado provisório, como “tentativa”, e reavaliadas nas iterações posteriores até que uma decisão definitiva possa ser tomada. O processo continua até que todas as variáveis sejam classificadas como relevantes ou irrelevantes, ou até que um número máximo de iterações seja atingido (KURSA; RUDNICKI, 2010).

As vantagens do algoritmo Boruta destacam-se especialmente em aplicações quimiométricas, áreas em que a seleção de variáveis é complexa devido ao grande número de preditores e à forte colinearidade, características comuns em dados espectrais. Diferentemente de métodos que selecionam apenas um subconjunto mínimo de variáveis para otimizar o desempenho preditivo, o Boruta identifica todas as variáveis informativas relevante para a previsão. Essa abordagem é fundamental para a interpretabilidade do modelo e facilitar a detecção de biomarcadores espectrais com significado químico ou biológico (KURSA, 2014). Sua natureza baseada em Random Forest o torna robusto a ruídos e outliers, e o uso de variáveis sombra fornece um critério estatístico rigoroso para a seleção, evitando a inclusão de variáveis irrelevantes que poderiam ser erroneamente consideradas importantes por outros métodos (KURSA; JANKOWSKI; RUDNICKI, 2010; KURSA; RUDNICKI, 2010).

3 REFERÊNCIAS

AGIOMAVRITI, Aikaterini-Artemis; NIKOLOPOULOU, Maria P.; BARTZANAS, Thomas; CHORIANOPOULOS, Nikos; DEMESTICHAS, Konstantinos; GELASAKIS, Athanasios I. Spectroscopy-Based Methods and Supervised Machine Learning Applications for Milk Chemical Analysis in Dairy Ruminants. **Chemosensors**, [S. l.], v. 12, n. 12, p. 263, 2024. ISSN: 2227-9040. DOI: 10.3390/chemosensors12120263. Disponível em: <https://www.mdpi.com/2227-9040/12/12/263>.

ALENEZI, Maha S.; TARTOR, Yasmine H.; EL-SHERBINI, Mohammed; PET, Elena; AHMADI, Mirela; ABDELKHALEK, Adel. Antibiotic Residues in Milk and Milk-Based Products Served in Kuwait Hospitals: Multi-Hazard Risk Assessment. **Antibiotics**, [S. l.], v. 13, n. 11, p. 1073, 2024. ISSN: 2079-6382. DOI: 10.3390/antibiotics13111073. Disponível em: <https://www.mdpi.com/2079-6382/13/11/1073>.

ALHUSSIEN, Mohammed Naif; DANG, Ajay Kumar. Milk somatic cells, factors influencing their release, future prospects, and practical utility in dairy animals: An overview. **Veterinary World**, [S. l.], v. 11, n. 5, p. 562–577, 2018. ISSN: 09728988. DOI: 10.14202/vetworld.2018.562-577. Disponível em: <http://www.veterinaryworld.org/Vol.11/May-2018/1.html>.

ANVISA. Instrução Normativa nº 162, de 1º de julho de 2022. Estabelece a ingestão diária aceitável (IDA), a dose de referência aguda (DRfA) e os limites máximos de resíduos (LMR) para insumos farmacêuticos ativos (IFA) de medicamentos veterinários em alimentos de Brasília, DF, Brasil, 2022. p. 146.

APARNATHI, KD; PATEL, Ajit; MEHTA, Bhavbhuti M.; PATEL, Deep; PRAJAPATI, JB. Gas chromatographic analysis of triglycerides - The reference method for testing purity of milk fat and perspectives on its use in India: Review. **Indian Journal of Dairy Science**, [S. l.], v. 76, n. 5, p. 409–426, 2023. ISSN: 00195146. DOI: 10.33785/IJDS.2023.v76i05.001. Disponível em: <https://epubs.icar.org.in/index.php/IJDS/article/view/137452>.

AULDIST, M. Milk Quality and Udder Health | Effect on Processing Characteristics. In: **Encyclopedia of Dairy Sciences**. [s.l.] : Elsevier, 2011. p. 902–907. DOI: 10.1016/B978-0-12-374407-4.00354-X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B978012374407400354X>.

BACIGALE, Samy Bashizi; AYAGIRWE, Rodrigue Basengere; MUTWEDU, Valence Bwana; MUGUMAARHAHAMA, Yannick; MUGISHO, Janvier Zirhumana; NZIKU, Zabron; FOFANA, Mamadou; UDOMKUN, Patchimaporn; MIGNOUNA, Jacob. Assessing milk products quality, safety, and influencing factors along the dairy value chain in eastern Democratic Republic of the Congo. **Frontiers in Sustainable Food Systems**, [S. l.], v. 7, 2023. ISSN: 2571-581X. DOI: 10.3389/fsufs.2023.1105515. Disponível em: <https://www.frontiersin.org/articles/10.3389/fsufs.2023.1105515/full>.

BERRUGA, María Isabel; DE LA VARA, Juan Ángel; LICÓN, Carmen C.; GARZÓN, Ana Isabel; GARCÍA, Andrés José; CARMONA, Manuel; CHONCO, Louis; MOLINA, Ana.

Physicochemical, Microbiological and Technological Properties of Red Deer (*Cervus elaphus*) Milk during Lactation. **Animals**, [S. l.], v. 11, n. 3, p. 906, 2021. ISSN: 2076-2615. DOI: 10.3390/ani11030906. Disponível em: <https://www.mdpi.com/2076-2615/11/3/906>.

BIANCOLILLO, Alessandra; MARINI, Federico; RUCKEBUSCH, Cyril; VITALE, Raffaele. Chemometric Strategies for Spectroscopy-Based Food Authentication. **Applied Sciences**, [S. l.], v. 10, n. 18, p. 6544, 2020. ISSN: 2076-3417. DOI: 10.3390/app10186544. Disponível em: <https://www.mdpi.com/2076-3417/10/18/6544>.

BRASIL. **Ministério da Agricultura e Pecuária. Plano Nacional de Controle de Resíduos e Contaminantes PNCRC / Animal**. 2017. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/inspecao/produtos-animal/plano-de-nacional-de-controle-de-residuos-e-contaminantes>. Acesso em: 3 fev. 2026.

BREIMAN, Leo. Random Forests. **Machine Learning**, [S. l.], v. 45, n. 1, p. 5–32, 2001. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. Disponível em: <https://link.springer.com/10.1023/A:1010933404324>.

BUTLER, Holly J.; SMITH, Benjamin R.; FRITZSCH, Robby; RADHAKRISHNAN, Pretheepan; PALMER, David S.; BAKER, Matthew J. Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. **The Analyst**, [S. l.], v. 143, n. 24, p. 6121–6134, 2018. ISSN: 0003-2654. DOI: 10.1039/C8AN01384E. Disponível em: <https://xlink.rsc.org/?DOI=C8AN01384E>.

CEBI, Nur; BEKIROGLU, Hatice; ERARSLAN, Azime; RODRIGUEZ-SAONA, Luis. Rapid Sensing: Hand-Held and Portable FTIR Applications for On-Site Food Quality Control from Farm to Fork. **Molecules**, [S. l.], v. 28, n. 9, p. 3727, 2023. ISSN: 1420-3049. DOI: 10.3390/molecules28093727. Disponível em: <https://www.mdpi.com/1420-3049/28/9/3727>.

CHEN, Jiaci et al. Comparative study on microbiological, physicochemical and nutritional properties of whole cow milk by thermal and non-thermal processing technologies. **Food Bioscience**, [S. l.], v. 59, p. 104012, 2024. ISSN: 22124292. DOI: 10.1016/j.fbio.2024.104012. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2212429224004425>.

CINAR, Mahmut; SERBESTER, Ugur; CEYHAN, Ayhan; GORGULU, Murat. Effect of Somatic Cell Count on Milk Yield and Composition of First and Second Lactation Dairy Cows. **Italian Journal of Animal Science**, [S. l.], v. 14, n. 1, p. 3646, 2015. ISSN: 1828-051X. DOI: 10.4081/ijas.2015.3646. Disponível em: <http://www.tandfonline.com/doi/full/10.4081/ijas.2015.3646>.

COSTA, Lucyana Vieira et al. Antibiotic residues in raw cow's milk: A systematic review of the last decade. **Foods**, [S. l.], v. 13, n. 23, p. 3758, 2024. ISSN: 2304-8158. DOI: 10.3390/foods13233758. Disponível em: <https://www.mdpi.com/2304-8158/13/23/3758>.

DA SILVA, Leandro P.; OLIVEIRA, Micael D. L.; VILLA, Javier E. L. A comparison of decision tree-based algorithms for food discrimination using vibrational spectroscopy. **Food Chemistry**, [S. l.], v. 488, p. 144909, 2025. ISSN: 03088146. DOI: 10.1016/j.foodchem.2025.144909. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0308814625021600>.

DAYANANDA, B.; OWEN, S.; KOLOBARIC, A.; CHAPMAN, J.; COZZOLINO, D. Pre-processing Applied to Instrumental Data in Analytical Chemistry: A Brief Review of the Methods and Examples. **Critical Reviews in Analytical Chemistry**, [S. l.], v. 54, n. 8, p. 2745–2753, 2024. ISSN: 1040-8347. DOI: 10.1080/10408347.2023.2199864. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/10408347.2023.2199864>.

DE FREITAS, Alexandre GM; DE MAGALHÃES, Bárbara EA; MINHO, Lucas AC; LEÃO, Danilo J.; SANTOS, Leandro S.; AUGUSTO DE ALBUQUERQUE FERNANDES, Sérgio. FTIR spectroscopy with chemometrics for determination of tylosin residues in milk. **Journal of the Science of Food and Agriculture**, [S. l.], v. 101, n. 5, p. 1854–1860, 2021. ISSN: 0022-5142. DOI: 10.1002/jsfa.10799. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/jsfa.10799>.

DE JONG, Sijmen. SIMPLS: An alternative approach to partial least squares regression. **Chemometrics and Intelligent Laboratory Systems**, [S. l.], v. 18, n. 3, p. 251–263, 1993. ISSN: 01697439. DOI: 10.1016/0169-7439(93)85002-X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/016974399385002X>.

DE SANTANA, Felipe Bachion; BORGES NETO, Waldomiro; POPPI, Ronei J. Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. **Food Chemistry**, [S. l.], v. 293, p. 323–332, 2019. ISSN: 03088146. DOI: 10.1016/j.foodchem.2019.04.073. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0308814619307368>.

DENOBILO, Michela; DE SOUZA NASCIMENTO, Elizabeth. Validação de método para determinação de resíduos oxitetraciclina, tetraciclina, clortetaciclina e doxiciclina, em leite, por cromatografia líquida de alta eficiência. **Revista Brasileira de Ciências Farmaceuticas/Brazilian Journal of Pharmaceutical Sciences**, [S. l.], v. 40, n. 2, p. 209–

218, 2004. ISSN: 15169332.

DI ROCCO, Melissa; MOLONEY, Mary; HAREN, Deirdre; GUTIERREZ, Montserrat; EARLEY, Seán; BERENDSEN, Bjorn; FUREY, Ambrose; DANAHER, Martin. Improving the chromatographic selectivity of β -lactam residue analysis in milk using phenyl-column chemistry prior to detection by tandem mass spectrometry. **Analytical and Bioanalytical Chemistry**, [S. l.], v. 412, n. 18, p. 4461–4475, 2020. ISSN: 1618-2642. DOI: 10.1007/s00216-020-02688-4. Disponível em: <https://link.springer.com/10.1007/s00216-020-02688-4>.

DOS SANTOS, Cleidiane Pereira; BARBOSA DA SILVA, Rosângela Silveira; SOUZA, Romicy Dermondes; DE ALBUQUERQUE FERNANDES, Sérgio Augusto. Mid-infrared spectroscopy combined with chemometrics for the detection of ciprofloxacin and Bronopol® residues in milk. **Food Chemistry**, [S. l.], v. 504, p. 147826, 2026. ISSN: 03088146. DOI: 10.1016/j.foodchem.2025.147826. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0308814625050782>.

DULIŃSKI, Robert; GANCARZ, Marek; SHAKHOVSKA, Nataliya; BYCZYŃSKI, Łukasz. The Application of Fourier Transform Infrared Spectroscopy and Chemometrics in Identifying Signatures for Sheep's Milk Authentication. **Processes**, [S. l.], v. 13, n. 2, p. 518, 2025. ISSN: 2227-9717. DOI: 10.3390/pr13020518. Disponível em: <https://www.mdpi.com/2227-9717/13/2/518>.

ERMETIN, Orhan; KUL, Ertuğrul; OKUYUCU, Ibrahim Cihangir. The relationship of somatic cell count with milk yield and composition in different stages of lactation in Holstein cows. **Veterinarski arhiv**, [S. l.], v. 94, n. 1, p. 21–32, 2024. ISSN: 03725480. DOI: 10.24099/vet.arhiv.2213. Disponível em: <https://vetarhiv.vef.unizg.hr/papers/2024-94-1-3.pdf>.

FAO. Code of hygienic practice for milk and milk products 2004. p. 1–33. Disponível em: https://www.fao.org/fileadmin/user_upload/livestockgov/documents/CXP_057e.pdf.

FAO. Codex Alimentarius: Maximum residue limits (MLRs) and risk management recommendations (RMRS) for residues of veterinary drugs in foods 2023. p. 1–53. Disponível em: <https://www.fao.org/fao-who-codexalimentarius/sh-proxy/en/?lnk=1&url=https%253A%252F%252Fworkspace.fao.org%252Fsites%252Fcodex%252FStandards%252FCXM%2B2%252FMRL2e.pdf>.

FAO. **Dairy Market Review: Overview of global market developments in 2024**. Rome. Disponível em: <https://openknowledge.fao.org/handle/20.500.14283/cd5441en>.

FOROUTAN, Aidin et al. Chemical composition of commercial cow's milk. **Journal of**

Agricultural and Food Chemistry, [S. l.], v. 67, n. 17, p. 4897–4914, 2019. ISSN: 0021-8561. DOI: 10.1021/acs.jafc.9b00204. Disponível em: <https://pubs.acs.org/doi/10.1021/acs.jafc.9b00204>.

FREITAS, Alexandre Gomes Marques De; ALMIR CAVALCANTE MINHO, Lucas; ELIZABETH ALVES DE MAGALHÃES, Bárbara; NEI LOPES DOS SANTOS, Walter; SOARES SANTOS, Leandro; AUGUSTO DE ALBUQUERQUE FERNANDES, Sérgio. Infrared spectroscopy combined with random forest to determine tylosin residues in powdered milk. **Food Chemistry**, [S. l.], v. 365, p. 130477, 2021. ISSN: 03088146. DOI: 10.1016/j.foodchem.2021.130477. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0308814621014837>.

FUSCO, Vincenzina; CHIEFFI, Daniele; FANELLI, Francesca; LOGRIECO, Antonio F.; CHO, Gyu-Sung; KABISCH, Jan; BÖHNLEIN, Christina; FRANZ, Charles M. A. P. Microbial quality and safety of milk and milk products in the 21st century. **Comprehensive Reviews in Food Science and Food Safety**, [S. l.], v. 19, n. 4, p. 2013–2049, 2020. ISSN: 1541-4337. DOI: 10.1111/1541-4337.12568. Disponível em: <https://ift.onlinelibrary.wiley.com/doi/10.1111/1541-4337.12568>.

GAI, Nan; UNIACKE-LOWE, Therese; O'REGAN, Jonathan; FAULKNER, Hope; KELLY, Alan L. Effect of Protein Genotypes on Physicochemical Properties and Protein Functionality of Bovine Milk: A Review. **Foods**, [S. l.], v. 10, n. 10, p. 2409, 2021. ISSN: 2304-8158. DOI: 10.3390/foods10102409. Disponível em: <https://www.mdpi.com/2304-8158/10/10/2409>.

HOUHOU, Rola; BOCKLITZ, Thomas. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. **Analytical Science Advances**, [S. l.], v. 2, n. 3–4, p. 128–141, 2021. ISSN: 2628-5452. DOI: 10.1002/ansa.202000162. Disponível em: <https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/ansa.202000162>.

IBGE. **Pesquisa da Pecuária Municipal: Produção de Leite**. 2025. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9107-producao-da-pecuaria-municipal.html>. Acesso em: 9 fev. 2026.

IRANZAD, Reza; LIU, Xiao. A review of random forest-based feature selection methods for data science education and applications. **International Journal of Data Science and Analytics**, [S. l.], v. 20, n. 2, p. 197–211, 2025. ISSN: 2364-415X. DOI: 10.1007/s41060-024-00509-w. Disponível em: <https://link.springer.com/10.1007/s41060-024-00509-w>.

JOLLIFFE, Ian T.; CADIMA, Jorge. Principal component analysis: a review and recent

developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, [S. l.], v. 374, n. 2065, p. 20150202, 2016. ISSN: 1364-503X. DOI: 10.1098/rsta.2015.0202. Disponível em:

<https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>.

KARAMAN, Ayse Demet; AYKAS, Didem Peren; RODRIGUEZ-SAONA, Rodriguez; ALVAREZ, Valente. Determination of the Compositional Differences in Organic and Conventional Milk using FTIR Spectroscopy. **Turkish Journal of Agriculture - Food Science and Technology**, [S. l.], v. 9, n. 3, p. 452–459, 2021. ISSN: 2148-127X. DOI: 10.24925/turjaf.v9i3.452-459.3365. Disponível em:

<http://www.agrifoodscience.com/index.php/TURJAF/article/view/3365>.

KAROUI, Romdhane; DOWNEY, Gerard; BLECKER, Christophe. Mid-Infrared Spectroscopy Coupled with Chemometrics: A Tool for the Analysis of Intact Food Systems and the Exploration of Their Molecular Structure–Quality Relationships – A Review. **Chemical Reviews**, [S. l.], v. 110, n. 10, p. 6144–6168, 2010. ISSN: 0009-2665. DOI: 10.1021/cr100090k. Disponível em: <https://pubs.acs.org/doi/10.1021/cr100090k>.

KASKOUS, Shehadeh; FARSCHTSCHI, Sabine; PFAFFL, Michael W. Physiological Aspects of Milk Somatic Cell Count in Small Ruminants—A Review. **Dairy**, [S. l.], v. 4, n. 1, p. 26–42, 2022. ISSN: 2624-862X. DOI: 10.3390/dairy4010002. Disponível em: <https://www.mdpi.com/2624-862X/4/1/2>.

KUDRINA, Marina; KOGEVNIKOVA, Irina; KHUDYAKOVA, Natalya. Nutritional value of cow's milk. **Bulletin of KSAU**, [S. l.], n. 12, p. 229–236, 2025. ISSN: 1819-4036. DOI: 10.36718/1819-4036-2022-12-229-236. Disponível em: <http://kgau.editorum.ru/en/nauka/article/94149/view>.

KURSA, Miron B.; JANKOWSKI, Aleksander; RUDNICKI, Witold R. Boruta – A System for Feature Selection. **Fundamenta Informaticae**, [S. l.], v. 101, n. 4, p. 271–285, 2010. ISSN: 0169-2968. DOI: 10.3233/FI-2010-288. Disponível em: <https://journals.sagepub.com/doi/10.3233/FI-2010-288>.

KURSA, Miron B.; RUDNICKI, Witold R. Feature selection with the boruta package. **Journal of Statistical Software**, [S. l.], v. 36, n. 11, 2010. ISSN: 1548-7660. DOI: 10.18637/jss.v036.i11. Disponível em: <http://www.jstatsoft.org/v36/i11/>.

KURSA, Miron Bartosz. Robustness of Random Forest-based gene selection methods. **BMC Bioinformatics**, [S. l.], v. 15, n. 1, p. 8, 2014. ISSN: 1471-2105. DOI: 10.1186/1471-2105-15-

8. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-8>.

LARKIN, Peter J. **Infrared and Raman Spectroscopy**. 2. ed ed., [s.l.] : Elsevier, 2018. ISBN: 9780128041628. DOI: 10.1016/C2015-0-00806-1. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/C20150008061>.

LEE, Loong Chuen; LIONG, Choong-Yeun; JEMAIN, Abdul Aziz. Effects of data pre-processing methods on classification of ATR-FTIR spectra of pen inks using partial least squares-discriminant analysis (PLS-DA). **Chemometrics and Intelligent Laboratory Systems**, [S. l.], v. 182, p. 90–100, 2018. ISSN: 01697439. DOI: 10.1016/j.chemolab.2018.09.001. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743917307840>.

LOIANNO, Valerio. On the measurement of the mutual diffusivity of binary gas mixtures with FTIR Spectroscopy. **Chemical Engineering Science**, [S. l.], v. 270, p. 118546, 2023. ISSN: 00092509. DOI: 10.1016/j.ces.2023.118546. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0009250923001021>.

LUZ, Renata; BARROS, Marilisa Berti de Azevedo; DE ASSUMPÇÃO, Daniela; DOMENE, Semíramis Martins Álvares; BARROS FILHO, Antonio de Azevedo. Do adults and older people follow current guidelines for milk consumption? **Public Health Nutrition**, [S. l.], v. 24, n. 14, p. 4622–4629, 2021. ISSN: 1368-9800. DOI: 10.1017/S1368980021000057. Disponível em: https://www.cambridge.org/core/product/identifier/S1368980021000057/type/journal_article.

MA, Y.; RYAN, C.; BARBANO, D. M.; GALTON, D. M.; RUDAN, M. A.; BOOR, K. J. Effects of Somatic Cell Count on Quality and Shelf-Life of Pasteurized Fluid Milk. **Journal of Dairy Science**, [S. l.], v. 83, n. 2, p. 264–274, 2000. ISSN: 00220302. DOI: 10.3168/jds.S0022-0302(00)74873-9. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0022030200748739>.

MAHMOOD, T.; ABBAS, M.; ILYAS, S.; AFZAL, N.; NAWAZ, R. Quantification of fluoroquinolone (enrofloxacin, norfloxacin and ciprofloxacin) residues in cow milk. **Ijcb**, [S. l.], v. 10, n. 2377, p. 10–15, 2016. Disponível em: www.iscientific.org/Journal.html.

MANIKANDAN, G.; PRAGADEESH, B.; MANOJKUMAR, V.; KARTHIKEYAN, A. L.; MANIKANDAN, R.; GANDOMI, Amir H. Classification models combined with Boruta feature selection for heart disease prediction. **Informatics in Medicine Unlocked**, [S. l.], v.

44, p. 101442, 2024. ISSN: 23529148. DOI: 10.1016/j.imu.2023.101442. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2352914823002885>.

MAPA. INSTRUÇÃO NORMATIVA Nº 77, DE 26 DE NOVEMBRO DE 20182018a. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750141/do1-2018-11-30-instrucao-normativa-n-77-de-26-de-novembro-de-2018-52749887.

MAPA. INSTRUÇÃO NORMATIVA Nº 76, DE 26 DE NOVEMBRO DE 20182018b. 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-2018-52749894IN_76.

MATEOS-APARICIO, Gregoria. Partial Least Squares (PLS) Methods: Origins, Evolution, and Application to Social Sciences. **Communications in Statistics - Theory and Methods**, [S. l.], v. 40, n. 13, p. 2305–2317, 2011. ISSN: 0361-0926. DOI: 10.1080/03610921003778225. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/03610921003778225>.

MCCARTHY, O. J. Physical and Physico-Chemical Properties of Milk. In: **Encyclopedia of Dairy Sciences**. [s.l.] : Elsevier, 2022. p. 603–614. DOI: 10.1016/B978-0-12-818766-1.00334-2. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B9780128187661003342>.

MEGLEN, Robert R. Examining large databases: a chemometric approach using principal component analysis. **Marine Chemistry**, [S. l.], v. 39, n. 1–3, p. 217–237, 1992. ISSN: 03044203. DOI: 10.1016/0304-4203(92)90103-H. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/030442039290103H>.

MEHMOOD, Tahir; LILAND, Kristian Hovde; SNIPEN, Lars; SÆBØ, Solve. A review of variable selection methods in Partial Least Squares Regression. **Chemometrics and Intelligent Laboratory Systems**, [S. l.], v. 118, p. 62–69, 2012. ISSN: 01697439. DOI: 10.1016/j.chemolab.2012.07.010. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743912001542>.

MENDES, Eduarda; DUARTE, Noélia. Mid-Infrared Spectroscopy as a Valuable Tool to Tackle Food Analysis: A Literature Review on Coffee, Dairies, Honey, Olive Oil and Wine. **Foods**, [S. l.], v. 10, n. 2, p. 477, 2021. ISSN: 2304-8158. DOI: 10.3390/foods10020477. Disponível em: <https://www.mdpi.com/2304-8158/10/2/477>.

MOKARI, Azadeh; GUO, Shuxia; BOCKLITZ, Thomas. Exploring the Steps of Infrared (IR) Spectral Analysis: Pre-Processing, (Classical) Data Modelling, and Deep Learning. **Molecules**,

[S. l.], v. 28, n. 19, p. 6886, 2023. ISSN: 1420-3049. DOI: 10.3390/molecules28196886. Disponível em: <https://www.mdpi.com/1420-3049/28/19/6886>.

MURPHY, Mary M.; BARRAJ, Leila M.; TOTH, Lindsey D.; HARKNESS, Laura S.; BOLSTER, Douglas R. Daily intake of dairy products in Brazil and contributions to nutrient intakes: a cross-sectional study. **Public Health Nutrition**, [S. l.], v. 19, n. 3, p. 393–400, 2016. ISSN: 1368-9800. DOI: 10.1017/S1368980015001780. Disponível em: https://www.cambridge.org/core/product/identifier/S1368980015001780/type/journal_article.

NANDIYANTO, Asep Bayu Dani; RAGADHITA, Risti; FIANDINI, Meli. Interpretation of Fourier Transform Infrared Spectra (FTIR): A Practical Approach in the Polymer/Plastic Thermal Decomposition. **Indonesian Journal of Science and Technology**, [S. l.], v. 8, n. 1, p. 113–126, 2022. ISSN: 2527-8045. DOI: 10.17509/ijost.v8i1.53297. Disponível em: <https://ejournal.upi.edu/index.php/ijost/article/view/53297>.

NAZRI, MUHAMMAD ZULHELMI. Application of Fourier Transform Infrared (FTIR) Spectroscopy, Gas Chromatography and Chemometric Analyses for Halal Authentication in Food Products. **Journal of Academia**, [S. l.], v. 13, n. 2, p. 209–223, 2025. ISSN: 2289-6368. DOI: 10.24191/joa.v13i2.6201. Disponível em: <https://journal.uitm.edu.my/ojs/index.php/JOA/article/view/6201>.

NUNES, Lauane; TAVARES, Guilherme M. Thermal treatments and emerging technologies: Impacts on the structure and techno-functional properties of milk proteins. **Trends in Food Science & Technology**, [S. l.], v. 90, p. 88–99, 2019. ISSN: 09242244. DOI: 10.1016/j.tifs.2019.06.004. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0924224418308550>.

PARMAR, Aakash; KATARIYA, Rakesh; PATEL, Vatsal. A Review on Random Forest: An Ensemble Classifier. *In*: [s.l.: s.n.]. p. 758–763. DOI: 10.1007/978-3-030-03146-6_86. Disponível em: http://link.springer.com/10.1007/978-3-030-03146-6_86.

PIRES, Ana José; PEREIRA, Gonçalo; FANGUEIRO, David; BEXIGA, Ricardo; OLIVEIRA, Manuela. When the solution becomes the problem: a review on antimicrobial resistance in dairy cattle. **Future Microbiology**, [S. l.], v. 19, n. 10, p. 903–929, 2024. ISSN: 1746-0913. DOI: 10.2217/fmb-2023-0232. Disponível em: <https://www.tandfonline.com/doi/full/10.2217/fmb-2023-0232>.

POSSA, Gabriela; CASTRO, Michelle Alessandra De; SICHIERI, Rosely; FISBERG, Regina Mara; FISBERG, Mauro. Dairy products consumption in Brazil is associated with

socioeconomic and demographic factors: Results from the National Dietary Survey 2008-2009. **Revista de Nutrição**, [S. l.], v. 30, n. 1, p. 79–90, 2017. ISSN: 1415-5273. DOI: 10.1590/1678-98652017000100008. Disponível em:

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-52732017000100079&lng=en&tlng=en.

PUGA-TORRES, Byron; ARAGÓN VÁSQUEZ, Eduardo; RON, Lenin; ÁLVAREZ, Vanessa; BONILLA, Salomé; GUZMÁN, Aldair; LARA, Diego; DE LA TORRE, David. Milk Quality Parameters of Raw Milk in Ecuador between 2010 and 2020: A Systematic Literature Review and Meta-Analysis. **Foods**, [S. l.], v. 11, n. 21, p. 3351, 2022. ISSN: 2304-8158. DOI: 10.3390/foods11213351. Disponível em: <https://www.mdpi.com/2304-8158/11/21/3351>.

RACHAH, Amira; REKSEN, Olav; TAFINTSEVA, Valeria; STEHR, Felicia Judith Marie; RUKKE, Elling-Olav; PRESTLØKKEN, Egil; MARTIN, Adam; KOHLER, Achim; AFSETH, Nils Kristian. Exploring Dry-Film FTIR Spectroscopy to Characterize Milk Composition and Subclinical Ketosis throughout a Cow's Lactation. **Foods**, [S. l.], v. 10, n. 9, p. 2033, 2021. ISSN: 2304-8158. DOI: 10.3390/foods10092033. Disponível em: <https://www.mdpi.com/2304-8158/10/9/2033>.

RINNAN, Åsmund; BERG, Frans Van den; ENGELSEN, Søren Balling. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, [S. l.], v. 28, n. 10, p. 1201–1222, 2009. ISSN: 01659936. DOI: 10.1016/j.trac.2009.07.007. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0165993609001629>.

RODRÍGUEZ-BERMÚDEZ, Ruth; FOUZ, Ramiro; RICO, Margarita; CAMINO, Fernando; SOUZA, Taile Katiele; MIRANDA, Marta; DIÉGUEZ, Francisco Javier. Factors Affecting Fatty Acid Composition of Holstein Cow's Milk. **Animals**, [S. l.], v. 13, n. 4, p. 574, 2023. ISSN: 2076-2615. DOI: 10.3390/ani13040574. Disponível em: <https://www.mdpi.com/2076-2615/13/4/574>.

SACHI, Sabbya; FERDOUS, Jannatul; SIKDER, Mahmudul; HUSSANI, S. Antibiotic residues in milk: Past, present, and future. **Journal of Advanced Veterinary and Animal Research**, [S. l.], v. 6, n. 3, p. 315, 2019. ISSN: 2311-7710. DOI: 10.5455/javar.2019.f350. Disponível em: <https://ejmanager.com/fulltextpdf.php?mno=28703>.

SAJI, Rakendhu; GANDHI, Kamal; SHARMA, Rajan; BAJAJ, Rajesh; MANN, Bimlesh; RAMANI, Akshay. Detection of flunixin residues in milk using ATR- FTIR spectroscopy

coupled with chemometrics. **Journal of Food Measurement and Characterization**, [S. l.], v. 18, n. 8, p. 6736–6746, 2024. ISSN: 2193-4126. DOI: 10.1007/s11694-024-02686-5. Disponível em: <https://link.springer.com/10.1007/s11694-024-02686-5>.

SANTOS, Roberta de Oliveira; GORGULHO, Bartira Mendes; CASTRO, Michelle Alessandra De; FISBERG, Regina Mara; MARCHIONI, Dirce Maria; BALTAR, Valéria Troncoso. Principal Component Analysis and Factor Analysis: differences and similarities in Nutritional Epidemiology application. **Revista Brasileira de Epidemiologia**, [S. l.], v. 22, 2019. ISSN: 1980-5497. DOI: 10.1590/1980-549720190041. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2019000100439&tlng=en.

SHINDE, Pramila P.; SHAH, Seema. A Review of Machine Learning and Deep Learning Applications. In: 2018 FOURTH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION CONTROL AND AUTOMATION (ICCUBEA) 2018, **Anais [...]**. : IEEE, 2018 p. 1–6. ISBN: 978-1-5386-5257-2. DOI: 10.1109/ICCUBEA.2018.8697857. Disponível em: <https://ieeexplore.ieee.org/document/8697857/>.

SILVA, Janiquelli Barbosa; ELIAS, Bianca Caroline; MAIS, Laís Amaral; WARKENTIN, Sarah; KONSTANTYNER, Tulio; OLIVEIRA, Fernanda Luisa Ceragioli. Factors associated with inadequate milk consumption among adolescents: National school health survey - Pense 2012. **Revista Paulista de Pediatria**, [S. l.], v. 38, 2020. ISSN: 1984-0462. DOI: 10.1590/1984-0462/2020/38/2018184. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-05822020000100409&tlng=en.

SIVAKESAVA, S.; IRUDAYARAJ, J. Rapid Determination of Tetracycline in Milk by FT-MIR and FT-NIR Spectroscopy. **Journal of Dairy Science**, [S. l.], v. 85, n. 3, p. 487–493, 2002. ISSN: 00220302. DOI: 10.3168/jds.S0022-0302(02)74099-X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S002203020274099X>.

STANEK, Piotr; ŻÓŁKIEWSKI, Paweł; JANUŚ, Ewa. A Review on Mastitis in Dairy Cows Research: Current Status and Future Perspectives. **Agriculture**, [S. l.], v. 14, n. 8, p. 1292, 2024. ISSN: 2077-0472. DOI: 10.3390/agriculture14081292. Disponível em: <https://www.mdpi.com/2077-0472/14/8/1292>.

TENG, Z. W.; YANG, G. Q.; WANG, L. F.; FU, T.; LIAN, H. X.; SUN, Y.; HAN, L. Q.; ZHANG, L. Y.; GAO, T. Y. Effects of the circadian rhythm on milk composition in dairy cows:

Does day milk differ from night milk? **Journal of Dairy Science**, [S. l.], v. 104, n. 7, p. 8301–8313, 2021. ISSN: 00220302. DOI: 10.3168/jds.2020-19679. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0022030221005142>.

TOUW, W. G.; BAYJANOV, J. R.; OVERMARS, L.; BACKUS, L.; BOEKHORST, J.; WELS, M.; VAN HIJUM, S. A. F. T. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? **Briefings in Bioinformatics**, [S. l.], v. 14, n. 3, p. 315–326, 2013. ISSN: 1467-5463. DOI: 10.1093/bib/bbs034. Disponível em: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs034>.

TSENKOVA, R.; ATANASSOVA, S.; OZAKI, Y.; TOYODA, K.; ITOH, K. Near-infrared spectroscopy for biomonitoring: influence of somatic cell count on cow's milk composition analysis. **International Dairy Journal**, [S. l.], v. 11, n. 10, p. 779–783, 2001. ISSN: 09586946. DOI: 10.1016/S0958-6946(01)00110-8. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0958694601001108>.

UE. Commission Delegated Regulation (EU) 2018/149 of 15 November 2017 amending Delegated Regulation (EU) 2016/1238 with regard to the compositional requirements and quality characteristics of milk and milk products eligible for public intervention and aid for. 2017. Disponível em: https://eur-lex.europa.eu/eli/reg_del/2018/149/oj. Acesso em: 21 jun. 2025.

WILKIE, Daisy; WHITE, Brad; HEIDARI, Golnaz; NAFFA, Rafea; PEDDIE, Gaile; ROWLANDS, Gareth J.; PLIEGER, Paul G. Methods for Untargeted Analysis of Milk Metabolites: Influence of Extraction Method and Optimization of Separation. **Metabolites**, [S. l.], v. 15, n. 9, p. 597, 2025. ISSN: 2218-1989. DOI: 10.3390/metabo15090597. Disponível em: <https://www.mdpi.com/2218-1989/15/9/597>.

WODAJO TIRFIE, Fikadu. A Review of Genetic and Non-Genetic Parameter Estimates for Milk Composition of Cattle. **Animal and Veterinary Sciences**, [S. l.], 2023. DOI: 10.11648/j.av.s.20231103.12. Disponível em: <http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=212&doi=10.11648/j.av.s.20231103.12>.

WOLD, Svante; ESBENSEN, Kim; GELADI, Paul. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, [S. l.], v. 2, n. 1–3, p. 37–52, 1987. ISSN: 01697439. DOI: 10.1016/0169-7439(87)80084-9. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0169743987800849>.

WOLD, Svante; SJÖSTRÖM, Michael; ERIKSSON, Lennart. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, [S. l.], v. 58, n. 2, p. 109–130, 2001. ISSN: 01697439. DOI: 10.1016/S0169-7439(01)00155-1. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0169743901001551>.

YAN, Chunsheng. A review on spectral data preprocessing techniques for machine learning and quantitative analysis. **iScience**, [S. l.], v. 28, n. 7, p. 112759, 2025. ISSN: 25890042. DOI: 10.1016/j.isci.2025.112759. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S258900422501020X>.

YI, Lunzhao; DONG, Naiping; YUN, Yonghuan; DENG, Baichuan; REN, Dabing; LIU, Shao; LIANG, Yizeng. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. **Analytica Chimica Acta**, [S. l.], v. 914, p. 17–34, 2016. ISSN: 00032670. DOI: 10.1016/j.aca.2016.02.001. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0003267016301866>.

ZHOU, Hongfang; XIN, Yinbo; LI, Suli. A diabetes prediction model based on Boruta feature selection and ensemble learning. **BMC Bioinformatics**, [S. l.], v. 24, n. 1, p. 224, 2023. ISSN: 1471-2105. DOI: 10.1186/s12859-023-05300-5. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05300-5>.

ZOU, Yunxia; CHEN, Yifei; MENG, Qingyong; WANG, Yachun; ZHANG, Yali. Cow Milk Fatty Acid and Protein Composition in Different Breeds and Regions in China. **Molecules**, [S. l.], v. 29, n. 21, p. 5142, 2024. ISSN: 1420-3049. DOI: 10.3390/molecules29215142. Disponível em: <https://www.mdpi.com/1420-3049/29/21/5142>.

II – OBJETIVOS GERAIS

3 Objetivo geral

Desenvolver e validar modelos baseados em MIR, associados a métodos quimiométricos e de aprendizado de máquina, para monitorar dois indicadores críticos de qualidade do leite, a presença de resíduos de ciprofloxacina e a contagem de células somáticas (CCS), visando uma análise rápida, não destrutiva e sustentável.

4 Objetivos específicos

- Comparar diferentes abordagens de modelagem, abrangendo desde técnicas clássicas de regressão multivariada, como os Mínimos Quadrados Parciais (PLS), até algoritmos avançados de aprendizado de máquina, como o Random Forest (RF).
- Desenvolver modelos de classificação para identificar amostras de leite que estejam em desconformidade com os limites regulatórios, como o Limite Máximo de Resíduos para antibióticos e os limites de CCS estabelecidos pela legislação.
- Avaliar e aplicar métodos de seleção de variáveis, como a Análise de Componentes Principais (PCA) e o algoritmo Boruta, para identificar as regiões do espectro MIR mais informativas e otimizar o desempenho preditivo dos modelos.
- Identificar as assinaturas espectrais e os grupos funcionais associados às variações de CCS e à presença de ciprofloxacina na matriz complexa do leite líquido.
- Validar a eficiência das ferramentas desenvolvidas como métodos de triagem para aplicação em larga escala no controle de qualidade industrial e na segurança alimentar.

III – CAPÍTULO I

ARTIGO 1 - Detection and classification of ciprofloxacin residues in bovine milk by FTIR spectroscopy and multivariate analysis

1 DETECTION AND CLASSIFICATION OF

2 CIPROFLOXACIN RESIDUES IN BOVINE

3 MILK BY FTIR SPECTROSCOPY AND

4 MULTIVARIATE ANALYSIS

5 *Romicy Dermondes Souza*¹, *Rosângela Silveira Barbosa*², *Marcelo Bonnet Alvarenga*²,
 6 *Fabiano Barreto*³, *Wellington da Silva Lyra*⁴, *Sérgio de Sousa Castro*¹, *Paulo Henrique*
 7 *Gonçalves Dias Diniz*⁵, *Sérgio Augusto de Albuquerque Fernandes*^{1*}

8 ¹Department of Rural and Animal Technology, State University of Southwest Bahia, 45700-
 9 000, Bahia, Itapetinga, Brazil

10 ²Embrapa Temperate Climate, 96010-971, Rio Grande do Sul, Pelotas, Brazil

11 ³Federal Agricultural Defense Laboratory of Rio Grande do Sul, Ministry of Agriculture,
 12 Livestock and Supply, 91780-580, Rio Grande do Sul, Porto Alegre, Brazil

13 ⁴Department of Chemistry, Center for Exact and Nature Sciences, Federal University of Paraíba
 14 - Campus I, 58051-900, Paraíba, João Pessoa, Brazil

15 ⁵Post-Graduate Program in Pure and Applied Chemistry, Federal University of Western Bahia,
 16 47810-059, Bahia, Barreiras, Brazil

17

18 *Corresponding author. E-mail: dermondesromicy@gmail.com

19 Contributing authors: rosangela.barbosa@embrapa.br; marcelo.bonnet@embrapa.br;

20 Fabiano.barreto@agro.gov.br; spectru@gmail.com; scastro@uesb.edu.br;

21 paulo.diniz@ufob.edu.br; sfernandes@uesb.edu.br;

1 **ABSTRACT**

2 This study evaluated the potential of mid-infrared spectroscopy combined with chemometric
3 and machine learning methods to detect ciprofloxacin residues in liquid cow's milk, aiming to
4 propose a fast and low-cost alternative to conventional methods. A total of 605 milk samples
5 fortified with concentrations ranging from 0 to 250 µg/L of the antibiotic were analyzed, and
6 the spectrum variables were selected by PCA and Boruta, followed by modeling with PLS and
7 random forest for regression, and linear discriminant analysis and random forest for
8 classification into two categories, below and above the maximum residue limit (100 µg/L).
9 Regression models showed limited performance in accurately quantifying the antibiotic.
10 However, classification models, particularly Random Forest, demonstrated the ability to
11 consistently distinguish between samples with concentrations above and below the established
12 regulatory limit. The analysis showed that the natural variability in milk composition influences
13 the performance of the models. The FTIR approach with machine learning showed promise as
14 a rapid screening method for detecting ciprofloxacin residues in milk, although its application
15 for accurate quantification still needs refinement. The technique has potential for routine
16 monitoring in the dairy industry.

17 **KEYWORDS**

18 Chemometrics, food safety, antibiotics, quality control.

19

20

21

22

23

24

25

1 INTRODUCTION

2 The use of ciprofloxacin in dairy cows is mainly related to the treatment and prevention
3 of mastitis, the most prevalent disease in dairy cows, causing economic losses to the agro-
4 industrial milk system ^{1,2}. Despite being widely used, the inappropriate use of ciprofloxacin
5 can result in residues of the antibiotic in milk, which poses risks to public health and contributes
6 to an increase in antimicrobial resistance. Several studies have identified the presence of
7 ciprofloxacin in milk samples, with some of them exceeding the maximum limits established ³.
8 ⁴. Prolonged exposure to ciprofloxacin residues can promote the emergence of resistant bacteria
9 such as *Escherichia coli* and *Staphylococcus aureus* ⁵. In addition, resistance genes can be
10 passed on between milk bacteria and the human microbiota, increasing the risk of antimicrobial
11 resistance spreading ⁶. Given the potential problems generated by ciprofloxacin residues in
12 milk, various methods have been developed to identify and quantify its residues.

13 The methods commonly used to detect and quantify ciprofloxacin in milk are high-
14 performance liquid chromatography (HPLC) and immunological tests. HPLC is considered the
15 standard for its sensitivity and precision, but it requires expensive equipment, difficult sample
16 preparation and trained personnel, so it is slow and expensive for everyday use ⁷.
17 Immunochromatographic tests are easy, quick and useful for screening, but they have lower
18 sensitivity and can react with other similar antibiotics ⁸. Therefore, given the problems
19 presented, there is a need for fast, cheap and reliable methods to check for antibiotics in milk.

20 Fourier transform infrared spectroscopy (FTIR) has several advantages for antibiotic
21 detection: it is a fast, non-destructive technique that requires little sample preparation and is
22 capable of identifying and quantifying residues directly in complex matrices such as milk ^{9,10}.
23 The combination of FTIR with chemometrics, i.e. statistical and machine learning methods
24 applied to spectral data, makes it possible to extract relevant information from complex and
25 overlapping spectra, increasing the sensitivity, specificity and accuracy of analysis ^{11, 12}.

Although there is research applying advanced spectroscopic techniques such as Raman coupled with machine learning to identify quinolone antibiotics such as ciprofloxacin in dairy products, these studies do not use FTIR as the main method ¹³. Another study used integrated FTIR and chemometrics methods to quantify tylosin residues in freeze-dried milk, but did not analyze its application to quinolone class antibiotics or implement the method in liquid milk ¹⁴. Most research focuses on techniques such as fluorescence, electrochemistry or chromatography ¹⁵. Thus, there is no research investigating the potential of FTIR combined with machine learning algorithms for the direct, rapid and automated analysis of ciprofloxacin in liquid milk. Based on the above, this study aims to analyze the potential of the FTIR technique, in combination with chemometric and machine learning methods, to identify and classify ciprofloxacin residues in cow's milk.

MATERIALS AND METHODS

Material collection

Milk samples were collected at Embrapa Clima Temperado's Terras Baixas Experimental Station (ETB), in the municipality of Capão do Leão, Rio Grande do Sul, Brazil (31° 49' 27" S and 52° 26' 35" W). For the experiments, lactating Jersey cows were selected. The cows were healthy, with no history of mastitis or other diseases and had not been receiving antibiotics in the 30 days prior to collection. First, the teats were cleaned with iodine solution. They were then dried with disposable paper towels. The first three jets of milk from each teat were discarded. The milk samples were placed in vials. After collection, the samples were placed in boxes with recyclable ice. The temperature was kept between 0°C and 10°C. The samples were transported to the Milk Analysis Laboratory (LabLeite) within 24 hours. During transportation, the cold chain was maintained. On arrival at the laboratory, the samples were stored at temperatures between 0°C and 10°C. Before analysis, the composition of the milk was

analyzed using Fourier Transform Mid-Infrared Spectroscopy (FTIR-MIR), following the official analysis protocol of the Brazilian Milk Quality Network (RBQL), as adopted by Embrapa.

Experimental design

The design of the fortification of the milk samples is shown in Figure 1. The fortification process was carried out using an analytical standard of ciprofloxacin (Sigma-Aldrich, MO, USA), with 11 different concentration levels being established: 0, 25, 50, 75, 100, 125, 150, 175, 200, 225 and 250 $\mu\text{g/L}$. To ensure the robustness of the study, each concentration level was represented by 55 individual samples, totaling 605 fortified milk samples for analysis.

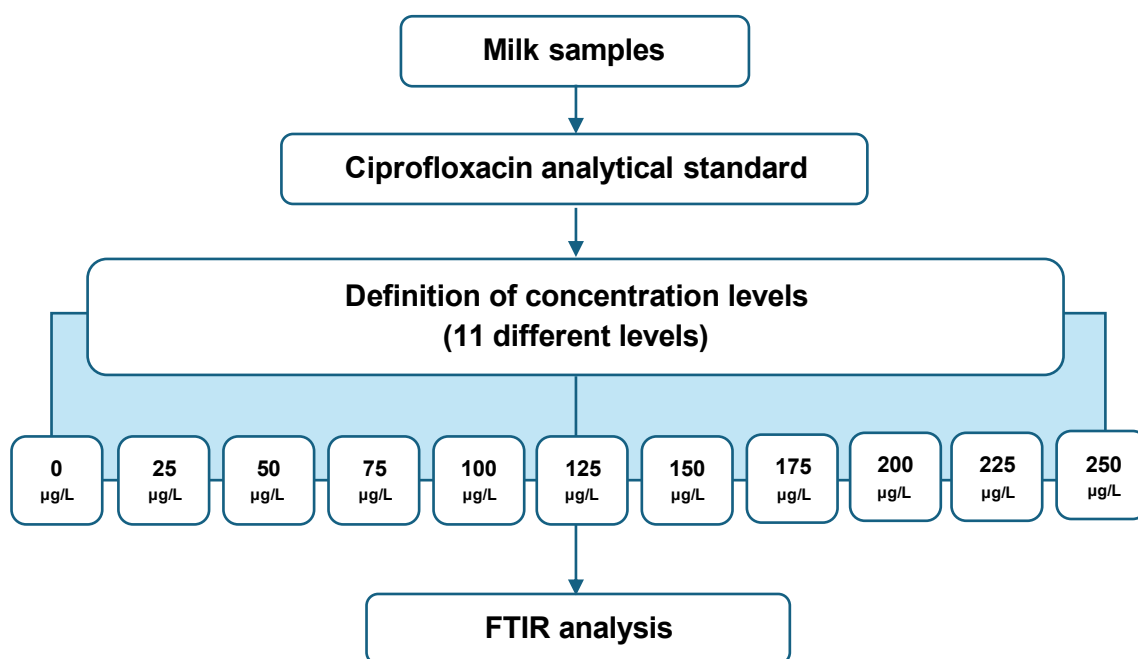


Figure 1. Experimental design for fortifying milk samples with ciprofloxacin

FTIR analysis

The spectra of the samples were obtained using Fourier Transform Infrared Spectroscopy with Attenuated Total Reflectance (FTIR-ATR), using a Cary 630 FTIR

spectrometer (Agilent Technologies Inc., Santa Clara, CA, USA). Analyses were conducted in the mid-infrared region, with a wavenumber range from 4000 cm^{-1} to 600 cm^{-1} , a resolution of 4 cm^{-1} , and 64 scans per sample, using a diamond crystal. Spectra were acquired in absorbance mode, and the maximum absorbance values associated with their respective wavenumber ranges were used as variables for statistical analysis and subsequent comparison of the spectroscopic profile. The graphs were generated using OriginPro 8.5 software for the characterization of the spectra.

Selection of spectral bands

Before analysis, the spectra were pre-processed by baseline correction using the asymmetric least squares method and Savitzky-Golay smoothing ¹⁶. Principal component analysis (PCA) was carried out using R software (version 4.4.3), where the data was standardized before analysis. The importance of each wave number was calculated by adding up the absolute values of the loadings of the first principal components that explained 95% of the total variation. For the random forest (RF) analysis, we used the Boruta method to choose the wave numbers. This method was done with the *Boruta* package in R. Before using Boruta, the data was normalized using mean and min-max scaling. The algorithm was run up to 100 times and the confidence level was set at 0.01. This classification compared importance with random attributes called “shadow features”. After selection by Boruta, a correlation analysis was carried out between the selected variables. This analysis eliminated redundant variables. Variables with a correlation greater than 0.8 were removed.

Predictive analytics

Partial Least Squares (PLS) regression analysis was carried out using the pls package in the R environment. To find the optimum number of components, 10-fold cross-validation was

carried out. The model was created using the spectral variables that had already been selected by PCA. For the RF regression, the randomForest package was used in the R environment. The model was trained with 500 decision trees. The wavelength variables were those selected by the Boruta method. In both cases, the data was randomly divided into training (70%) and test (30%) sets. Performance was evaluated using the test set. The Root Mean Square Error (RMSE), coefficient of determination (R^2) and Mean Absolute Error (MAE) metrics were used.

Classification analysis

Classification analyses were carried out using two methods, Linear Discriminant Analysis (LDA) and RF. For both, the data was divided into a training set (70%) and a test set (30%). The samples were classified into two categories: “Below the MRL” (≤ 100) and “Above the MRL” (> 100), according to the MRL established by Anvisa ¹⁷. The RF model was created using 500 decision trees. For the LDA, the chances of each class were calculated using Bayes' rule. The metrics for both models were calculated based on the confusion matrix. These measurements were made following previous work by Varoquaux and Colliot ¹⁸ and Hicco, Tötsch and Jurman ¹⁹.

$$\text{Recall} = \frac{TP}{TP + FN} \quad [1]$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad [2]$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad [3]$$

$$F1 - \text{score} = \frac{2TP}{2 \times TP + FP + FN} \quad [4]$$

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{specificity}}{2} \quad [5]$$

Where TP are the true positives, TN the true negatives, FP the false positives and FN the false negatives.

RESULTS AND DISCUSSION

Milk composition

The composition of the milk samples used in this study is shown in Table 1. The results show that the composition values are consistent with the typical ranges for bovine milk, although the fat content is slightly higher, which is in line with Jersey cows, known for producing milk with a higher fat concentration. The protein, lactose and total solids values are within the common ranges reported in commercial and regional studies, reflecting good nutritional quality ^{20,21}. The SCC value suggests moderate levels of somatic cells, however, it is below the quality standards established by Brazilian legislation for refrigerated raw milk ²².

Table 1. Average composition of bovine milk used in the study

Component	Mean $\pm$ SD
Fat	5.061 $\pm$ 0.463
Protein	3.751 $\pm$ 0.028
Lactose	4.129 $\pm$ 0.027
Total solids	13.929 $\pm$ 0.430
Cells (x1000)/mL	362.736 $\pm$ 59.604

Spectra in the mid-infrared

Figure 2 shows the mid-infrared absorption spectrum of ciprofloxacin. The peak at 1018 cm⁻¹ indicates stretching of the C-F group, common in fluoroquinolones, or C-O of a carboxyl or ether ^{23,24}. The peak at 1639 cm⁻¹ shows the stretching of the C=O group of the quinolone or the C=C group in the aromatic ring ^{23,25}. Small peaks were observed in the region near 1404 cm⁻¹, which may be associated with the asymmetric deformation of the carboxylate

group (COO⁻), but may also involve C-N or C-C bond vibrations of the quinolone ring or the piperazine ring^{23–25}. The peaks observed at 2843 and 2955 cm⁻¹ are both attributed to the C-H stretching of aliphatic groups, such as those present in the piperazine ring or side chains^{23,24}. The results are in line with the findings in the literature^{26–28}.

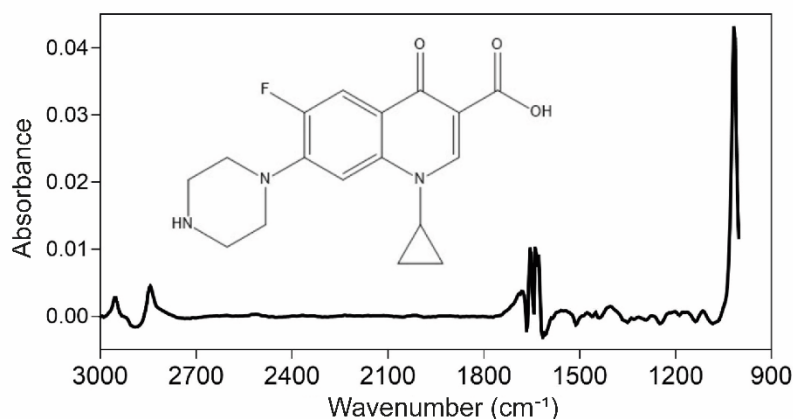


Figure 2. Mid-infrared absorption spectrum of ciprofloxacin

The bands in the FTIR spectrum of milk with different concentrations of ciprofloxacin show subtle changes in some regions (Figure 3). The bands between 1045-1076 cm⁻¹ are typical of carbohydrates, especially lactose, the main sugar in milk^{29,30}. The 1551 and 1647 cm⁻¹ regions are associated with amide II and I, both markers of proteins, mainly caseins and whey proteins. Amide III was also observed at 1273 cm⁻¹^{31,32}. The bands at 1458, 1747, 2854 and 2924 cm⁻¹ are associated with lipids^{33,34}. The band at 1246 cm⁻¹ suggests phosphoproteins³¹.

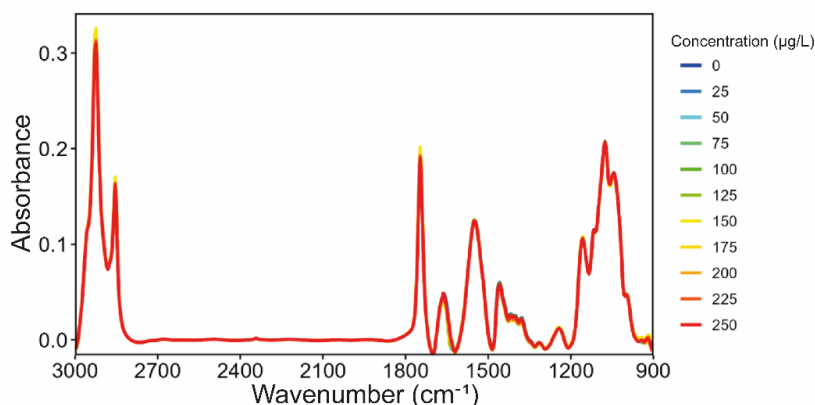


Figure 3. Mid-infrared spectrum of raw milk with different concentrations of ciprofloxacin (0 - 250 µg/L)

Variable selection

The PCA and Boruta methods were used to select variables. For PCA, the number of components needed to explain 70% of the variance was 13 principal components, so the variables were selected using the sum of the absolute loadings (Figure 4). For Boruta, the samples selected were those classified by the algorithm as confirmed. For the classification, 118 variables were confirmed, while for the regression, 137 variables were confirmed. To reduce the number of variables selected by Boruta and reduce the influence of collinearity, variables with a correlation above 0.8 were eliminated. In total, 8 variables were selected for further analysis.

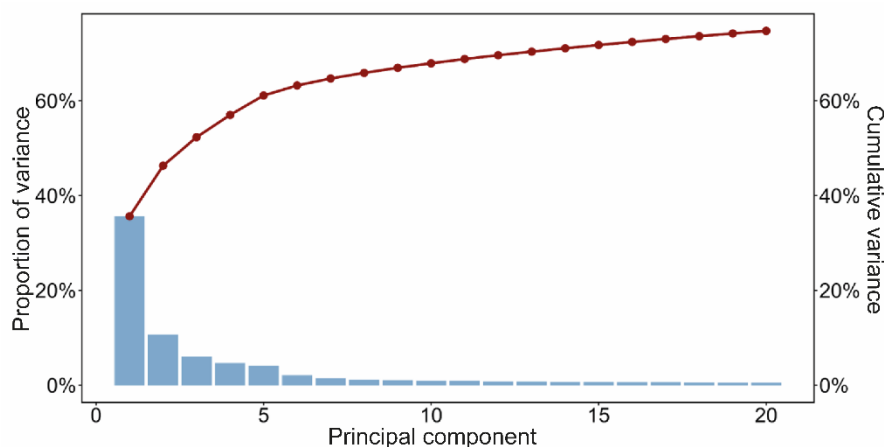


Figure 4. Variance explained by principal component of milk samples impregnated with ciprofloxacin at different concentrations.

The PCA and Boruta analysis identified wave numbers that correspond to the milk components and some specific ciprofloxacin markers (Table 2). One difference between the methods was the selection of carbonyl-related bands ($\text{C}=\text{O}$) at wave numbers $1639\text{--}1720\text{ cm}^{-1}$ ^{23, 25}. PCA identified this group, but this region was not selected by Boruta. Both methods detected changes in the aliphatic C-H stretching regions, $(2781\text{--}2970\text{ cm}^{-1})$ ^{23, 24}, however, each method chose different wavenumbers, which may be related to small changes in shape caused

by ciprofloxacin. Both approaches showed that the regions involving proteins (amides I, II and III) and carbohydrates.

Table 2. Set of wave numbers selected by PCA and Boruta for milk samples containing different concentrations of ciprofloxacin

PCA	Boruta	Functional Group
1045, 1076, 1095, 1119, 1138, 1142	1126, 1192	C-O stretch (lactose) / C-F ^{23, 24, 29, 30}
1246, 1273	1273	Amide III (phosphoproteins, caseins) / C-N ^{25, 31, 32}
1404, 1447	1466	CH ₂ /CH ₃ deformations (lipids) / COO ⁻ asymmetric ^{33, 34}
1551, 1593, 1612	1616, 1620, 1628	Amide I, II (proteins) / aromatic C=C ^{31, 32}
1639, 1686, 1697	—	C=O (quinolone, esters, fatty acids) ^{23, 25}
2843, 2854, 2924, 2955	2797, 2970	Aliphatic C-H stretches ^{23, 33, 34}

The PCA and Boruta methods identify different numbers of waves in FTIR spectra because they use different approaches to choose the most relevant variables. PCA is an unsupervised method that seeks to maximize the variance of the data by identifying linear combinations of variables that explain most of the variation in the spectrum, without considering classes or concentrations^{35, 36}. Boruta, on the other hand, is a supervised method based on RF, which selects variables based on their importance in predicting a target variable³⁷. Studies using FTIR and variable selection methods, such as that by Freitas et al.³⁸, shows that supervised methods find specific spectral bands.

PCA, which does not use supervision, chooses regions with greater overall variation. These regions are related to milk components. Boruta, on the other hand, focuses on bands that help separate classes or indicate antibiotic concentration, even if they have less variation^{39, 40}. Therefore, in our study, PCA mainly highlighted the regions of 1095–1142 cm⁻¹ and 1593–1697 cm⁻¹, regions associated with phosphate groups, sugars, and amides. Boruta selected bands in other regions, including wave numbers below 1000 cm⁻¹ and above 2700 cm⁻¹. The variability observed in milk composition, mainly in fat (5.061 ± 0.463%), may have generated variations

in the spectra in regions characteristic of aliphatic groups. In PCA, which identifies directions of maximum variance in data in an unsupervised manner, these lipid variations likely determined the structure of the first principal components^{39,41}. Consequently, some of the spectral variables most strongly associated with fat, such as the bands at 1447, 2843, and 2955 cm^{-1} . In contrast, the Boruta algorithm, being supervised, attempted to select variables correlated with the concentration of ciprofloxacin, even in the presence of this matrix variability. However, it is possible that spectral interference from lipids may have overlapped the peaks associated with the antibiotic, such as the 1639–1697 cm^{-1} range, a region attributed to the C=O bond of the quinolone structure. Therefore, it is possible that Boruta may have interpreted these regions as uninformative or dominated by noise, leading to their exclusion^{40,42}.

Predictive models

The predictive performance of the RF and PLS regression models in quantifying ciprofloxacin in liquid milk samples, with concentrations ranging from 0 to 250 $\mu\text{g/L}$, indicates that the RF model employed the variables chosen by Boruta and the PLS model used the variables chosen by PCA (Figure 5). RF showed less variation in points compared to the PLS model. This result suggests that, in the quantification of ciprofloxacin in milk, the application of nonlinear machine learning methods, such as RF, demonstrates superior predictive potential compared to traditional linear methods such as PLS. However, the results achieved for the PLS and RF regression models in the quantification of ciprofloxacin in liquid milk showed limited performance (Table 3).

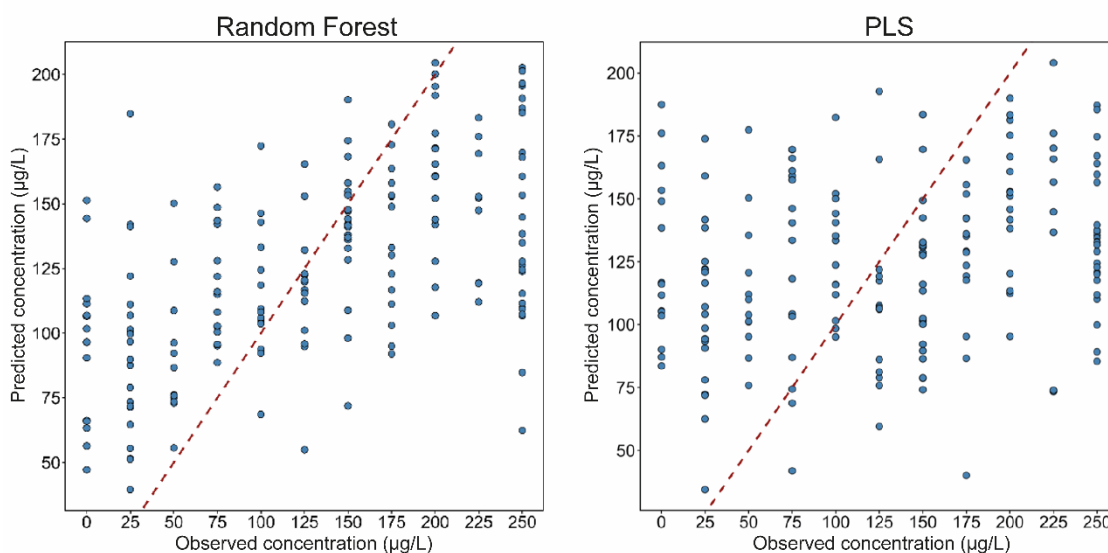


Figure 5. Predictive performance of random forest regression and PLS for quantifying ciprofloxacin at concentrations of 0–250 µg/L in liquid milk

The RF model performed better than PLS, as lower mean square errors and mean absolute errors were observed. Even so, both models had low coefficients of determination (R^2). In the RF model, it was 0.325, and for PLS, it was 0.062. These figures show that the models have little ability to explain variations in the data. Even so, RF seems to better predict the results. However, both models have difficulties accurately measuring ciprofloxacin in milk. A study conducted by Pandey et al.⁴³ shows that PLS can achieve high accuracy, with R^2 values above 0.99, when measuring residues such as ciprofloxacin in aqueous solutions. RF also performed well in predicting residues with concentrations up to 100 µg/L in freeze-dried milk samples, with R^2 above 0.95⁴⁴. However, the same precision was not achieved when measuring ciprofloxacin in liquid milk using these methods in the present study. The low precision may have been caused by the complexity of milk, spectral interference, variation between samples, and the need to carefully select the important spectral variables^{44, 45}.

The variability in milk composition may have limited the performance of the prediction models. In PLS, whose construction is based on maximizing the covariance between the spectra and the analyte concentration³⁴, This variation related to the presence of other milk components

may have impaired the relationship between the signal and the concentration of ciprofloxacin, resulting in a much lower coefficient of determination. RF, although less sensitive to noise and capable of capturing nonlinear relationships, was also affected. The variable selection step with Boruta may have excluded potential ciprofloxacin bands due to signals caused by the variability of the lipid fraction of milk. This exclusion may have limited the model's ability to capture subtle patterns associated with the analyte, compromising its generalization ability.

Table 3. Performance metrics of PLS and Random Forest regression models in the quantification of ciprofloxacin in liquid milk

Metric	PLS	Random forest
RMSE	80.58	68.15
MAE	68.26	54.18
R ²	0.062	0.325

Classification models

For classification analysis, milk samples were divided according to the maximum residue limit (MRL) for ciprofloxacin defined by Anvisa ¹⁷. The groups were: “Below the MRL” ($\leq 100 \mu\text{g/L}$) and “Above the MRL” ($> 100 \mu\text{g/L}$). The classification results are shown in Table 4. The RF method performed well and was correct in about 76% of the classifications, in addition to high precision, specificity, and recall. These results indicate that the RF method can accurately identify positive samples and also minimize errors when identifying negative ones. The F1-Score of 0.784 shows that the method has a good overall balance. In comparison, the LDA method had lower results in all metrics. The higher accuracy of RF is because this method can better handle complex data that does not follow a straight line and has a lot of different information, such as FTIR spectra. It is also more resistant to noise ³⁸. In contrast, LDA assumes linear relationships between variables and can be affected by multicollinearity and irregular data distribution. Therefore, this analysis performs less well on complex data sets

and is less accurate ⁴⁶. When compared with regression results, classification models performed better. This shows that the use of qualitative methods may be more suitable for detecting the presence of ciprofloxacin in milk.

Table 4. Performance metrics of LDA and RF models in the classification of milk samples containing ciprofloxacin

Metric	LDA	Random Forest
Balanced accuracy	0.239	0.761
Precision	0.212	0.780
Recall	0.263	0.788
Specificity	0.220	0.728
F1-Score	0.235	0.784

CONCLUSION

The application of FTIR spectroscopy associated with machine learning algorithms can aid in the analysis of ciprofloxacin in cow's milk. The models developed to evaluate the amount of ciprofloxacin did not perform satisfactorily. However, the models used to classify samples performed better, especially the random forest model, which can accurately differentiate milk samples with values above and below the MRL. This demonstrates that this approach can serve as a screening method for detecting ciprofloxacin residues in milk.

AUTHOR CONTRIBUTIONS

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Romicy Dermondes Souza: Writing – original draft, Formal analysis, Investigation.

Rosângela Silveira Barbosa: Resources, Supervision.

Marcelo Bonnet Alvarenga: Data Curation, Validation.

- 1 ***Fabiano Barreto***: Writing – review & editing.
- 2 ***Wellington da Silva Lyra***: Investigation.
- 3 ***Sérgio de Sousa Castro***: Methodology. Conceptualization
- 4 ***Paulo Henrique Gonçalves Dias Diniz***: Methodology, Writing – review & editing.
- 5 ***Sérgio Augusto de Albuquerque Fernandes***: Supervision, Project administration, Writing –
- 6 review & editing, Conceptualization.

7

8 **ACKNOWLEDGMENT**

9 This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível
 10 Superior – Brasil (CAPES) – Finance Code 001. The authors would also like to thank Empresa
 11 Brasileira de Pesquisa Agropecuária (Embrapa) and the Ministério da Agricultura, Pecuária e
 12 Abastecimento (MAPA) for their support.

13

14 **ABBREVIATIONS**

15

16 FTIR, Fourier transform infrared spectroscopy; MIR, mid-infrared; ATR, attenuated total
 17 reflectance; PLS, partial least squares; RF, random forest; PCA, principal component analysis;
 18 LDA, linear discriminant analysis; MRL, maximum residue limit; SD, standard deviation;
 19 RMSE, root mean square error; MAE, mean absolute error; R^2 , coefficient of determination;
 20 SCC, somatic cell count; ANVISA, Agência Nacional de Vigilância Sanitária; CAPES,
 21 Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; MAPA, Ministério da
 22 Agricultura, Pecuária e Abastecimento; Embrapa, Empresa Brasileira de Pesquisa
 23 Agropecuária.

24

25 **REFERENCES**

(1) Lubna; Hussain, T.; Shami, A.; Rafiq, N.; Khan, S.; Kabir, M.; Khan, N. U.; Khattak, I.; Kamal, M.; Usman, T. Antimicrobial Usage and Detection of Multidrug-Resistant *Staphylococcus Aureus*: Methicillin- and Tetracycline-Resistant Strains in Raw Milk of Lactating Dairy Cattle. *Antibiotics*, **2023**, *12* (4), 673. DOI: 10.3390/antibiotics12040673.

(2) Nazarov, M.; Popovich, E.; Mashyanova, S. Medical Prevention of Mastitis in Cows Two Months Before Calving in a Sustainable Agriculture; 2022; pp 330–339. DOI: 10.1007/978-3-030-91405-9_36.

(3) Rame Hau, E. E. R.; Sharma, M.; Khanal, B. K. S.; Sly, P. D.; Mikkelsen, D.; Clark, N.; Soares Magalhães, R. J. S. The Prevalence and Risk Factors Associated with the Presence of Antibiotic Residues in Milk from Peri-Urban Dairy Cattle Farms in Kathmandu, Nepal. *Antibiotics*, **2025**, *14* (1), 98. DOI: 10.3390/antibiotics14010098.

(4) Rahman, A.; Rinky, F.; Rahman, A.; Mousume, S. N.; Das, R.; Sharmin, T.; Sohag, M. H.; Bari, M. L. Risk Assessment of Antibiotic Residues in Raw Cow's Milk in Bangladesh. *Food Addit. Contam. Part B*, **2025**, *18* (2), 99–110. DOI: 10.1080/19393210.2024.2433945.

(5) Joubrane, K.; Jammoul, A.; Daher, R.; Ayoub, S.; El Jed, M.; Hneino, M.; El Hawari, K.; Al Iskandarani, M.; Daher, Z. Microbiological Contamination, Antimicrobial Residues, and Antimicrobial Resistance in Raw Bovine Milk in Lebanon. *Int. Dairy J.*, **2022**, *134*, 105455. DOI: 10.1016/j.idairyj.2022.105455.

(6) Tóth, A. G.; Csabai, I.; Krikó, E.; Tőzsér, D.; Maróti, G.; Patai, Á. V.; Makrai, L.; Szita, G.; Solymosi, N. Antimicrobial Resistance Genes in Raw Milk for Human Consumption. *Sci. Rep.*, **2020**, *10* (1), 7464. DOI: 10.1038/s41598-020-63675-4.

(7) de Faria, L. V.; Lisboa, T. P.; Alves, G. F.; de Farias, D. M.; Matos, M. A. C.; Muñoz, R. A. A.; Matos, R. C. Electrochemical Study of Different Sensors for Simple and Fast Quantification of Ciprofloxacin in Pharmaceutical Formulations and Bovine Milk. *Electroanalysis*, **2020**, *32* (10), 2266–2272. DOI: 10.1002/elan.202060211.

- (8) Liu, L.; Luo, L.; Suryoprabowo, S.; Peng, J.; Kuang, H.; Xu, C. Development of an Immunochromatographic Strip Test for Rapid Detection of Ciprofloxacin in Milk Samples. *Sensors*, **2014**, *14* (9), 16785–16798. DOI: 10.3390/s140916785.
- (9) Ribeiro da Cunha, B.; Fonseca, L. P.; Calado, C. R. C. Metabolic Fingerprinting with Fourier-Transform Infrared (FTIR) Spectroscopy: Towards a High-Throughput Screening Assay for Antibiotic Discovery and Mechanism-of-Action Elucidation. *Metabolites*, **2020**, *10* (4), 145. DOI: 10.3390/metabo10040145.
- (10) Kurrey, R.; Mahilang, M.; Kanti Deb, M.; Nirmalkar, J.; Shrivastava, K.; Pervez, S.; Kumar Rai, M.; Rai, J. A Direct DRS-FTIR Probe for Rapid Detection and Quantification of Fluoroquinolone Antibiotics in Poultry Egg-Yolk. *Food Chem.*, **2019**, *270*, 459–466. DOI: 10.1016/j.foodchem.2018.07.129.
- (11) Kochan, K.; Jiang, J.-H.; Kostoulas, X.; Lai, E.; Richardson, Z.; Pebotuwa, S.; Heraud, P.; Wood, B. R.; Peleg, A. Y. Fast and Accurate Prediction of Antibiotic Susceptibility in Clinical Methicillin-Resistant *S. Aureus* Isolates Using ATR-FTIR Spectroscopy: A Model Validation Study. *Anal. Chem.*, **2025**, *97* (11), 6041–6048. DOI: 10.1021/acs.analchem.4c06086.
- (12) Ribeiro da Cunha, B.; Fonseca, L. P.; Calado, C. R. C. Simultaneous Elucidation of Antibiotic Mechanism of Action and Potency with High-Throughput Fourier-Transform Infrared (FTIR) Spectroscopy and Machine Learning. *Appl. Microbiol. Biotechnol.*, **2021**, *105* (3), 1269–1286. DOI: 10.1007/s00253-021-11102-7.
- (13) Yang, Z.; Chen, G.; Ma, C.; Gu, J.; Zhu, C.; Li, L.; Gao, H. Magnetic Fe₃O₄@COF@Ag SERS Substrate Combined with Machine Learning Algorithms for Detection of Three Quinolone Antibiotics: Ciprofloxacin, Norfloxacin and Levofloxacin. *Talanta*, **2023**, *263*, 124725. DOI: 10.1016/j.talanta.2023.124725.
- (14) de Freitas, A. G.; de Magalhães, B. E.; Minho, L. A.; Leão, D. J.; Santos, L. S.; Augusto de Albuquerque Fernandes, S. FTIR Spectroscopy with Chemometrics for Determination of Tylosin Residues in Milk. *J. Sci. Food Agric.*, **2021**, *101* (5), 1854–1860. DOI:

- 1 10.1002/jsfa.10799.
- 2 (15) Vali Zade, S.; Neymeyr, K.; Sawall, M.; Khoshayand, M. R.; Abdollahi, H.
- 3 Identification and Quantification of Ciprofloxacin, Enrofloxacin and Difloxacin Residues in
- 4 Milk Using Excitation-Emission Fluorescence and Second-Order Standard Addition Methods.
- 5 *Chemom. Intell. Lab. Syst.*, **2023**, 242, 104988. DOI: 10.1016/j.chemolab.2023.104988.
- 6 (16) Dong, Z.; Xu, J. Baseline Estimation Using Optimized Asymmetric Least Squares (O-
- 7 ALS). *Measurement*, **2024**, 233, 114731. DOI: 10.1016/j.measurement.2024.114731.
- 8 (17) ANVISA. *Instrução Normativa Nº 162, de 1º de Julho de 2022. Estabelece a Ingestão*
- 9 *Diária Aceitável (IDA), a Dose de Referência Aguda (DRfA) e Os Limites Máximos de Resíduos*
- 10 *(LMR) Para Insumos Farmacêuticos Ativos (IFA) de Medicamentos Veterinários Em Alimentos*
- 11 *De*; Diário Oficial da União: seção 1: Brasília, DF, Brasil, 2022; p 146.
- 12 (18) Varoquaux, G.; Colliot, O. Evaluating Machine Learning Models and Their Diagnostic
- 13 Value; 2023; pp 601–630. DOI: 10.1007/978-1-0716-3195-9_20.
- 14 (19) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is
- 15 More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-
- 16 Class Confusion Matrix Evaluation. *BioData Min.*, **2021**, 14 (1), 13. DOI: 10.1186/s13040-021-
- 17 00244-z.
- 18 (20) Bonifaz-García, N. F.; Argüello–Argüello, A. F.; Arias–Sandoval, C. A.; Simbaña–Díaz,
- 19 P. E. Compositional, Hygienic and Sanitary Quality of Bovine Milk in Three Regions of
- 20 Ecuador. *Rev. la Fac. Med. Vet. y Zootec.*, **2025**, 72 (1). DOI: 10.15446/rfmvz.v72n1.114348.
- 21 (21) Foroutan, A.; Guo, A. C.; Vazquez-Fresno, R.; Lipfert, M.; Zhang, L.; Zheng, J.; Badran,
- 22 H.; Budinski, Z.; Mandal, R.; Ametaj, B. N.; et al. Chemical Composition of Commercial Cow's
- 23 Milk. *J. Agric. Food Chem.*, **2019**, 67 (17), 4897–4914. DOI: 10.1021/acs.jafc.9b00204.
- 24 (22) MAPA. *INSTRUÇÃO NORMATIVA Nº 76, DE 26 DE NOVEMBRO DE 2018*; 2018.
- 25 (23) SHADIQUE, M. A.; PRATIK, P.; SINGH, P. K. Synthesis, Structural Characterization
- 26 and Antibacterial Studies of Fluoroquinolone Drug-Ciprofloxacin. *J. Ultra Chem.*, **2020**, 16 (5),

58–62. DOI: 10.22147/juc/160502.

(24) Tozar, T.; Boni, M.; Staicu, A.; Pascu, M. L. Optical Characterization of Ciprofloxacin Photolytic Degradation by UV-Pulsed Laser Radiation. *Molecules*, **2021**, 26 (8), 2324. DOI: 10.3390/molecules26082324.

(25) Kumari, P.; Yadav, A. K. Structural Characterization of Some Fluoroquinolone Drugs. *J. Res. Chem.*, **2025**, 6 (1), 205–207. DOI: 10.22271/reschem.2025.v6.i1c.193.

(26) Sundaram, N. M.; Rajendran, N. Biodegradation and Cytotoxicity of Ciprofloxacin-Loaded Hydroxyapatite-Polycaprolactone Nanocomposite Film for Sustainable Bone Implants. *Int. J. Nanomedicine*, **2015**, 119. DOI: 10.2147/IJN.S79995.

(27) Rajalakshmi, K.; Gunasekaran, S.; Kumaresan, S. Vibrational Spectra, Electronic and Quantum Mechanical Investigations on Ciprofloxacin. *Indian J. Phys.*, **2014**, 88 (7), 733–744. DOI: 10.1007/s12648-014-0468-8.

(28) Sahoo, S.; Chakraborti, C. K.; Mishra, S. C.; Nanda, U. N.; Naik, S. FTIR and XRD Investigations of Some Fluoroquinolones. *Int. J. Pharm. Pharm. Sci.*, **2011**, 3 (3), 165–170.

(29) Lafi, A. G. Al; Naser, I.; Abboud, H. Application of Two Dimensional Correlation Fourier Transform Infrared Spectroscopy (2D-COS-FTIR) to the Quantification of Carbohydrates in Milk Powder. *J. Food Compos. Anal.*, **2024**, 134, 106541. DOI: 10.1016/j.jfca.2024.106541.

(30) Balan, B.; Dhaulaniya, A. S.; Jamwal, R.; Yadav, A.; Kelly, S.; Cannavan, A.; Singh, D. K. Rapid Detection and Quantification of Sucrose Adulteration in Cow Milk Using Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy Coupled with Multivariate Analysis. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, **2020**, 240, 118628. DOI: 10.1016/j.saa.2020.118628.

(31) Kristoffersen, K. A.; Afseth, N. K.; Böcker, U.; Lindberg, D.; de Vogel-van den Bosch, H.; Ruud, M. L.; Wubshet, S. G. Average Molecular Weight, Degree of Hydrolysis and Dry-Film FTIR Fingerprint of Milk Protein Hydrolysates: Intercorrelation and Application in

- 1 Process Monitoring. *Food Chem.*, **2020**, *310*, 125800. DOI: 10.1016/j.foodchem.2019.125800.
- 2 (32) Daniloski, D.; McCarthy, N. A.; O'Callaghan, T. F.; Vasiljevic, T. Authentication of β -
3 Casein Milk Phenotypes Using FTIR Spectroscopy. *Int. Dairy J.*, **2022**, *129*, 105350. DOI:
4 10.1016/j.idairyj.2022.105350.
- 5 (33) Toledo-Alvarado, H. O.; Tempelman, R. J.; Lopez-Cruz, M.; VandeHaar, M. J.; Santos,
6 J. E. P.; Peñagaricano, F.; Khanal, P.; de los Campos, G. Selection Indexes for Residual Feed
7 Intake Derived from Milk Spectra. *J. Dairy Sci.*, **2024**, *107* (12), 11164–11179. DOI:
8 10.3168/jds.2023-24425.
- 9 (34) Rachah, A.; Reksen, O.; Tafintseva, V.; Stehr, F. J. M.; Rukke, E.-O.; Prestløkken, E.;
10 Martin, A.; Kohler, A.; Afseth, N. K. Exploring Dry-Film FTIR Spectroscopy to Characterize
11 Milk Composition and Subclinical Ketosis throughout a Cow's Lactation. *Foods*, **2021**, *10* (9),
12 2033. DOI: 10.3390/foods10092033.
- 13 (35) Amit; Jamwal, R.; Kumari, S.; Dhaulaniya, A. S.; Balan, B.; Singh, D. K. Application
14 of ATR-FTIR Spectroscopy along with Regression Modelling for the Detection of Adulteration
15 of Virgin Coconut Oil with Paraffin Oil. *LWT*, **2020**, *118*, 108754. DOI:
16 10.1016/j.lwt.2019.108754.
- 17 (36) Amit; Jamwal, R.; Kumari, S.; Kelly, S.; Cannavan, A.; Singh, D. K. Rapid Detection
18 of Pure Coconut Oil Adulteration with Fried Coconut Oil Using ATR-FTIR Spectroscopy
19 Coupled with Multivariate Regression Modelling. *LWT*, **2020**, *125*, 109250. DOI:
20 10.1016/j.lwt.2020.109250.
- 21 (37) Chen, L.; Yang, S.; Nan, Z.; Li, Y.; Ma, J.; Ding, J.; Lv, Y.; Yang, J. Detection of Dextran,
22 Maltodextrin and Soluble Starch in the Adulterated Lycium Barbarum Polysaccharides (LBPs)
23 Using Fourier-Transform Infrared Spectroscopy (FTIR) and Machine Learning Models.
24 *Heliyon*, **2023**, *9* (6), e17115. DOI: 10.1016/j.heliyon.2023.e17115.
- 25 (38) Freitas, A. G. M. de; Almir Cavalcante Minho, L.; Elizabeth Alves de Magalhães, B.;
26 Nei Lopes dos Santos, W.; Soares Santos, L.; Augusto de Albuquerque Fernandes, S. Infrared

- 1 Spectroscopy Combined with Random Forest to Determine Tylosin Residues in Powdered
2 Milk. *Food Chem.*, **2021**, *365*, 130477. DOI: 10.1016/j.foodchem.2021.130477.
- 3 (39) Šašić, S.; Ozaki, Y. Comparison of Principal Component Analysis and Generalized
4 Two-Dimensional Correlation Spectroscopy: Spectral Analysis of Synthetic Model System and
5 Near-Infrared Spectra of Milk. *Appl. Spectrosc.*, **2001**, *55* (1), 29–38. DOI:
6 10.1366/0003702011951399.
- 7 (40) Kursá, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.*,
8 **2010**, *36* (11). DOI: 10.18637/jss.v036.i11.
- 9 (41) Santos, R. de O.; Gorgulho, B. M.; Castro, M. A. de; Fisberg, R. M.; Marchioni, D. M.;
10 Baltar, V. T. Principal Component Analysis and Factor Analysis: Differences and Similarities
11 in Nutritional Epidemiology Application. *Rev. Bras. Epidemiol.*, **2019**, *22*. DOI: 10.1590/1980-
12 549720190041.
- 13 (42) Masrur Ahmed, A. A.; Deo, R. C.; Feng, Q.; Ghahramani, A.; Raj, N.; Yin, Z.; Yang, L.
14 Deep Learning Hybrid Model with Boruta-Random Forest Optimiser Algorithm for Streamflow
15 Forecasting with Climate Mode Indices, Rainfall, and Periodicity. *J. Hydrol.*, **2021**, *599*,
16 126350. DOI: 10.1016/j.jhydrol.2021.126350.
- 17 (43) Pandey, S.; Pandey, P.; Tiwari, G.; Tiwari, R.; Rai, A. FTIR Spectroscopy: A Tool for
18 Quantitative Analysis of Ciprofloxacin in Tablets. *Indian J. Pharm. Sci.*, **2012**, *74* (1), 86. DOI:
19 10.4103/0250-474X.102551.
- 20 (44) Freitas, A. G. M. de; Minho, L. A. C.; Magalhães, B. E. A. de; Santos, W. N. L. dos;
21 Santos, L. S.; Fernandes, S. A. de A. Infrared Spectroscopy Combined with Random Forest to
22 Determine Tylosin Residues in Powdered Milk. *Food Chem.*, **2021**, *365*, 130477. DOI:
23 10.1016/j.foodchem.2021.130477.
- 24 (45) Li, H.; Sheng, W.; Hassan, M. M.; Geng, W.; Chen, Q. Quantification of Antibiotics in
25 Food by Octahedral Gold-Silver Nanocages-Based SERS Sensor Coupling Multivariate
26 Calibration. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, **2024**, *320*, 124595. DOI:

1 10.1016/j.saa.2024.124595.

2 (46) Zhu, F.; Gao, J.; Yang, J.; Ye, N. Neighborhood Linear Discriminant Analysis. *Pattern*
3 *Recognit.*, **2022**, *123*, 108422. DOI: 10.1016/j.patcog.2021.108422.

4

IV – CAPÍTULO II

ARTIGO 2 - Evaluation of linear and nonlinear models for prediction and classification of somatic cell count in bovine milk using FTIR spectroscopy

1 EVALUATION OF LINEAR AND NONLINEAR
2 MODELS FOR PREDICTION AND
3 CLASSIFICATION OF SOMATIC CELL
4 COUNT IN BOVINE MILK USING FTIR
5 SPECTROSCOPY

6 *Romicy Dermondes Souza¹, Rosângela Silveira Barbosa², Marcelo Bonnet Alvarenga²,*
7 *Fabiano Barreto³, Wellington da Silva Lyra⁴, Sérgio de Sousa Castro¹, Paulo Henrique*
8 *Gonçalves Dias Diniz⁵, Sérgio Augusto de Albuquerque Fernandes^{1*}*

9 ¹Department of Rural and Animal Technology, State University of Southwest Bahia, 45700-
10 000, Bahia, Itapetinga, Brazil

11 ²Embrapa Temperate Climate, 96010-971, Rio Grande do Sul, Pelotas, Brazil

12 ³Federal Agricultural Defense Laboratory of Rio Grande do Sul, Ministry of Agriculture,
13 Livestock and Supply, 91780-580, Rio Grande do Sul, Porto Alegre, Brazil

14 ⁴Department of Chemistry, Center for Exact and Nature Sciences, Federal University of Paraíba
15 - Campus I, 58051-900, Paraíba, João Pessoa, Brazil

16 ⁵Post-Graduate Program in Pure and Applied Chemistry, Federal University of Western Bahia,
17 47810-059, Bahia, Barreiras, Brazil

18 *Corresponding author.

19 E-mail: sfernandes@uesb.edu.br

1 Contributing authors: dermondesromicy@gmail.com; rosangela.barbosa@embrapa.br;
2 marcelo.bonnet@embrapa.br; Fabiano.barreto@agro.gov.br; spectru@gmail.com;
3 scastro@uesb.edu.br; paulo.diniz@ufob.edu.br; sfernandes@uesb.edu.br.

4

5

6

1 **ABSTRACT**

2 The somatic cell count (SCC) is a key indicator of subclinical mastitis in dairy cows, a condition
3 that compromises milk quality and generates economic losses. Although flow cytometry is the
4 standard method for its quantification, FTIR spectroscopy emerges as a fast and low-cost
5 alternative. However, its effectiveness in heterogeneous populations without strict control of
6 interfering factors remains uncertain. Therefore, this study aimed to compare linear (PLS and
7 LDA) and nonlinear (Random Forest) models for the prediction and classification of SCC in
8 bovine milk samples using FTIR data. A total of 624 milk samples with a wide range of SCC
9 were analyzed. The FTIR spectra were preprocessed and subjected to variable selection by PCA
10 (unsupervised) and Boruta (supervised). PLS and Random Forest models were fitted for
11 regression, while LDA and Random Forest were used for binary classification ($\text{SCC} \geq 500,000$
12 cells/mL). Performance was evaluated using metrics such as R^2 , RMSE, sensitivity, specificity,
13 and F1-score, with 10 repeated train/test splits. Both regression models showed low R^2 (<0.30)
14 and high RMSE, indicating limitations in quantifying SCC. For classification, LDA achieved
15 higher sensitivity (0.795), while Random Forest showed higher specificity (0.908) and
16 precision (0.761). FTIR does not allow for precise quantification of SCC in heterogeneous
17 samples, but it may be useful as an initial screening tool when combined with discriminative
18 models, especially in routine laboratory settings where equipment is already available.

19 **KEYWORDS**

20 Boruta, Principal Component Analysis, Subclinical mastitis, Random Forest, Chemometrics

1 INTRODUCTION

2
3 Mastitis in cows is an inflammation of the mammary gland that affects the health of the
4 udder, the quality and quantity of milk produced, causing significant economic losses in the
5 dairy industry. The most common form is subclinical mastitis, which has no visible clinical
6 signs but reduces milk production and increases somatic cell count (SCC), an important
7 indicator of infection ¹. SCC refers to the number of somatic cells, mainly leukocytes and some
8 epithelial cells, present in 1 mL of milk. Elevated values indicate inflammation of the udder,
9 usually associated with mastitis ².

10 Several methods are used to quantify SCC, such as direct microscopy, biosensors and
11 flow cytometry, the most widely used method, which despite being fast and accurate, has high
12 cost of operation, in addition to requiring sophisticated equipment, specialized maintenance and
13 expensive reagents ³. In this context, middle infrared spectroscopy (MIR) appears as a
14 promising alternative to estimate SCC in an indirect, fast, non-destructive and low-cost way ⁴.
15 FTIR is an analytical technique widely used in milk quality control laboratories because it
16 allows the simultaneous quantification of several milk components, such as fat, protein, lactose
17 and urea, from a single spectrum. The technique is based on absorption of infrared radiation by
18 functional molecular groups, generating a chemical fingerprint of the sample. In addition, FTIR
19 spectra generate a large volume of multivariate data, which, when combined with advanced
20 variable selection and predictive modeling methods, allows to extract relevant information even
21 in the absence of specific spectral peaks for somatic cells ^{4, 5}.

22 Some studies ^{6, 7} sought to predict SCC indirectly through changes in biochemical
23 composition, but showed limited performance in precise and robust quantification, especially
24 in heterogeneous samples and without strict control of interfering factors. The literature shows
25 that models based on partial least squares regression (PLS) are predominant ^{6, 8} and some more
26 recent studies have incorporated somatic cell scores or phenotypic variables to improve mastitis

detection⁹. However, none of the studies explored a systematic comparison between variable selection approaches based on supervised criteria, such as the Boruta algorithm, a variable selection method based on Random Forest (RF) that evaluates the direct predictive importance of each wave number and the use of non-linear predictive methods, such as RF itself. RF is a machine learning algorithm used for classification and regression tasks that consists in building a forest of several trained decision trees on random subsets of the data. To make a forecast, the RF aggregates the results of all trees, where in the case of classification, it ranks by majority vote and in the case of regression, it determines the value by the average of the forecasts^{10,11}. The RF is robust to unbalanced data, missing values and can provide measures of importance of variables, assisting in the interpretation of the model¹². Because of its efficiency, versatility and accuracy, Random Forest is widely used in several areas, from financial forecasting to biomedical analysis^{13,14}. Based on the above, this study aims to evaluate and compare the performance of RF, a non-linear model, with linear models, in the prediction and classification of bovine milk samples containing different levels of SCC, using spectroscopy data in mid-infrared.

MATERIALS AND METHODS

Collection of samples

The milk samples used in this study were obtained at the Experimental Station Terras Baixas (ETB) of Embrapa Clima Temperado, located in Capão do Leão, Rio Grande do Sul (31°49'27" S and 52°26'35" O). The collection was planned to compose a diverse set of samples, with wide variation in SCC levels, an essential condition for the development and robust validation of predictive models. For this, jersey cows with different levels of breast health were included, from clinically healthy animals to animals with subclinical and clinical mastitis,

ensuring a comprehensive representation of the SCC interval observed in real milk production. All animals had remained without antibiotic treatment in the 30 days prior to collection. Before milking, the teats were sanitized with iodine solution and dried with disposable paper towel. The first milk jets were discarded and samples collected manually by a trained milker, with strict hand hygiene. Immediately after collection, the samples were packaged in thermal boxes maintained at 5 °C ($\pm$ 5 °C) and transported to the laboratory for analysis. A total of 624 milk samples were collected.

Analysis of samples

The SCC analysis was done by flow cytometry method, using automated equipment SomaScope (Delta Instruments, Drachten, Netherlands). The infrared spectra were obtained using the Fourier transform infrared spectroscopy (FTIR) LactoScope FTIR (Pertin Instruments, Hägersten, Sweden). The composition data were obtained by the same equipment. The spectra were obtained in the spectral range from 1000 to 3000 cm^{-1} , with a spectral resolution of 4 cm^{-1} . The spectra, after obtaining, were pre-processed by baseline correction by the method of asymmetric least squares and Savitzky-Golay smoothing. The process flowchart, from sample collection to analysis, is described in a simplified manner in Figure 1.

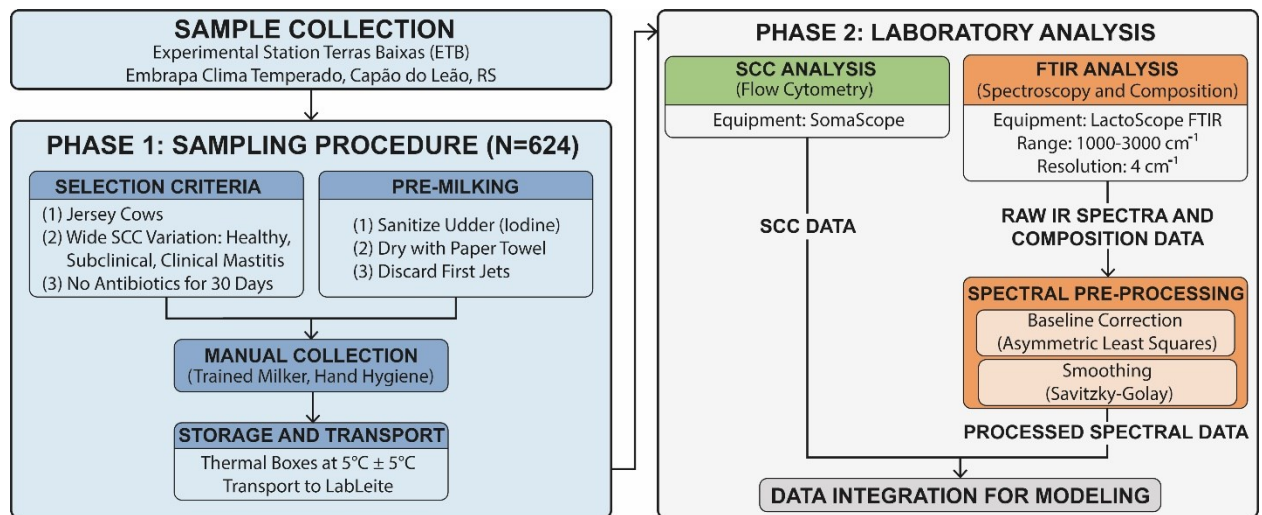


Figure 1. Flowchart of the experimental protocol for sample collection and laboratory analysis of milk samples

Selection of variables

Two methods were used for selection of variables, the PCA and Boruta. For both methods, we used the standardization of spectral data by centralization in the mean followed by min-max scaling. For the PCA, the importance of each wave number was quantified by the sum of the squares of the main component loadings that accumulate at least 70% of the explained variance. From this metric, we selected a subset of wave numbers that represent 80% of the importance. For Boruta, a configuration was performed with a significance level of 1% and a maximum of 100 iterations, using the continuous response variable as target. Only the variables classified as "Confirmed" were retained for the next step. For both methods, the selected variables were submitted to a collinearity analysis. The Pearson correlation matrix was calculated between all variables and then the elimination of highly correlated variables was carried out, adopting a strict threshold of $|r| > 0.9$. The collinearity was analyzed separately for each method. All procedures were implemented in R language (version 4.4.1).

Predictive analysis

The evaluation of predictive performance was performed comparing two algorithms: regression by PLS and RF. Both models were developed in the R environment. Initially, the data set was randomly divided, allocating 70% for training and 30% for testing. The division was stratified based on SCC values to ensure proportional representation of the different counting ranges in training and testing, and a seed (seed = 123) was fixed to ensure reproducibility. For PLS modeling, the *pls* package was used. The selection of input spectral variables was based on components previously identified by PCA. In order to avoid overfitting, the optimal number of latent components was defined by means of a 10-part (10-fold) cross validation. The RF model was implemented with the *randomForest* package, being configured to generate 500 decision trees. Critical RF hyperparameters, such as *mtry* and *nodesize*, were optimized by grid search with stratified cross-validation (5-fold) in the training set. In this case, the variables used were selected by the Boruta algorithm. The final predictive performance of both models was quantified using only the test set. The metrics used for this comparison were the root mean square error (RMSE) and the coefficient of determination (R^2). For greater robustness, the metrics were accompanied by standard deviations obtained from 10 independent repetitions of the training/test division.

Analysis of classification

The discriminative analyses were performed by the methods Linear Discriminant Analysis (LDA) and the RF algorithm. Initially, it was carried out based on a division of the data into training set (70%) and test set (30%). As in the regression, the division was stratified by SCC class and fixed with seed = 123 to ensure reproducibility. The samples were categorized into two classes based on the threshold of 500,000 cells/mL¹⁵. The RF model was configured

with 500 decision trees. Its hyperparameters were optimized by grid search with stratified cross validation (5-fold) in the training set, considering mtry, maximum depth of trees and nodesize. Already LDA model was implemented using the Bayes rule to estimate the probabilities of belonging to each class. The performance of both models was quantified by means of metrics derived from the confusion matrix, adopting the procedure described by Varoquaux and Colliot (2023) and Hicco, Tötsch and Jurman (2021). The metrics included sensitivity, specificity, precision, F1-score and accuracy, estimated from 10 repetitions of the training/test division.

RESULTS AND DISCUSSION

Composition of milk

Figure 2 shows graphs comparing SCC with fat, protein and lactose in the 624 milk samples analyzed. The results indicate that none of these relationships was statistically significant, since all p values were greater than 0.05. R^2 was very low for fat (0.002) and protein (0.003). This means that less than 1% of changes in the contents of these components are related to SCC. Most studies demonstrate only a weak or inconsistent relationship between SCC and milk fat/protein percentages. Therefore, the change in fat and protein levels due to SCC is small compared with other factors such as lactation stage, parity and season^{18–20}. In the case of lactose, although the correlation coefficient was higher ($r = -0.399$), R^2 was still low (0.160) and was not statistically significant ($p = 0.158$). This result is consistent with other studies^{21–23} which show that high SCC can cause a more significant reduction in lactose concentration than fat and protein.

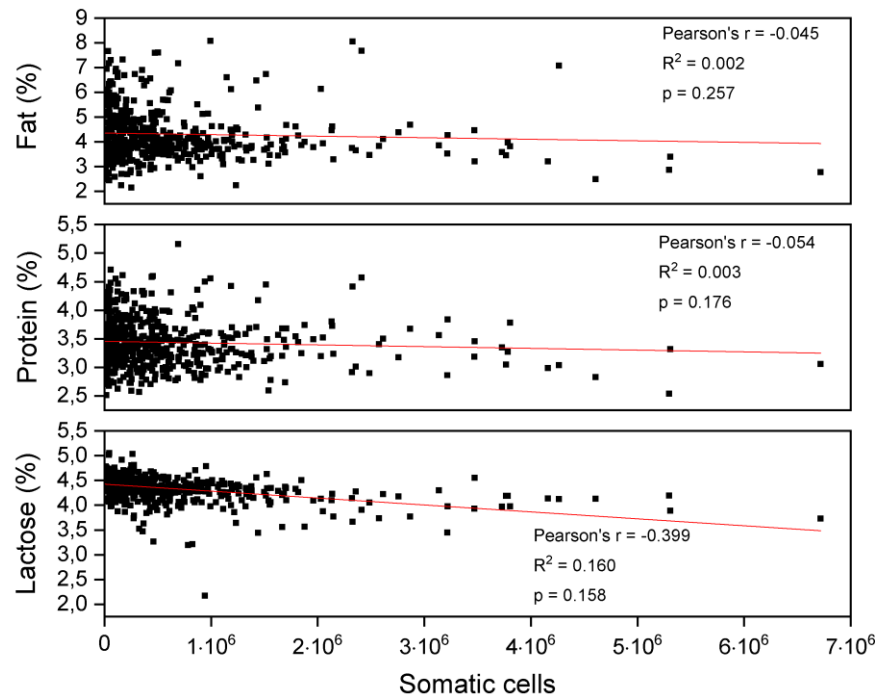


Figure 2. Relationship between the Somatic Cell Count (SCC) and milk components: fat, protein and lactose.

Table 1 presents the descriptive statistics of the main milk components analyzed in this study, including somatic cell count (SCC) and fat, protein and lactose levels. The mean and median values indicate that the sample is representative of a heterogeneous population, with wide variation in SCC levels, which reflects the intentional inclusion of healthy animals with subclinical or clinical mastitis, as planned in the methodology. Among the main components of milk, fat was the component with greater variation between samples ($CV = 25.1\%$), followed by protein ($CV = 12.8\%$) and lactose ($CV = 6.3\%$). These differences reflect the diversity of the group of cows analyzed, which included healthy animals with subclinical mastitis and clinical mastitis. This large variation in fat and protein content is consistent with what is already known in the literature that says that mastitis causes changes in milk metabolism, directly changing its composition ^{24, 25}.

Table 1. Descriptive statistics of milk components

Compound	Mean (sd)	Minimum–Maximum	Median	CV
SCC	560529.9 (31312.33)	6452 - 6718733	306933	139.430
Fat	4.319 (1.084)	2.140 – 8.074	4.034	25.098
Protein	3.439 (0.439)	2.511 – 5.155	3.368	12.764
Lactose	4.347 (0.274)	2.167 – 5.051	4.375	6.296

Analysis of the spectra

Figure 3 shows the absorption spectra in the FTIR for milk samples with different concentrations of SCC, where characteristic spectra of the milk components present in 1038, 1078, 1113, 1160, 1242, 1455, 1554, 1661, 1743, 2857 and 2928 cm^{-1} are noted. The peaks present in 2857 and 2928 cm^{-1} are associated with C-H bonds in fats, while wave numbers between 1000 and 1250 cm^{-1} usually relate to sugars such as lactose and protein structures ²⁶, ²⁷. Peaks around 1455 and 1554 cm^{-1} are linked to amide groups of milk proteins and values close to 1661 and 1743 cm^{-1} indicate carbonyl bonds, typical of fats ^{27,28}. No articles were found that associate these peaks with SCC. However, studies show that SCC is associated with changes in milk composition, such as proteins, lipids and metabolite ²⁹. Some studies ^{21,30} have shown that high SCC in milk is correlated with increased production of metabolites such as lactate, butyrate, isoleucine, acetate and β -hydroxybutyrate and reduced production of hypurate and fumarate, but the specific attributions of the spectral peaks were not detailed. In addition, FTIR can be used to predict milk characteristics and health conditions such as ketosis, but without direct attributions to somatic cell peaks ³¹.

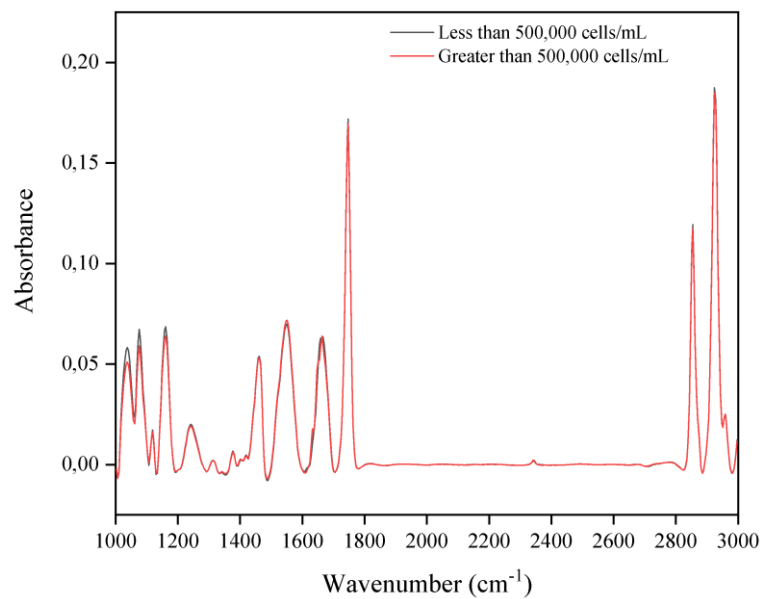


Figure 3. Spectrum in the mean infrared spectrum of milk samples with SCC values greater and less than 500,000 cells/mL

Selection of variables

The results of wave number selection by means of the Boruta and PCA algorithms is shown in Table 2. The Boruta algorithm identified 10 significant wave numbers, while the PCA selected a more comprehensive set with 52 wave numbers, mainly distributed in the regions between 1007-1126 cm^{-1} , 1223-1273 cm^{-1} , 1393-1435 cm^{-1} and 1508-1562 cm^{-1} . The region between 1000-1130 cm^{-1} , strongly represented in the selections of both algorithms, is associated with C-O and C-C vibrations characteristic of lactose and other carbohydrate structures^{26, 27}. This is because metabolic change caused by inflammation can lead to changes in the profile of sugars and other metabolites, such as the reduction of glucose available for lactose production³².

The 1250-1270 cm^{-1} region, selected by both methods, is associated with secondary protein structures, whose relevance for SCC prediction may be related to changes in protein

1 conformation due to the presence of inflammation-specific proteins. Wave numbers selected
2 between 1400-1435 cm^{-1} by both methods, close to the region of 1455 cm^{-1} , is associated with
3 amide groups of milk proteins ^{33, 34}. This is associated with inflammation of the mammary
4 gland, which is correlated to an increase in the concentration of immune-related proteins such
5 as immunoglobulins, haptoglobin, serum amyloid A and antibacterial proteins ^{35, 36}. In addition,
6 serum proteins such as bovine serum albumin and globulins leak into the milk due to increased
7 permeability of the blood-milk barrier caused by inflammation ³⁷.

8 The PCA selected multiple wave numbers in the region 1510-1560 cm^{-1} related to the
9 absorption of Amide II, variables not selected in Boruta. Similarly, the wave number 2881 cm^{-1} ,
10 selected only by Boruta, is in the characteristic region of C-H bonds present in fats ^{26, 27}. This
11 may have been caused by the different objectives of each method. The PCA, being
12 unsupervised, seeks to capture the parts of the spectrum that most vary between samples, even
13 if this variation is not directly linked to the SCC. Therefore, she identified several wave
14 numbers in the Amide II region (1510-1560 cm^{-1}), which may reflect changes in milk proteins
15 that may have been influenced by SCC but also by other factors. In contrast, the Boruta is a
16 method of selection of supervised characteristics able to identify s most relevant variables for
17 the prediction of SCC, excluding variables that do not have great predictive power, as well as
18 excluding variables do not contribute to classification or prediction analyses ^{38, 39}.

19 It should be noted that all selected wave numbers are associated with the functional
20 groups of milk components, and this difficulty in capturing predictive SCC patterns is related
21 to the limitations of the FTIR. Somatic cells, mainly leukocytes, do not present intense and
22 distinct spectral signatures in the mid-infrared. This is because their biomolecular constituents,
23 such as membrane proteins and nucleic acids, have spectral bands that overlap with those of
24 major milk components, such as caseins and lipids, making it difficult for FTIR to directly
25 differentiate these cells ⁴⁰⁻⁴².

Table 2. Wave numbers selected by the Boruta and PCA approaches associated with somatic cell count.

Selection method		Functional attribution	Reference
Boruta	PCA		
1022, 1061, 1080, 1088, 1122, 1126	1007, 1018, 1022, 1034, 1038, 1041, 1045, 1049, 1053, 1057, 1061, 1065, 1068, 1072, 1076, 1080, 1084, 1088, 1092	C-O, C-C and C-O-H of carbohydrates	26,27
1257, 1269	1223, 1227, 1261, 1265, 1273	Amina III	33,34
1435	1400, 1404, 1412, 1416, 1420, 1423, 1427, 1431, 1435	Deformation of CH /CH , COO (carboxylate)	43,44
—	1520, 1524, 1528, 1531, 1535, 1539, 1543, 1547, 1551, 1555, 1558, 1562	Amida II	27,28
2881	—	C-H elongation in fatty acids	26,27

Predictive methods

The comparative analysis between RF and LDA models for the prediction of SCC in milk samples (Figure 4) showed unsatisfactory performance in both approaches. The values of R^2 were 0.255 for RF and 0.293 for PLS, accompanied by elevated RMSE, 7.24 105 and 7.03 105 cells/mL, respectively. These results demonstrate that both the non-linear (RF) and the linear (PLS) model failed to establish a relationship between the characteristics of the infrared spectra and the concentration of SCC, even after applying techniques of prior selection of variables.

The limitation of the models in capturing prediction patterns may be directly related to FTIR constraints for the quantification of SCC in a complex matrix such as milk. Somatic cells, predominantly leukocytes, do not present distinct and intense spectral signatures in the medium infrared, since their biomolecular constituents, such as membrane proteins and nucleic acids, overlap the bands of major milk components, as caseins and lipids ^{40–42}. The literature ^{6,19} shows

that the prediction of SCC by the FTIR spectrum of milk is made indirectly from changes in milk composition, such as reduction of lactose, changes in protein fractions and changes in ion concentration, and not the proper bands of somatic cells. Another important aspect that may have influenced the performance of the models was the choice of variables. In the PLS, the selection based on PCA prioritized variables that most contribute to the variance of the spectra, using an unsupervised criterion that does not directly consider the relationship with SCC, the response variable⁴⁵. Although the PCA is effective in reducing redundancies, its isolated application may have kept variables irrelevant to the prediction, possibly eliminating spectral bands of low intensity, but potentially specific for the SCC. For RF, the selection was conducted by the Boruta algorithm, which evaluates the importance of the variables through comparison with shadow features. Although Boruta is robust in identifying relevant variables in non-linear data, its dependence on measures of individual importance may have eliminated features where relevance only exists in complex interactions. It is possible that Boruta, by being overly strict, may have discarded weak but potentially informative signals that would be crucial to the nonlinear model^{46–48}.

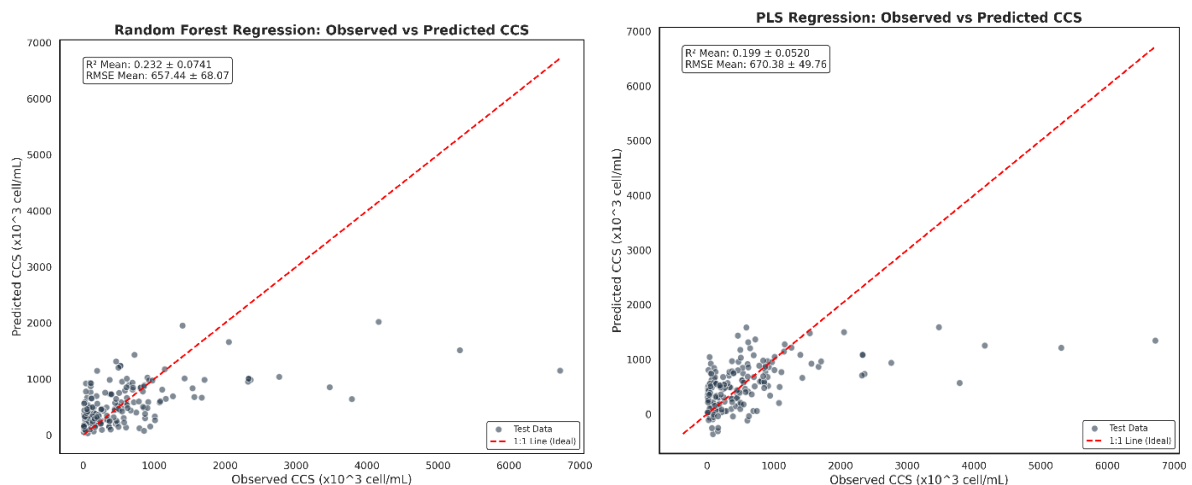


Figure 4. Comparison of the predictive performance of RF and PLS in estimating the concentration of SCC in milk.

Discriminatory methods

The comparison between RF and PLS models for the classification of milk samples, considering the SCC above and below 500,000 cells/mL is shown in Table 3. The results show that although both models exhibit similar accuracy, their discrimination capabilities varied in relation to sensitivity, specificity and precision. The sensitivity of LDA (0.795) was higher than that of RF (0.455), indicating that the linear model can correctly identify almost 80% of samples with SCC above 500,000 cells/mL, while RF cannot detect more than half of these cases. This difference between sensitivity is critical. A model with low sensitivity, such as the RF, would have the possibility of making mistakes in classifying pathological samples as healthy ⁵³, impairing the effectiveness of sanitary control. This finding is in line with the results of Rienesl et al. ⁹, which shows that models constructed from data from the mid-infrared spectra do not present a balance between sensitivity and specificity in the diagnosis of mastitis. On the other hand, RF, as a nonlinear model based on decision trees, is theoretically more robust to nonlinearities and complex interactions between variables. However, its lower performance in sensitivity in this study suggests that the correlations between FTIR spectra and SCC concentration may be mainly linear or that the model is facing overfitting in areas with noise or low biological importance ^{54, 55}.

The high specificity of RF (0.908) contrasts with that of LDA (0.725), indicating that RF is more rigorous to classify samples with SCC > 500,000 cells/mL, generating few false positives. The high specificity of RF can be partially attributed to its propensity to favor variables of high importance, which, even being discriminatory in certain subsets of data, may not be robust in all population distribution. Tsenkova et al. ⁸ observed that spectral models for SCC are influenced by animal individuality, i.e., unique spectral patterns of each cow. This behavior may lead to non-linear models learning individual characteristics rather than patterns associated with udder health condition. This is common in spectroscopy when dealing with heterogeneous samples with different breeds, lactation stages or diets, where SCC spectral

patterns can be influenced by interfering factors. Franzoi and collaborators ¹⁹ demonstrated that extreme levels of SCC, both very high and very low, can affect the milk composition in a non-uniform way between races, which may hinder the generalization of nonlinear models.

As for precision, RF achieved higher precision (0.761) compared to LDA (0.454). This indicates that although RF detects fewer true cases ($SCC > 500,000$ cells/mL), when detection occurs they are likely to be correct. In contrast, the LDA, although discriminating more true cases, generates a high rate of false positives, which can generate some implications, such as a confirmation by complementary methods in almost half of the positive cases, increasing costs and processing time ^{56,57}. The RF, although less sensitive, could be used with more confidence to filter samples for confirmation of suspected cases. In animal health contexts, as demonstrated by ⁹, the loss of a false negative case has more serious consequences than a false positive, which can be corrected by retest. Therefore, the higher sensitivity values of LDA, even with lower precision, makes it more suitable for udder health analysis than RF.

Table 3. Performance metrics of RFe PLS models for classification of milk samples based on SCC

Metric	RF	PLS
Balanced Accuracy	0.731	0.741
Precision	0.761	0.454
Sensitivity	0.455	0.795
Specificity	0.908	0.725
F1-Score	0.569	0.579

CONCLUSION

Despite the theoretical potential of middle infrared spectroscopy combined with chemometrics, the evaluated models presented limited performance for accurate quantification of SCC in heterogeneous populations. However, in practical terms, FTIR spectroscopy associated with random forest demonstrates viability as a quick and low-cost screening tool for

preliminary identification of cows with subclinical or clinical mastitis in dairy herds, especially for using existing equipment in most milk analysis laboratories, although it does not replace specific diagnostic methods to confirm positive cases.

AUTHOR CONTRIBUTIONS

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Romicy Dermondes Souza: Writing – original draft, Formal analysis, Investigation.

Rosângela Silveira Barbosa: .Resources, Supervision.

Marcelo Bonnet Alvarenga: Data Curation, Validation.

Fabiano Barreto: Writing – review & editing.

Wellington da Silva Lyra: Investigation.

Sérgio de Sousa Castro: Methodology. Conceptualization

Paulo Henrique Gonçalves Dias Diniz: Methodology, Writing – review & editing.

Sérgio Augusto de Albuquerque Fernandes: Supervision, Project administration, Writing – review & editing, Conceptualization.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors would also like to thank Empresa Brasileira de Pesquisa Agropecuária (Embrapa) and the Ministério da Agricultura, Pecuária e Abastecimento (MAPA) for their support.

ABBREVIATIONS

FTIR, Fourier transform infrared spectroscopy; MIR, mid-infrared; SCC, somatic cell count; PLS, partial least squares; RF, random forest; LDA, linear discriminant analysis; PCA, principal component analysis; RMSE, root mean square error; R^2 , coefficient of determination; CV, coefficient of variation; SD, standard deviation.

REFERENCES

- (1) Stanek, P.; Żółkiewski, P.; Januś, E. A Review on Mastitis in Dairy Cows Research: Current Status and Future Perspectives. *Agriculture* 2024, 14 (8), 1292. <https://doi.org/10.3390/agriculture14081292>.
- (2) Stocco, G.; Summer, A.; Cipolat-Gotet, C.; Zanini, L.; Vairani, D.; Dadousis, C.; Zecconi, A. Differential Somatic Cell Count as a Novel Indicator of Milk Quality in Dairy Cows. *Animals* 2020, 10 (5), 753. <https://doi.org/10.3390/ani10050753>.
- (3) Damm, M.; Holm, C.; Blaabjerg, M.; Bro, M. N.; Schwarz, D. Differential Somatic Cell Count—A Novel Method for Routine Mastitis Screening in the Frame of Dairy Herd Improvement Testing Programs. *J. Dairy Sci.* 2017, 100 (6), 4926–4940. <https://doi.org/10.3168/jds.2016-12409>.
- (4) Rachah, A.; Reksen, O.; Tafintseva, V.; Stehr, F. J. M.; Rukke, E.-O.; Prestløkken, E.; Martin, A.; Kohler, A.; Afseth, N. K. Exploring Dry-Film FTIR Spectroscopy to Characterize Milk Composition and Subclinical Ketosis throughout a Cow's Lactation. *Foods* 2021, 10 (9), 2033. <https://doi.org/10.3390/foods10092033>.
- (5) Pegolo, S.; Mota, L. F. M.; Bisutti, V.; Martinez-Castillero, M.; Giannuzzi, D.; Gallo, L.; Schiavon, S.; Tagliapietra, F.; Revello Chion, A.; Trevisi, E.; Negrini, R.; Ajmone Marsan, P.;

- 1 Cecchinato, A. Genetic Parameters of Differential Somatic Cell Count, Milk Composition, and
2 Cheese-Making Traits Measured and Predicted Using Spectral Data in Holstein Cows. *J. Dairy*
3 *Sci.* 2021, 104 (10), 10934–10949. <https://doi.org/10.3168/jds.2021-20395>.
- 4 (6) Tsenkova, R.; Atanassova, S.; Kawano, S.; Toyoda, K. Somatic Cell Count
5 Determination in Cow's Milk by near-Infrared Spectroscopy: A New Diagnostic Tool. *J. Anim.*
6 *Sci.* 2001, 79 (10), 2550. <https://doi.org/10.2527/2001.79102550x>.
- 7 (7) Tsenkova, R.; Atanassova, S.; Ozaki, Y.; Toyoda, K.; Itoh, K. Near-Infrared
8 Spectroscopy for Biomonitoring: Influence of Somatic Cell Count on Cow's Milk Composition
9 Analysis. *Int. Dairy J.* 2001, 11 (10), 779–783. [https://doi.org/10.1016/S0958-6946\(01\)00110-](https://doi.org/10.1016/S0958-6946(01)00110-8)
10 8.
- 11 (8) Tsenkova, R.; Meilina, H.; Kuroki, S.; Burns, D. H. Near Infrared Spectroscopy Using
12 Short Wavelengths and Leave-One-Cow-Out Cross-Validation for Quantification of Somatic
13 Cells in Milk. *J. Near Infrared Spectrosc.* 2009, 17 (6), 345–351.
14 <https://doi.org/10.1255/jnirs.868>.
- 15 (9) Rienesl, L.; Khayatdadeh, N.; Köck, A.; Egger-Danner, C.; Gengler, N.; Grelet, C.;
16 Dale, L. M.; Werner, A.; Auer, F.-J.; Leblois, J.; Sölkner, J. Prediction of Acute and Chronic
17 Mastitis in Dairy Cows Based on Somatic Cell Score and Mid-Infrared Spectroscopy of Milk.
18 *Animals* 2022, 12 (14), 1830. <https://doi.org/10.3390/ani12141830>.
- 19 (10) Salman, H. A.; Kalakech, A.; Steiti, A. Random Forest Algorithm Overview. *Babylonian*
20 *J. Mach. Learn.* 2024, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>.
- 21 (11) Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* 2016, 25 (2), 197–227.
22 <https://doi.org/10.1007/s11749-016-0481-7>.
- 23 (12) Breiman, L. Random Forests. *Mach. Learn.* 2001, 45 (1), 5–32.
24 <https://doi.org/10.1023/A:1010933404324>.

- 1 (13) Adedeji, A. A.; Okeke, A.; Rady, A. M. Utilization of FTIR and Machine Learning for
2 Evaluating Gluten-Free Bread Contaminated with Wheat Flour. *Sustainability* 2023, 15 (11),
3 8742. <https://doi.org/10.3390/su15118742>.
- 4 (14) Freitas, A. G. M. de; Almir Cavalcante Minho, L.; Elizabeth Alves de Magalhães, B.;
5 Nei Lopes dos Santos, W.; Soares Santos, L.; Augusto de Albuquerque Fernandes, S. Infrared
6 Spectroscopy Combined with Random Forest to Determine Tylosin Residues in Powdered
7 Milk. *Food Chem.* 2021, 365, 130477. <https://doi.org/10.1016/j.foodchem.2021.130477>.
- 8 (15) MAPA. INSTRUÇÃO NORMATIVA No 76, DE 26 DE NOVENBRO DE 2018; 2018.
9 [https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750137/doi1 -](https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750137/doi1-2018-11-30-instrucao-normativa-n-76-de-26-de-novembro-de-2018-52749894IN_76)
10 [2018-11-30-instrucao-normativa-n-76-de-26-de-novembro-de-2018-52749894IN 76](https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/52750137/doi1-2018-11-30-instrucao-normativa-n-76-de-26-de-novembro-de-2018-52749894IN_76).
- 11 (16) Varoquaux, G.; Colliot, O. Evaluating Machine Learning Models and Their Diagnostic
12 Value; 2023; pp 601–630. https://doi.org/10.1007/978-1-0716-3195-9_20.
- 13 (17) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (MCC) Is
14 More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-
15 Class Confusion Matrix Evaluation. *BioData Min.* 2021, 14 (1), 13.
16 <https://doi.org/10.1186/s13040-021-00244-z>.
- 17 (18) Pegolo, S.; Giannuzzi, D.; Bisutti, V.; Tessari, R.; Gelain, M. E.; Gallo, L.; Schiavon, S.;
18 Tagliapietra, F.; Trevisi, E.; Ajmone Marsan, P.; Bittante, G.; Cecchinato, A. Associations
19 between Differential Somatic Cell Count and Milk Yield, Quality, and Technological
20 Characteristics in Holstein Cows. *J. Dairy Sci.* 2021, 104 (4), 4822–4836.
21 <https://doi.org/10.3168/jds.2020-19084>.
- 22 (19) Franzoi, M.; Manuelian, C. L.; Penasa, M.; De Marchi, M. Effects of Somatic Cell Score
23 on Milk Yield and Mid-Infrared Predicted Composition and Technological Traits of Brown
24 Swiss, Holstein Friesian, and Simmental Cattle Breeds. *J. Dairy Sci.* 2020, 103 (1), 791–804.
25 <https://doi.org/10.3168/jds.2019-16916>.

- 1 (20) Ermetin, O.; Kul, E.; Okuyucu, I. C. The Relationship of Somatic Cell Count with Milk
2 Yield and Composition in Different Stages of Lactation in Holstein Cows. *Vet. Arh.* 2024, 94
3 (1), 21–32. <https://doi.org/10.24099/vet.arhiv.2213>.
- 4 (21) Cinar, M.; Serbester, U.; Ceyhan, A.; Gorgulu, M. Effect of Somatic Cell Count on Milk
5 Yield and Composition of First and Second Lactation Dairy Cows. *Ital. J. Anim. Sci.* 2015, 14
6 (1), 3646. <https://doi.org/10.4081/ijas.2015.3646>.
- 7 (22) Fonseca, M.; Kurban, D.; Roy, J.-P.; Santschi, D. E.; Molgat, E.; Dufour, S. Usefulness
8 of Differential Somatic Cell Count for Udder Health Monitoring: Association of Differential
9 Somatic Cell Count and Somatic Cell Score with Quarter-Level Milk Yield and Milk
10 Components. *J. Dairy Sci.* 2025, 108 (4), 3900–3916. <https://doi.org/10.3168/jds.2024-25402>.
- 11 (23) Wahyu Harjanti, D.; Sambodho, P. Effects of Mastitis on Milk Production and
12 Composition in Dairy Cows. *IOP Conf. Ser. Earth Environ. Sci.* 2020, 518 (1), 012032.
13 <https://doi.org/10.1088/1755-1315/518/1/012032>.
- 14 (24) Chegini, A.; Ghavi Hossein-Zadeh, N.; Hosseini-Moghadam, H.; Shadparvar, A. A.
15 Effect of Somatic Cell Count on Milk Fat and Protein in Different Parities and Stages of
16 Lactation in Holstein Cows. *Acta Agric. Slov.* 2017, 110 (1).
17 <https://doi.org/10.14720/aas.2017.110.1.5>.
- 18 (25) Kul, E.; Şahin, A.; Atasever, S.; Uğurlutepe, E.; Soydaner, M. The Effects of Somatic
19 Cell Count on Milk Yield and Milk Composition in Holstein Cows. *Vet. Arh.* 2019, 89 (2), 143–
20 154. <https://doi.org/10.24099/vet.arhiv.0168>.
- 21 (26) Toledo-Alvarado, H.; Pérez-Cabal, M. A.; Tempelman, R. J.; Cecchinato, A.; Bittante,
22 G.; de los Campos, G.; Vazquez, A. I. Association between Days Open and Milk Spectral Data
23 in Dairy Cows. *J. Dairy Sci.* 2021, 104 (3), 3665–3675. <https://doi.org/10.3168/jds.2020-19031>.
- 24 (27) Bittante, G.; Cecchinato, A. Genetic Analysis of the Fourier-Transform Infrared Spectra
25 of Bovine Milk with Emphasis on Individual Wavelengths Related to Specific Chemical Bonds.
26 *J. Dairy Sci.* 2013, 96 (9), 5991–6006. <https://doi.org/10.3168/jds.2013-6583>.

- 1 (28) Wang, Q.; Hulzebosch, A.; Bovenhuis, H. Genetic and Environmental Variation in
2 Bovine Milk Infrared Spectra. *J. Dairy Sci.* 2016, 99 (8), 6793–6803.
3 <https://doi.org/10.3168/jds.2015-10488>.
- 4 (29) Sundekilde, U. K.; Poulsen, N. A.; Larsen, L. B.; Bertram, H. C. Nuclear Magnetic
5 Resonance Metabonomics Reveals Strong Association between Milk Metabolites and Somatic
6 Cell Count in Bovine Milk. *J. Dairy Sci.* 2013, 96 (1), 290–299.
7 <https://doi.org/10.3168/jds.2012-5819>.
- 8 (30) Podhorecká, K.; Borková, M.; Šulc, M.; Seydlová, R.; Dragounová, H.; Švejcárová, M.;
9 Peroutková, J.; Elich, O. Somatic Cell Count in Goat Milk: An Indirect Quality Indicator. *Foods*
10 2021, 10 (5), 1046. <https://doi.org/10.3390/foods10051046>.
- 11 (31) Klein, S.-L.; Scheper, C.; May, K.; König, S. Genetic and Nongenetic Profiling of Milk
12 β -Hydroxybutyrate and Acetone and Their Associations with Ketosis in Holstein Cows. *J. Dairy*
13 *Sci.* 2020, 103 (11), 10332–10346. <https://doi.org/10.3168/jds.2020-18339>.
- 14 (32) Caboni, P.; Manis, C.; Ibba, I.; Contu, M.; Coroneo, V.; Scano, P. Compositional Profile
15 of Ovine Milk with a High Somatic Cell Count: A Metabolomics Approach. *Int. Dairy J.* 2017,
16 69, 33–39. <https://doi.org/10.1016/j.idairyj.2017.02.001>.
- 17 (33) Nandiyanto, A. B. D.; Ragadhita, R.; Fiandini, M. Interpretation of Fourier Transform
18 Infrared Spectra (FTIR): A Practical Approach in the Polymer/Plastic Thermal Decomposition.
19 *Indones. J. Sci. Technol.* 2022, 8 (1), 113–126. <https://doi.org/10.17509/ijost.v8i1.53297>.
- 20 (34) Enders, A.; North, N.; Fensore, C.; Velez-Alvarez, J.; Allen, H. Functional Group
21 Identification for Ftir Spectra Using Image-Based Machine Learning Models. March 11, 2021.
22 <https://doi.org/10.26434/chemrxiv.14188679.v1>.
- 23 (35) Turk, R.; Rošić, N.; Kuleš, J.; Horvatić, A.; Gelemanovic, A.; Galen, A.; Ljubić, B. B.;
24 Benić, M.; Stevanović, V.; Mrljak, V.; Chadwick, C. C.; Eckersall, P. D. Milk and Serum
25 Proteomes in Subclinical and Clinical Mastitis in Simmental Cows. *J. Proteomics* 2021, 244,
26 104277. <https://doi.org/10.1016/j.jprot.2021.104277>.

- 1 (36) Eckersall, P. D.; Young, F. J.; Nolan, A. M.; Knight, C. H.; McComb, C.; Waterston, M.
2 M.; Hogarth, C. J.; Scott, E. M.; Fitzpatrick, J. L. Acute Phase Proteins in Bovine Milk in an
3 Experimental Model of Staphylococcus Aureus Subclinical Mastitis. *J. Dairy Sci.* 2006, 89 (5),
4 1488–1501. [https://doi.org/10.3168/jds.S0022-0302\(06\)72216-0](https://doi.org/10.3168/jds.S0022-0302(06)72216-0).
- 5 (37) Tiantong, A.; Eardmusic, S.; Arunvipas, P.; Lee, J.-W.; Inyawilert, W. The Influence of
6 Subclinical Mastitis on the Protein Composition and Protease Activities of Raw Milk from
7 Lactating Thai-Crossbred Dairy Cows. *Vet. World* 2023, 1363–1368.
8 <https://doi.org/10.14202/vetworld.2023.1363-1368>.
- 9 (38) Oktafia, O.; Nugroho, R. S. A. Comparison Pf Support Vector Machine(SVM), Xgboost
10 and Random Forest for Sentiment Analysis of Bumble App User Comments. *Proxies J. Inform.*
11 2024, 6 (1), 32–46. <https://doi.org/10.24167/proxies.v6i1.12453>.
- 12 (39) Santos, R. de O.; Gorgulho, B. M.; Castro, M. A. de; Fisberg, R. M.; Marchioni, D. M.;
13 Baltar, V. T. Principal Component Analysis and Factor Analysis: Differences and Similarities
14 in Nutritional Epidemiology Application. *Rev. Bras. Epidemiol.* 2019, 22.
15 <https://doi.org/10.1590/1980-549720190041>.
- 16 (40) Du, C.; Ren, X.; Chu, C.; Ding, L.; Nan, L.; Sabek, A.; Hua, G.; Yan, L.; Zhang, Z.;
17 Zhang, S. Assessing the Relationship between Somatic Cell Count and the Milk Mid-infrared
18 Spectrum in Chinese Holstein Cows. *Vet. Rec.* 2023, 193 (11).
19 <https://doi.org/10.1002/vetr.3560>.
- 20 (41) Fleming, A.; Schenkel, F. S.; Ali, R. A.; Corredig, M.; Carta, S.; Gregu, C. M.;
21 Malchiodi, F.; Macciotta, N. P. P.; Miglior, F. Phenotypic Investigation of Fine Milk
22 Components in Bovine Milk and Their Prediction Using Mid-Infrared Spectroscopy. *Can. J.*
23 *Anim. Sci.* 2019, 99 (2), 218–227. <https://doi.org/10.1139/cjas-2018-0058>.
- 24 (42) Grewal, M. K.; Chandrapala, J.; Donkor, O.; Apostolopoulos, V.; Stojanovska, L.;
25 Vasiljevic, T. Fourier Transform Infrared Spectroscopy Analysis of Physicochemical Changes

- 1 in UHT Milk during Accelerated Storage. *Int. Dairy J.* 2017, 66, 99–107.
2 <https://doi.org/10.1016/j.idairyj.2016.11.014>.
- 3 (43) Conceição, D.; Gonçalves, B.-H.; da Hora, F.; Faleiro, A.; Santos, L.; Ferrão, S. Use of
4 FTIR-ATR Spectroscopy Combined with Multivariate Analysis as a Screening Tool to Identify
5 Adulterants in Raw Milk. *J. Braz. Chem. Soc.* 2018. [https://doi.org/10.21577/0103-](https://doi.org/10.21577/0103-5053.20180208)
6 5053.20180208.
- 7 (44) Jaiswal, P.; Jha, S. N.; Borah, A.; Gautam, A.; Grewal, M. K.; Jindal, G. Detection and
8 Quantification of Soymilk in Cow–Buffalo Milk Using Attenuated Total Reflectance Fourier
9 Transform Infrared Spectroscopy (ATR–FTIR). *Food Chem.* 2015, 168, 41–47.
10 <https://doi.org/10.1016/j.foodchem.2014.07.010>.
- 11 (45) Šašić, S.; Ozaki, Y. Comparison of Principal Component Analysis and Generalized Two-
12 Dimensional Correlation Spectroscopy: Spectral Analysis of Synthetic Model System and
13 Near-Infrared Spectra of Milk. *Appl. Spectrosc.* 2001, 55 (1), 29–38.
14 <https://doi.org/10.1366/0003702011951399>.
- 15 (46) Kursá, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.*
16 2010, 36 (11). <https://doi.org/10.18637/jss.v036.i11>.
- 17 (47) Zhou, H.; Xin, Y.; Li, S. A Diabetes Prediction Model Based on Boruta Feature Selection
18 and Ensemble Learning. *BMC Bioinformatics* 2023, 24 (1), 224.
19 <https://doi.org/10.1186/s12859-023-05300-5>.
- 20 (48) Maurya, N. S.; Kushwah, S.; Kushwaha, S.; Chawade, A.; Mani, A. Prognostic Model
21 Development for Classification of Colorectal Adenocarcinoma by Using Machine Learning
22 Model Based on Feature Selection Technique Boruta. *Sci. Rep.* 2023, 13 (1), 6413.
23 <https://doi.org/10.1038/s41598-023-33327-4>.
- 24 (49) Kassaw, A. K.; Bekele, G.; Kassaw, A. K.; Yimer, A. Prediction of Acute Respiratory
25 Infections Using Machine Learning Techniques in Amhara Region, Ethiopia. *Sci. Rep.* 2024,
26 14 (1), 27968. <https://doi.org/10.1038/s41598-024-76847-3>.

- 1 (50) Jafarzadeh, H.; Mahdianpari, M.; Gill, E.; Mohammadimanesh, F.; Homayouni, S.
2 Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral
3 and Polsar Data: A Comparative Evaluation. *Remote Sens.* 2021, 13 (21), 4405.
4 <https://doi.org/10.3390/rs13214405>.
- 5 (51) Mateos-Aparicio, G. Partial Least Squares (PLS) Methods: Origins, Evolution, and
6 Application to Social Sciences. *Commun. Stat. - Theory Methods* 2011, 40 (13), 2305–2317.
7 <https://doi.org/10.1080/03610921003778225>.
- 8 (52) Wang, J.; Zhou, S. <scp>Particle Swarm Optimization-XGBoost-based</Scp>
9 Modeling of <scp>radio-frequency</Scp> Power Amplifier under Different Temperatures. *Int.*
10 *J. Numer. Model. Electron. Networks, Devices Fields* 2024, 37 (2).
11 <https://doi.org/10.1002/jnm.3168>.
- 12 (53) Mazer, R. H.; Ulson, J. A. C.; Medeiros, M. I. M. de. Identificação de Adulterantes No
13 Leite Bovino Por Meio de Redes Neurais Artificiais e Técnica Cromática. *Rev. Científica*
14 *Multidiscip. Núcleo do Conhecimento* 2021, 53–68.
15 [https://doi.org/10.32749/nucleodoconhecimento.com.br/engenharia-eletrica/adulterantes-no-](https://doi.org/10.32749/nucleodoconhecimento.com.br/engenharia-eletrica/adulterantes-no-leite)
16 [leite](https://doi.org/10.32749/nucleodoconhecimento.com.br/engenharia-eletrica/adulterantes-no-leite).
- 17 (54) Hapfelmeier, A.; Ulm, K. A New Variable Selection Approach Using Random Forests.
18 *Comput. Stat. Data Anal.* 2013, 60, 50–69. <https://doi.org/10.1016/j.csda.2012.09.020>.
- 19 (55) Charoenthanakitkul, A.; Viboonsang, P.; Kosolsombat, S. Optimizing Malware
20 Detection with Random Forest, XGBoost, LightGBM, and LLM-Reporting. In 2025 IEEE
21 International Conference on Cybernetics and Innovations (ICCI); IEEE, 2025; pp 1–5.
22 <https://doi.org/10.1109/ICCI64209.2025.10987272>.
- 23 (56) Wentzel, J.; Gall, C.; Bourn, M.; De Beer, J.; du Plessis, F.; Fosgate, G. T. Carnivore
24 Detection at the Domestic/Wildlife Interface within Mpumalanga Province, South Africa.
25 *Animals* 2021, 11 (9), 2535. <https://doi.org/10.3390/ani11092535>.

- 1 (57) Thompson, J. S.; Green, M. J.; Hyde, R.; Bradley, A. J.; O'Grady, L. The Use of Machine
- 2 Learning to Predict Somatic Cell Count Status in Dairy Cows Post-Calving. *Front. Vet. Sci.*
- 3 2023, 10. <https://doi.org/10.3389/fvets.2023.1297750>.