Universidade Federal de Minas Gerais

Escola de Veterinária

Departamento de Tecnologia e Inspeção de Produtos de Origem Animal

USO DE ESPECTROFOTOMETRIA FTIR (*Fourier Transform Infrared*) e MINERAÇÃO DE DADOS PARA A DETECÇÃO E IDENTIFICAÇÃO DE ADULTERANTES NO LEITE CRU

Wanessa Luciene Fonseca Tavares

BELO HORIZONTE

2019

Wanessa Luciene Fonseca Tavares

USO DE ESPECTROFOTOMETRIA FTIR (*Fourier Transform Infrared*) e MINERAÇÃO DE DADOS PARA A DETECÇÃO E IDENTIFICAÇÃO DE ADULTERANTES NO LEITE CRU

Tese apresentada à Escola de Veterinária da Universidade Federal de Minas Gerais como requisito para obtenção do grau de Doutor em Ciência Animal

Área de concentração: Tecnologia e Inspeção de Produtos de Origem Animal

Orientador: Prof. Dr. Leorges Moraes da Fonseca

Coorientadoras: Profa. Dra. Mônica Pinho Cerqueira
Profa. Dra. Mônica Oliveira Leite

Belo Horizonte

Escola de Veterinária da UFMG

2019

Tavares, Wanessa Luciene Fonseca, 1982.
T231u Uso de Espectrofotometria FTIR (*Fourier Trasnform Infrared*) e mineração de dados para a detecção e identificação de adulterantes no leite cru. / Wanessa Luciene Fonseca Tavares – 2019.

98p.: il.

Orientador: Leorges Moraes da Fonseca
Coorientadoras: Mônica Pinho Cerqueira
Mônica Oliveira Leite
Tese de Doutorado apresentado à Escola de Veterinária da Universidade Federal de Minas Gerais.

1- Leite– Análise -Teses - 2- Leite– Qualidade - Teses – 3 – Leite – Adulterantes - Teses –
I– Fonseca, Leorges Moraes da - II – Cerqueira, Mônica Pinho – III – Leite, Mônica Oliveira – IV -
Universidade Federal de Minas Gerais, Escola de Veterinária.

CDD – 637.048

Bibliotecária responsável Cristiane Patrícia Gomes – CRB2569

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

WANESSA LUCIENE FONSECA TAVARES VICENTINI

Tese submetida à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em CIÊNCIA ANIMAL, como requisito para obtenção do grau de DOUTOR em CIÊNCIA ANIMAL, área de concentração em TECNOLOGIA E INSPEÇÃO DE PRODUTOS DE ORIGEM ANIMAL.

Aprovada em 24 de Maio de 2019 , pela banca constituída pelos membros:

Prof. Leorges Moraes da Fonseca
Presidente – Orientador

Prof. Sérgio Vale Aguiar Campos
Departamento de Ciências da Computação - UFMG

Profª. Bruna Maria Salotti de Souza
Escola de Veterinária - UFMG

Prof. Marco Antônio Sloboda Cortez
Universidade Federal Fluminense - UFF

Drª. Elisa Helena Paz Andrade
Secretaria Municipal de Contagem

Universidade Federal de Minas Gerais
Escola de Veterinária
Av. Antônio Carlos, 6627 -
Caixa Postal 567 – CEP 30123-970
Belo Horizonte – Minas Gerais
Telefone: (31) 3409-2057/2059(fax)
www.vet.ufmg.br
E-mail cap@vet.ufmg.br

Escola de Veterinária
UFMG

Ao Professor desta casa Wander de Assis Tavares, que, por tão sorte a minha, foi também meu amado pai

AGRADECIMENTOS

A Deus, em Sua infinita bondade, pelo maravilhoso presente da jornada da vida.

Aos meus pais Wander e Célia, por continuarem sendo o meu alicerce.

À minha querida Escola de Veterinária, da admirada UFMG, pelo suporte e estrutura.

Ao meu brilhante mestre Prof. Leorges, pela oportunidade e confiança concedidas, pela orientação, apoio e incentivo nas dificuldades.

À amiga Daniela, do doutorado para a vida, pela parceria forte em todos os momentos.

Ao colega Habib, pela valiosa contribuição e pelos conhecimentos compartilhados.

Aos queridos professores, colegas da pós-graduação e servidores do DTIPOA pelo auxílio no desenvolvimento deste projeto, em especial Profa. Mônica Pinho, Profa. Mônica Leite, Prof. Ronon, Prof. Marcelo, Profa. Cláudia, Juliana Lima, Marcinha e Miltinho.

Aos colaboradores do Laboratório de Análise da Qualidade do Leite - LabUFMG: Rosemary, Fabiana, Wendel, Márcio, Wanderlan, Raquel, Taynara, Maria e Dona Maria, Adriana e Claudianne, pela competência e presteza nos trabalhos realizados, sempre com sorriso largo. Trabalhar com vocês foi sensacional!

À minha família e amigos, por compreenderem e respeitarem os necessários momentos de ausência, e por torcerem pelo meu sucesso.

Ao Giovanni, pelo aprendizado e pelas experiências compartilhadas.

E a você, que direta ou indiretamente contribuiu para a realização deste trabalho, minha sincera gratidão!

“Não haverá borboletas se a vida não passar por longas e silenciosas metamorfoses”

Rubem Alves

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

ABNT	Associação Brasileira de Normas Técnicas
ATR	*Attenuated Total Reflection*
CNN	*Convolutional Neural Network*
CBT	Contagem Bacteriana Total
CCS	Contagem de Células Somáticas
CSV	*Comma-Separated Values*
ELISA	*Enzyme-Linked Immunosorbent Assay*
FEPAM	Fundação Estadual de Proteção Ambiental
FTIR	*Fourier Transform Infrared Spectroscopy*
GBDT	*Gradient Boosting Decision Tree*
GBM	*Gradient Boosting Machine*
ISO	*International Standard Organization*
LabUFMG	Laboratório de Análise da Qualidade do Leite UFMG
LDA	*Linear Discriminant Analysis*
MAPA	Ministério da Agricultura, Pecuária e Abastecimento
MIR	*Mid-Infrared Spectroscopy*
MP-RS	Ministério Público do Rio Grande do Sul
MRL	*Multiple Linear Regression*
MUN	*Milk Urea Nitrogen*
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
PLS	*Partial Least Squares*
PLSDA	*Partial Least Squares Discrimination Analysis*
RBQL	Rede Brasileira de Laboratórios de Controle da Qualidade do Leite
RF	*Random Forest*
RNA	Redes Neurais Artificiais
SNF	*Solids Non Fat*
SNG	Sólidos Não Gordurosos
SIMCA	*Soft Independent Modeling of Class Analogy*
ST	Sólidos Totais

RESUMO

O leite é um alimento de alto valor biológico frequentemente envolvido em fraudes, cuja prática gera não apenas prejuízos econômicos, mas também riscos à saúde do consumidor. Atualmente, os métodos de detecção de adulterantes no leite cru previstos pela legislação brasileira apresentam limitações em relação à sua sensibilidade analítica, além de serem demorados, consumirem grandes quantidades de reagentes e gerarem resíduos poluentes. Por esses motivos, vêm sendo substituídos por métodos instrumentais mais eficientes, a exemplo da espectroscopia FTIR (*Fourier Transform Infrared*), associadas a técnicas de mineração de dados, que permitem a detecção e identificação desses adulterantes. O objetivo deste trabalho foi detectar e identificar as adulterações nos dados espectrais das amostras de leite analisadas pelo espectrofotômetro FTIR, por meio das classificações por aprendizagem profunda e aprendizagem em conjunto. Foram avaliadas 9.788 amostras de leite, das quais 2.376 foram adicionadas de amido, sacarose, bicarbonato de sódio, peróxido de hidrogênio e formaldeído, em concentrações, temperaturas e tempos de armazenamento distintos. Diferentes classificadores foram utilizados para treinar modelos capazes de reconhecer as alterações provocadas pelos adulterantes nas características de composição normal do leite. Foram realizadas as classificações binárias e multiclasse com os subconjuntos selecionados de treinamento e testes para os classificadores *Gradient Boosting Machine* (GBM), *Random Forests* (RF) e *Convolutional Neural Networks* (CNN). A classificação foi realizada usando dois tipos de dados: o espectro total do infravermelho foi analisado pela CNN, e os componentes numéricos extraídos do equipamento, pelos classificadores GBM e RF. Para os métodos em conjunto (GBM e RF), as precisões de classificação variaram de 93,18% a 98,72%. Já a CNN proposta produziu precisões de até 99,34%. Ambos os métodos apresentaram alta precisão, porém a CNN obteve melhores resultados, uma vez que utiliza um conjunto de dados mais denso (coordenadas espectrais). Assim, de acordo com a arquitetura CNN proposta, pode-se predizer, com mais de 99% de acurácia, que a amostra analisada está ou não adulterada (método de triagem) e, ainda mais, em caso positivo, identificar qual o adulterante adicionado no modelo treinado. Portanto, o presente trabalho contribui sobremaneira para a fiscalização agropecuária nacional, uma vez que fornece respaldo metodológico para a detecção de fraudes, visando à garantia de autenticidade, qualidade e de saúde pública na cadeia produtiva do leite.

Palavras-chave: FTIR, mineração de dados, adulterantes, leite.

ABSTRACT

Milk is a high biological value food often involved in fraud, whose practice generates not only economic losses, but also risks to consumer health. Currently, the methods for detecting adulterants in raw milk provided by Brazilian legislation have limitations in relation to their analytical sensitivity, as well as being time consuming, consuming large quantities of reagents and generating pollutant residues. For these reasons, they have been replaced by more efficient instrumentation methods, such as FTIR spectroscopy, associated with data mining techniques that allow the detection and identification of these adulterants. The objective of this work was to detect and identify the adulterations in the spectral data of the milk samples analyzed by the FTIR spectrophotometer, through the classifications by deep and ensemble learning. A total of 9,788 milk samples were evaluated, of which 2,376 were adulterated with starch, sucrose, sodium bicarbonate, hydrogen peroxide and formaldehyde at different concentrations, temperatures and storage times. Different classifiers were used to train models capable of recognizing the alterations caused by the adulterants in the characteristics of normal milk composition. Binary and multiclass classifications were performed with the selected training and test subsets for the Gradient Boosting Machine (GBM), Random Forests (RF) and Convolutional Neural Networks (CNN) classifiers. The classification was performed using two types of data: the total infrared spectrum was analyzed by CNN, and the numerical components extracted from the equipment, by GBM and RF classifiers. For the ensemble methods (GBM and RF), the classification accuracies ranged from 93.18% to 98.72%. The CNN proposal, however, produced precision of up to 99.34%. Both methods presented high precision, but the CNN obtained better results, since it uses a more dense set of data (spectral coordinates). In other words, according to the proposed CNN architecture, one can predict with >99% accuracy that the analyzed sample is unadulterated (screening method) and, even more so, to identify which adulterant is added in the trained model, greatly contributing to the agricultural inspection, aiming at the guarantee of authenticity, quality and public health.

Key words: *FTIR, data mining, adulterants, milk.*

1. INTRODUÇÃO

O leite é um dos alimentos mais completos da natureza e sua importância é baseada em seu elevado valor nutritivo, como a riqueza de proteínas, vitaminas, gorduras, sais minerais e compostos com alta digestibilidade. A qualidade nutricional do leite está estreitamente relacionada às características físico-químicas, sensoriais e microbiológicas, sendo que as análises físico-químicas visam avaliar o valor alimentar e o rendimento industrial, e ainda detectar possíveis fraudes (SILVEIRA *et al*., 2004; SANTOS *et al*., 2011; RIBEIRO *et al*., 2018).

O setor laticinista tem passado por uma crescente demanda por produtos lácteos de alta qualidade, levando à progressiva adaptação desse importante segmento às exigências do mercado consumidor (OLIVEIRA *et al*., 2012).

No Brasil, desde a criação da Rede Brasileira de Laboratórios de Controle da Qualidade do Leite (RBQL), a qualidade do leite vem sendo monitorada por laboratórios oficiais do país para dar suporte às Instruções Normativas Nº76/2018 e Nº77/2018 do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Neste contexto, o leite de cada produtor tem sido avaliado pelo menos uma vez por mês em um dos laboratórios oficiais do país. Apesar desse monitoramento, as técnicas analíticas utilizadas em todas as Unidades Operacionais têm sido objeto de questionamento e de estudo devido às variáveis envolvidas, algumas delas inerentes às próprias técnicas e equipamentos analíticos, ou devido às causas de variação da produção e composição do leite.

Assim, a confiabilidade dos resultados obtidos por meio do contínuo monitoramento da qualidade de leite do tanque e do controle oficial de rebanhos leiteiros impacta todos os elos que compõem a cadeia produtiva do leite. Portanto, a confiabilidade técnica dos resultados analíticos do Programa Nacional de Melhoria da Qualidade do Leite é essencial para a continuidade do progresso das políticas nacionais de avanço no setor laticinista. Ainda mais, com o suporte já existente nesta rede de laboratórios, há um grande potencial para o desenvolvimento de novos ensaios analíticos a serem executados nos equipamentos já existentes, a exemplo da pesquisa de substâncias adulterantes do leite.

Dessa forma, o presente trabalho visa utilizar os ensaios analíticos já existentes nos equipamentos de infravermelho por FTIR (*Fourier Transform Infrared Spectroscopy*)

para o monitoramento de substâncias adulterantes no leite, em atendimento à crescente demanda analítica que tem surgido nos últimos anos para este tipo de equipamento.

Hipóteses:

i. A presença de analitos estranhos à composição normal do leite cru pode levar a alterações na análise do equipamento CombiScope™ (*Delta Instruments*), baseado na espectroscopia de infravermelho por metodologia de *Fourier Transform Infrared* (FTIR).
ii. Estas alterações podem ser diferenciadas dos outros componentes do leite e mensuradas utilizando-se absorção diferenciada em comprimentos de ondas específicos do espectro do infravermelho.
iii. Por meio da técnica de mineração de dados, é possível quantificar estas absorções diferenciadas e estimar qualitativamente a presença de substâncias estranhas ao leite.
iv. Os dados espectrais produzidos por esta técnica podem ser explorados pela utilização de métodos de aprendizado de máquina, como as redes neurais artificiais e as árvores de decisão, com o objetivo de criar modelos que representam as características das amostras de leite puro e de leite adulterado.

2. OBJETIVOS

i. Realizar experimentos com métodos de classificação para reconhecer padrões na composição do leite analisado pelo equipamento CombiScope™ FTIR (*Delta Instruments*), a fim de prever possíveis adulterações por substâncias estranhas.
ii. Detectar e identificar as adulterações nos dados espectrais das amostras de leite analisadas, através das classificações por aprendizagem profunda e aprendizagem em conjunto.

3. REVISÃO DE LITERATURA

3.1 Adulterantes em leite

O leite é o primeiro e mais completo alimento de todos os infantes mamíferos. O seu consumo permanece essencial para o desenvolvimento físico e intelectual do ser humano nas diferentes fases da vida, por ser uma excelente fonte de cálcio e proteínas de alto valor biológico, em quantidades e proporções ideais ao funcionamento do organismo. Por este motivo, o leite é um dos alimentos mais frequentemente envolvidos em fraudes (SANTOS *et al*., 2013; KAMAL & KAROUI, 2015; AZAD & AHMED, 2016; POONIA *et al*., 2017).

De acordo com a literatura, há uma maior incidência de fraudes em leite em países em desenvolvimento e subdesenvolvidos, devido à ausência de monitoramento e fiscalização adequados (AZAD & AHMED, 2016), como Brasil (GONDIM *et al*., 2017), China (ZOU *et al*., 2014), Tailândia (KASEMSUMRAN *et al*., 2007), Paquistão (JAWAID *et al*., 2013) e Índia (KARTHEEK *et al*., 2011), por exemplo.

No Brasil, segundo o Art. 501 do Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal/RIISPOA/2017, considera-se impróprio para qualquer tipo de aproveitamento o leite cru quando, na seleção da matéria prima, apresente resíduos de produtos inibidores, de neutralizantes de acidez, de reconstituintes de densidade ou do índice crioscópico, de conservadores, de agentes inibidores do crescimento microbiano ou de outras substâncias estranhas à sua composição (BRASIL, 2017).

Atualmente, a legislação brasileira determina a pesquisa diária de substâncias neutralizantes da acidez, reconstituintes de densidade ou do índice crioscópico, e de substâncias conservadoras na recepção do leite cru refrigerado (IN Nº77/2018 – MAPA). São estabelecidas provas específicas oficiais para a pesquisa de amido, sacarose, peróxido de hidrogênio e formaldeído (IN Nº30/2018 – MAPA).

Além de lesar o consumidor economicamente, por fornecer um produto de qualidade inferior à declarada, a adulteração do leite pode representar risco à saúde, dependendo

da substância utilizada na fraude (SHARMA & PARADAKAR, 2010; AZAD & AHMED, 2016).

As fraudes mais praticadas no leite são o aumento de volume, a adição de reconstituintes da densidade, neutralizantes da acidez e substâncias conservantes. Somada à motivação financeira, a prática de fraude pode ser incitada pela dificuldade de detecção pelas provas de rotina (TRONCO, 2008; KARTHEEK *et al*., 2011; AZAD & AHMED, 2016).

A principal fraude detectada no leite é a adição de água, cujo objetivo é aumentar o volume total. Essa fraude dificilmente é realizada de forma isolada, já que pode ser rapidamente detectada por provas de rotina, como densidade e crioscopia. Para mascarar a adição de água são utilizadas substâncias denominadas reconstituintes de densidade, tais como a sacarose e o amido, com o objetivo de aumentar o percentual de SNG (sólidos não gordurosos). A adição de água reduz a densidade do leite e aumenta seu ponto de congelamento, enquanto que a adição de reconstituintes produz o efeito inverso. No entanto, leites fraudados com quantidades equilibradas de água e reconstituintes podem não apresentar alterações nestas provas (SANTOS & FONSECA, 2007; TRONCO, 2008). A adição fraudulenta de amido no leite pode causar diarreia devido aos efeitos do amido não digerido no cólon. Além disso, a adição de amido e também de sacarose pode ser bastante prejudicial para pacientes diabéticos, podendo levá-los a óbito (AZAD & AHMED, 2016).

A obtenção do leite em condições de higiene insatisfatórias, bem como a insuficiência de refrigeração do leite após a ordenha, resulta em elevadas contagens de microrganismos que metabolizam a lactose, transformando-a em ácido lático. Com o aumento da acidez titulável, o leite não é recebido pela indústria, causando prejuízos econômicos. Para mascará-la, são adicionadas substâncias neutralizantes, como o bicarbonato de sódio. Em contrapartida, a elevação da acidez pela produção de ácido lático é também responsável por limitar o crescimento microbiano, e a sua neutralização pela adição dessas substâncias favorece consequentemente o aumento da população microbiana, determinando uma queda ainda maior na qualidade do leite (FRANCO & LANDGRAF, 2008). A adição fraudulenta de bicarbonatos no leite pode causar alterações na sinalização hormonal que regula o desenvolvimento e a reprodução do indivíduo (AZAD & AHMED, 2016).

Outro tipo de fraude relacionada à contaminação do leite é a adição de conservantes, como peróxido de hidrogênio e formaldeído, com o objetivo de reduzir ou eliminar os microrganismos presentes no leite, prevenindo as alterações decorrentes de sua multiplicação. A adição de peróxido de hidrogênio no leite promove alterações na sua qualidade nutricional, reduzindo de forma significativa as concentrações das vitaminas A, B1 e C. Em baixas concentrações, apresenta rápida degradação no leite devido à ação de enzimas naturalmente presentes. Em grandes quantidades, os peróxidos no leite podem causar complicações gastrointestinais, que podem levar à gastrite e inflamação do intestino. Já o formaldeído é uma molécula altamente reativa e irritante para o trato respiratório, olhos, pele e sistema gastrointestinal, além de ser caracterizada como substância cancerígena (AZAD & AHMED, 2016). Para a indústria, a adição dessas substâncias é particularmente prejudicial, pois podem interferir na produção de derivados, inibindo a multiplicação das culturas láticas adicionadas durante a sua fabricação, demonstrando as falhas no controle de qualidade da indústria e dos órgãos fiscalizadores (TRONCO, 2008).

- **Ocorrências de adulterantes em leite cru**

No Brasil, várias pesquisas demonstram a ocorrência de fraudes no leite cru. No ano de 2007, foi deflagrada pela Polícia Federal a Operação Ouro Branco, na qual foi desvendado um esquema de fraude no leite praticado por duas cooperativas mineiras. O leite era adicionado de uma solução química composta por citrato de sódio, hidróxido de sódio, cloreto de sódio, sacarose, fosfatos, bicarbonatos e peróxido de hidrogênio, com o objetivo de mascarar os defeitos na qualidade microbiológica do leite e aumentar o volume total (BOTELHO, 2015).

No estudo de FREITAS FILHO *et al*. (2009), realizado na cidade de Garanhuns – PE, foi detectada a presença de cloro (13%) e peróxido de hidrogênio (20%) em 15 amostras de leite cru analisadas naquela região.

FIRMINO e colaboradores (2010) avaliaram a qualidade físico-química e a presença de adulterantes no leite cru da região de Rio Pomba - MG. Das 20 amostras analisadas, 48,3% foram positivas para a presença de formol, nitrato, cloreto, sacarose, urina e soro lácteo.

No trabalho de MENDES *et al*. (2010), a qualidade do leite informal comercializado no município de Mossoró – RN foi avaliada por meio de análises físico-químicas e de pesquisa de fraudes. Foram analisadas 32 amostras, das quais 50% apresentaram alterações nos valores de crioscopia, indicando fraude por adição de água.

No estudo de BERSOT *et al*. (2010), realizado em Pelotina – PR, foi detectada a adição de água em 20% das 30 amostras de leite cru analisadas, além de outras alterações de qualidade físico-química e microbiológica, como acidez elevada e alta contagem bacteriana.

MONTANHINI & HEIN (2013) avaliaram 23 amostras de leite cru no município de Piraí do Sul – PR, das quais 34,78% apresentaram alterações de densidade, sugerindo a fraude por adição de água.

No trabalho de SANDA *et al* (2013), foi avaliada a qualidade do leite clandestino distribuído na cidade de Pires do Rio – GO, onde também foram observadas amostras com indício de fraude por adição de água e de substâncias neutralizantes.

No ano de 2013, a Polícia Federal iniciou a Operação Leite Compensado, deflagrada em sucessivas etapas pelo Ministério Público do Rio Grande do Sul (MP-RS), com a participação do Grupo de Atuação Especial de Combate ao Crime Organizado (GAECO), Ministério da Agricultura, Pecuária e Abastecimento (MAPA), Receita Estadual e Fundação Estadual de Proteção Ambiental (FEPAM). As empresas envolvidas foram acusadas de crime organizado e comercialização de leite e derivados lácteos impróprios para consumo humano, incluindo a utilização de fertilizantes contendo ureia e formaldeído (BOTELHO, 2015).

No exterior, um dos mais graves exemplos de fraude no leite ocorreu no ano de 2008 na China, que tragicamente exportou produtos lácteos contaminados a outros países da Ásia. A fraude consistia na adição da substância melamina (2,4,6-triamino-1,3,5-triazina), utilizada na produção de resinas, colas e plásticos, com o objetivo de aumentar o teor de proteína, por meio da dosagem do nitrogênio total, usualmente mensurada por métodos padrão, como Kjeldahl ou Dumas. Como a molécula de melamina é constituída por aproximadamente 66% de nitrogênio, foi possível mascarar a grande quantidade de água adicionada ao leite, processado posteriormente a leite em pó. A ingestão dessa substância pode levar à falência renal e, eventualmente, à morte,

principalmente de indivíduos vulneráveis, como os infantes. Foram reportadas mais de 300 mil crianças com alterações renais naquele país, das quais 6 vieram a óbito (SHARMA & PARADAKAR, 2010; JAWAID *et al*., 2013; YANG *et al*., 2013; ZOU *et al*., 2014).

- **Métodos de detecção de fraudes em leite**

Os testes qualitativos clássicos para detectar a adulteração no leite, estabelecidos como métodos oficiais pela autoridade reguladora, incluem determinações independentes para cada analito (BRASIL, 2018). Estes métodos de bancada exigem uma grande quantidade de testes e reagentes, consomem tempo e geram grandes quantidades de resíduos. Por este motivo, estão sendo substituídos por técnicas instrumentais não destrutivas, que consomem menos reagentes, geram menos resíduos e podem detectar vários analitos simultaneamente, economizando tempo e recursos (SOUZA *et al*., 2011; SILVA *et al*., 2013; BOTELHO *et al*., 2015; GONDIM *et al*., 2015; SILVA *et al*., 2015).

Os métodos baseados nas técnicas espectroscópicas, combinadas às técnicas quimiométricas multivariadas, oferecem estas vantagens e se tornam ferramentas poderosas para a determinação dos adulterantes no leite (ZOU *et al*., 2014; BOTELHO *et al*., 2015). Uma das técnicas mais usadas na indústria alimentar é a espectroscopia no infravermelho, cujas vantagens incluem a análise de amostras com pouca ou nenhuma preparação, facilidade de uso, rápida obtenção de dados e uso como técnica de "impressão digital" (SOUZA *et al*., 2011; JAWAID *et al*., 2013; GONDIM *et al*., 2017).

O pressuposto básico por trás da aplicação de técnicas espectroscópicas depende da geração da "impressão digital" do analito. Um produto lácteo específico como o leite, que apresenta uma determinada composição química, quando exposto a uma fonte de luz, apresentará um espectro característico, que é resultado da absorção por vários constituintes químicos diferentes. Como uma "impressão digital", cada molécula apresentará o seu próprio espectro na região do infravermelho, tornando o método viável para identificar diferentes tipos de amostras. Devido às pequenas variações de composição exata, é necessário comparar o espectro do analito a uma biblioteca de

espectros representativos, a fim de estabelecer sua qualidade ou autenticidade (KAROUI & BAERDEMAERKER, 2007).

A espectroscopia no infravermelho mede o comprimento de onda e a intensidade da luz infravermelha absorvida por uma amostra. Este método é baseado nas vibrações dos átomos de uma molécula. O espectro infravermelho de uma amostra é registrado pela passagem de um feixe de luz infravermelha através da mesma e, em seguida, pela determinação de qual parte da radiação incidente é absorvida em um comprimento de onda particular. A energia na qual aparece um espectro de absorção em picos corresponde à freqüência de uma vibração da molécula da amostra. Esta técnica permite determinar grupos funcionais presentes em uma amostra, uma vez que cada grupo funcional de uma molécula possui uma freqüência vibracional única. Ao levar em consideração os efeitos de todos os diferentes grupos funcionais, resulta-se um espectro representando uma "impressão digital" molecular única, que pode ser usada para confirmar a detecção da adulteração de uma amostra (KAMAL & KAROUI, 2015).

O infravermelho pode ser dividido em três regiões espectrais: o infravermelho próximo de alta energia (NIR) (~ 14.000 e 4.000 cm^{-1}), o infravermelho médio (MIR) (~ 4.000 e 400 cm^{-1}) e o infravermelho distante (~ 400 e 10 cm^{-1}). A espectroscopia infravermelho médio (MIR) pode ser dividida em quatro grandes regiões: a região de ligações simples (4.000 e 2.500 cm^{-1}), região de ligações triplas (2.500 e 2.000 cm^{-1}), região de ligações duplas (2.000 e 1.500 cm^{-1}) e a região de outras deformações de ligações (1.500 e 400 cm^{-1}). Uma transição entre os estados fundamentais leva a apenas um tipo de resposta vibracional, derivada da absorbância MIR. Portanto, o tipo particular de ligação orgânica é a representação exclusiva de picos existentes em um espectro de absorbância MIR (KAMAL & KAROUI, 2015).

Em estudos recentes, vários pesquisadores utilizaram a espectroscopia MIR para determinar a autenticidade e a detecção de adulteração nos produtos lácteos.

No estudo de SANTOS *et al*. (2013), foi avaliada a aplicação da refletância total atenuada (ATR), utilizando a microespectroscopia do infravermelho médio (MIR-*microspectroscopy*), na detecção e quantificação da adulteração no leite. As amostras foram adulteradas com soro lácteo, peróxido de hidrogênio, urina sintética, ureia e leite

sintético (mistura de óleos vegetais emulsificantes com detergentes e ureia), em cinco concentrações diferentes. Os resultados mostraram que a microespectroscopia do infravermelho médio detectou diferenças entre os espectros das amostras fraudadas e amostras controle, indicando que esta pode ser uma ferramenta simples, rápida e não destrutiva para a detecção e quantificação da adulteração do leite.

Em Botelho *et al*. (2015), foi proposto um novo método de triagem para a detecção simultânea dos adulterantes amido, citrato de sódio, formaldeído, água e sacarose em leite usando espectroscopia de infravermelho médio atenuada (ATR). No estudo de Coitinho e colaboradores (2017), foi utilizado um equipamento compacto (*MilkoScan FT1*) que adota uma metodologia de espectroscopia de infravermelho por transformada de Fourier para monitorar a adulteração em leite cru. As amostras foram adulteradas com amido, bicarbonato de sódio, citrato de sódio, formaldeído, sacarose e água. Já em Gondim *et al*. (2017), foi proposta uma estratégia sequencial para detectar os seguintes adulterantes utilizando espectroscopia de infravermelho médio: formaldeído, peróxido de hidrogênio, bicarbonato, carbonato, cloreto, citrato, hidróxido, hipoclorito, amido, sacarose e água.

Além da espectroscopia, outras técnicas para detecção e quantificação de adulterantes no leite foram descritas pela literatura, tais como: cromatografias líquida e gasosa, eletroforese capilar, espectrometria de massa, ressonância magnética nuclear, ensaio imuno-absorvente ligado a enzima (ELISA), reação em cadeia da polimerase (PCR), medições de condutância de frequência única, sensor piezoelétrico baseado em enzima, língua eletrônica, isótopo estável, biossensor potenciométrico, imagens digitais combinadas com parâmetros de cores, sensor de elemento de fase constante, dentre outras (KAROUI & BAERDEMAERKER, 2007; SANTOS *et al*., 2013; DAS *et al*., 2015; KAMAL & KAROUI, 2015; AZAD & AHMED, 2016; POONIA *et al*., 2017).

- **Perspectivas futuras**

Espera-se que o desenvolvimento das técnicas analíticas enfatize não apenas o refinamento de métodos existentes, mas também os estágios anteriores de preparação de amostras. O futuro das técnicas de autenticação da qualidade do leite abrange o campo das análises automatizadas, evitando protocolos complexos. Acredita-se que a

evolução dos métodos oficiais para as análises de adulteração no leite seja baseada em técnicas instrumentais modernas. Dada a heterogeneidade de alguns problemas, as estatísticas multivariadas tradicionais devem ser substituídas por algoritmos eficientes para auxiliar no reconhecimento de padrões e na classificação de produtos autênticos ou adulterados. O futuro está ligado ao crescente desenvolvimento de soluções analíticas que combinam poderosos dispositivos analíticos a softwares de processamento de dados eficazes (POONIA *et al*., 2017).

3.2 Espectroscopia

A espectroscopia compreende o estudo da interação da radiação eletromagnética, cujo principal objetivo é a determinação dos níveis de energia e transições de espécies atômicas e moleculares. As transições eletrônicas estão associadas a mudanças de orbital atômico, *spin* eletrônico e momento angular total. Já para as moléculas, onde existem as ligações químicas, as transições eletrônicas envolvem mudança de energia dos elétrons de valência, ou seja, variações nas populações eletrônicas de orbitais moleculares. Em razão da existência das ligações químicas, as moléculas também possuem energia vibracional e rotacional (SKOOG *et al*., 2002). Geralmente as transições vibracionais estão relacionadas com a região do infravermelho (SALA, 2008).

O astrônomo William Herschel descobriu a radiação infravermelha em 1800. Sabendo que a luz solar continha todas as cores do espectro e que era também uma fonte de calor, Herschel queria descobrir quais eram as cores responsáveis pelo aquecimento dos objetos. Ele idealizou então um experimento utilizando um prisma e termômetros para medir as temperaturas das diferentes cores. À medida que movia o termômetro do violeta para o vermelho no espectro criado pela luz do sol atravessando o prisma, ele observou um aumento da temperatura, sugerindo que a temperatura mais alta ocorria além da luz vermelha. Como a radiação que causou esse aquecimento não era visível, Herschel denominou essa radiação invisível de infravermelho (BIGGS *et al.*, 1987; SILVEIRA, 2004).

Uma das primeiras aplicações da espectroscopia no infravermelho, como ferramenta analítica, foi durante o período da Segunda Guerra Mundial, onde já se sabia que os

espectros armazenavam diversas informações sobre a amostra e apresentavam um elevado potencial para serem empregados nos mais diversos tipos de análises químicas e/ou físicas. Porém era praticamente impossível extrair informações quantitativas ou reforçar hipóteses sobre a estrutura química das espécies. Nos meados dos anos oitenta, com o desenvolvimento da microeletrônica e a popularização dos microcomputadores, houve um significativo avanço nas análises instrumentais possibilitando a aquisição de um grande número de dados de uma mesma amostra. Entretanto, a modelagem desses dados tornou-se mais complexa e só foi solucionada com a aplicação de técnicas quimiométricas, as quais contribuíram para a utilização da espectroscopia como ferramenta de análise em aplicações qualitativas e quantitativas na química analítica (COSTA FILHO & POPPI, 2002).

Como técnica de análise quantitativa para o controle de alimentos e produtos industrializados, a espectroscopia no infravermelho obteve crescimento após o advento da Transformação de Fourier e a utilização do interferômetro de Michelson, o que tornou o método mais rápido e robusto (SALIBA *et al.*,2003).

A região do infravermelho médio e as técnicas de refletância para análises em alimentos vêm sendo amplamente utilizadas desde o início dos anos noventa. Nesta década, inúmeros estudos revelaram as aplicações da espectroscopia na análise de leite, carne, óleos, gorduras e frutas, avaliando aspectos quantitativos e qualitativos, bem como para obtenção de certificação da qualidade (FERRÃO *et al*., 2004).

Nas últimas décadas, foram realizados vários trabalhos utilizando as técnicas de espectroscopia no infravermelho em conjunto com métodos de calibração multivariada, com diferentes propósitos, dentre as mais diversas áreas da ciência (BORIN *et al*., 2006; NICOLAOU *et al*., 2010; SANTOS *et al*., 2013; ZHANG *et al*, 2014; CAPUANO *et al*., 2015; CARVALHO *et al*., 2015). Este acentuado crescimento e aperfeiçoamento nos últimos anos revelam o contínuo interesse por métodos analíticos rápidos, não destrutivos e não invasivos, com a não formação de subprodutos químicos tóxicos.

Da mesma forma, os métodos espectroscópicos, juntamente com as ferramentas quimiométricas, têm sido cada vez mais aplicados à indústria de produtos lácteos, como uma alternativa para substituir os procedimentos de referência. Com análises mais rápidas e precisas, objetiva-se atender ao aumento da demanda de mercado do leite e

garantir os requisitos e parâmetros de qualidade e inocuidade, previstos em legislação específica.

- **Espectroscopia no infravermelho**

Várias técnicas permitem obter informações sobre a estrutura molecular, níveis de energia e ligações químicas, podendo-se citar como exemplo: espectroscopia Raman e infravermelho. A espectroscopia estuda a interação da radiação eletromagnética com a matéria, sendo um dos seus principais objetivos a determinação dos níveis de energia de átomos ou moléculas. Quase todos os compostos que tenham ligações covalentes, sejam orgânicos ou inorgânicos, absorvem várias frequências de radiação eletromagnética na região do infravermelho (SALA, 2008).

A energia denominada infravermelho corresponde à região do espectro eletromagnético situada na faixa de números de onda entre 14.290 e 200 cm^{-1}. A região que apresenta número de ondas entre 4000 – 400 cm^{-1} é a mais comumente utilizada pela química orgânica, sendo denominada infravermelho médio (BARBOSA, 2013). Em resumo, a região de infravermelho compreende comprimentos de onda de 0,78 μm a 30 μm no espectro infravermelho. Essa faixa é dividida em três regiões, de acordo com o comprimento de onda ou o número de ondas da radiação, de acordo com a tabela abaixo (MORGANO *et al*., 2005).

Tabela 1: Divisões da região do infravermelho.

Região	**Intervalo de número de onda (v) – (cm^{-1})**	**Região em comprimento de onda (λ) – (μm)**	**Região de frequência (v) – (Hz)**	**Amostra a que se aplica**
Próximo (NIR)	12.800 – 4.000	780 – 2500	$3,8 \times 10^{14}$ a $1,2 \times 10^{14}$	Materiais comerciais sólidos ou líquidos e misturas gasosas.
Médio (MIR)	4.000 – 400	2500 – 5000	$1,2 \times 10^{14}$ a $6,0 \times 10^{12}$	Sólidos, líquidos ou gases puros, misturas complexas de líquidos, sólidos ou gases.
Distante (FIR)	200 – 10	5000 – 100000	$3,8 \times 10^{12}$ a $3,0 \times 10^{11}$	Espécies inorgânicas ou organometálicas puras e amostras atmosféricas.

Fonte: Adaptado de WILLIAMS, 1987 e HOLLER *et al*., 2009.

Os espectros no infravermelho podem ser obtidos a partir de amostras sólidas, líquidas e gasosas, e a qualidade desses espectros depende do método de preparo da amostra e do tipo de acessório utilizado para sua obtenção. Os métodos mais empregados envolvem transmitância e refletância, sendo este último muito utilizado para análises qualitativas e quantitativas. O método por refletância pode apresentar acessórios para obtenção de espectros por Refletância Total Atenuada (ATR), Refletância Especular e Refletância Difusa (SKOOG *et al*., 2002).

A absorção de radiação IR (região espectral do infravermelho) é limitada principalmente a espécies moleculares que possuem pequenas diferenças de energia entre diversos estados vibracionais e rotacionais. Para absorver a radiação IR, uma molécula deve ser submetida a uma variação no momento dipolo durante seu movimento rotacional ou vibracional. Apenas sob estas circunstâncias, o campo elétrico alternado da radiação pode interagir com a molécula e causar variações na amplitude de um de seus movimentos (HOLLER *et al*., 2009).

Os espectros fornecem as transições (diferenças de energia entre os níveis) e a partir destas medidas determinam-se as posições relativas dos níveis energéticos. No caso de moléculas, a região espectral onde estas transições são observadas depende do tipo de níveis envolvidos: eletrônicos, vibracionais ou rotacionais. Normalmente as transições eletrônicas estão situadas na região do ultravioleta ou visível, as vibrações na região do infravermelho e as rotacionais na região de micro-ondas (SALA, 2008).

Assim como ocorre em outros tipos de absorção de energia, as moléculas, quando absorvem radiação no infravermelho atingem um estado de maior energia. A absorção de radiação no infravermelho é, como outros processos de absorção, um processo quantizado. No processo de absorção são absorvidas as frequências de radiação no infravermelho, que equivalem às frequências vibracionais de estiramento e dobramento das ligações na maioria das moléculas mais covalentes (PAVIA *et al*., 2012).

Uma maneira indireta de observar os espectros vibracionais, transferindo para a região do visível as informações que seriam obtidas no infravermelho, é através do espalhamento Raman, ou seja, do espalhamento inelástico de radiação eletromagnética monocromática que interage com as moléculas. As frequências vibracionais são determinadas pelas diferenças entre as frequências das radiações espalhadas e a da radiação incidente (SALA, 2008).

No processo de absorção são absorvidas as frequências de radiação no infravermelho que equivalem às frequências vibracionais naturais da molécula em questão, e a energia absorvida serve para aumentar a amplitude dos movimentos vibracionais das ligações na molécula. Observa-se, contudo, que nem todas as ligações em uma molécula são capazes de absorver energia no infravermelho, mesmo que a frequência de radiação seja exatamente igual a do movimento vibracional. Apenas as ligações que tem momento dipolo que muda em função do tempo são capazes de absorver radiação no infravermelho (PAVIA *et al*., 2012).

A espectroscopia no infravermelho se baseia na determinação de grupos funcionais de uma amostra, sendo que cada grupo absorve energia em frequências características. Essas frequências apresentam vibrações específicas, que podem ser de estiramento ou de deformação, as quais correspondem a níveis de energia da molécula (SILVERSTEIN *et al.*, 1994; BIGGS *et al*., 1987). Os grupos carbonila (C=O) das ligações éster dos triglicerídeos absorvem radiação no comprimento de onda de 5,73 μm, os grupos amida (CONH) das ligações peptídicas das proteínas em 6,46 μm e os grupos hidroxila (OH) da lactose em 9,53 μm. A quantidade de sólidos totais presente em uma amostra pode ser determinada pelo somatório do conteúdo dos componentes: gordura, proteína e lactose, acrescidos de uma constante de minerais, ou pela diferença da absorção de radiação em um comprimento de onda de 4,3 μm dos grupos hidroxila das moléculas de água (BIGGS *et al.*, 1987; SILVEIRA, 2004).

- **Reflexão no infravermelho por metodologia FTIR**

O avanço da espectroscopia no infravermelho médio, como técnica para análise quantitativa, deve-se principalmente à combinação da Transformação de Fourier e da nova geometria dos espectrofotômetros com a utilização do interferômetro de Michelson, tornando-os mais rápidos (DURIG e SULLIVAN, 1990; EIKREM, 1990; KALASINSKY, 1990; COATES, 1998). Um esquema ilustrativo para o interferômetro de Michelson do espectrofotômetro FTIR pode ser visualizado na Figura 01.

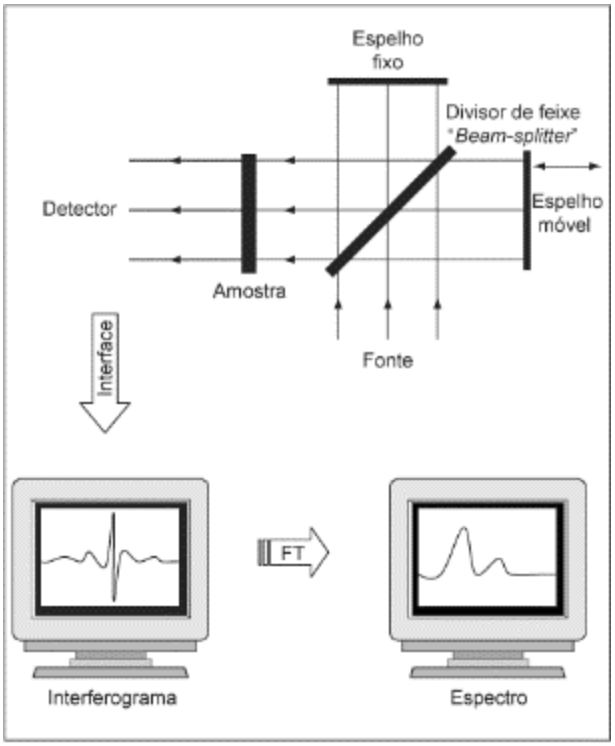


Figura 1: Esquema ilustrativo para o interferômetro de Michelson e do espectro resultante da aplicação da Transformação de Fourier. Fonte: WEBSTER et al., 2000.

O interferômetro de Michelson consiste basicamente em dois espelhos (um fixo e um móvel) e um divisor de feixe (*beam-splitter*), que transmite 50% da radiação incidente da fonte para o espelho móvel e reflete os outros 50% para o espelho fixo. Os espelhos, por sua vez, refletem os dois feixes para o divisor, onde se recombinam. Se os dois espelhos encontram-se equidistantes do divisor, as amplitudes combinam-se construtivamente. Se o espelho móvel mover-se a uma distância de $\lambda/4$ do divisor, as

amplitudes combinam-se destrutivamente. Para a radiação no infravermelho (policromática), a soma de todas as interações construtivas e destrutivas para cada componente resulta num sinal complexo denominado interferograma. Após a aquisição do interferograma, é aplicada a Transformação de Fourier que converte os dados obtidos no interferômetro em um espectro que relaciona a intensidade *versus* frequência (número de onda), como ilustrado na Figura 02 (MORGANO *et al*., 2005; BOTELHO *et al*., 2015).

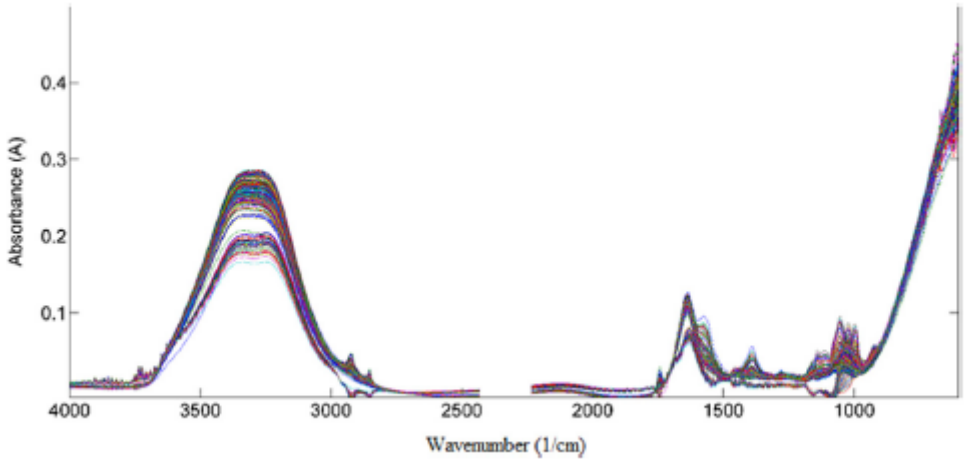


Figura 2: Espectros de amostras de leite obtidos por FTIR. Fonte: BOTELHO et al., 2015.

Como uma "impressão digital", cada molécula apresentará o seu próprio espectro na região do infravermelho, tornando o método viável para identificar diferentes tipos de amostras (análise qualitativa). Os picos presentes no gráfico do espectro correspondem às frequências de vibrações entre os átomos que compõem cada amostra. A altura desses picos corresponde à quantidade de determinada amostra (análise quantitativa) (MORGANO *et al*., 2005).

O espectrofotômetro com transformação de Fourier expõe a amostra a um único pulso de radiação e de medidas de resposta. O sinal resultante, chamado de indução bifásica, é uma medida direta da coerência temporal da luz e contém uma rápida decadência composta de todas as possíveis frequências. Como o sinal medido no interferômetro não pode ser interpretado diretamente, é necessária a técnica matemática chamada de Transformação de Fourier. Esta transformação é realizada pelo computador (algoritmos), apresentando ao usuário as informações desejadas para a análise espectral

(MORGANO *et al*., 2005).

- **Dados multivariados**

Os métodos de análise multivariada são assim chamados, pois, no caso em que técnicas espectroscópicas no infravermelho são empregadas, é possível manipular dados de absorvância espectral associados a mais de uma frequência ao mesmo tempo. Estes métodos têm tornado possível modelar propriedades químicas e físicas de amostras a partir de seus dados espectroscópicos (WILLIAMS *et al*., 1987).

As análises qualitativas e quantitativas utilizando espectroscopia na região do infravermelho expandiram-se a partir do momento em que os dados gerados por um espectrofotômetro FTIR puderam ser digitalizados, habilitando os métodos estatísticos na resolução de problemas de análise química. Inicialmente, muitas amostras não comportavam o isolamento de uma banda para análise, o que tornava necessário o uso de métodos de separação química tal como a cromatografia. Porém, a possibilidade de utilizar várias frequências do espectro tem aumentado o tipo de amostras que podem ser quantificadas por espectroscopia no infravermelho (HELFER *et al.*, 2006).

Nos métodos clássicos de análise univariável, somente a absorvância de uma frequência é associada à concentração, enquanto que métodos que usam simultaneamente duas ou mais frequências são conhecidos como métodos multivariáveis. A precisão dos métodos univariáveis é dependente da capacidade para identificar uma única banda isolada para cada componente. Os métodos multivariáveis, entretanto, podem ser utilizados igualmente quando estão sobrepostas informações espectrais de vários componentes através de várias regiões espectrais selecionadas (BIGGS *et al*., 1987).

Os dados multivariados geralmente são dispostos numa matriz **X** de valores, correspondendo a **m** variáveis para **n** amostras, conforme a figura 03.

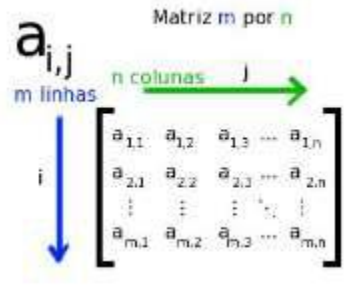


Figura 3: Representação de uma matriz de dados. Fonte: HELFER et al., 2006.

3.3 Mineração de dados

A mineração de dados (em inglês, *data mining*) é o processo de encontrar anomalias, padrões e correlações em grandes conjuntos de dados para prever resultados. A mineração de dados extrai informações de um conjunto de dados e o transforma em uma estrutura compreensível. É o processo computacional que envolve métodos na interseção de inteligência artificial, aprendizado de máquina, estatística e sistemas de banco de dados. A tarefa de mineração de dados real é a análise automática ou semi-automática de grandes quantidades de dados para extrair padrões interessantes previamente desconhecidos (KESAVARAJ & SUKUMARAN, 2013).

As técnicas de classificação em mineração de dados são capazes de processar uma grande quantidade de dados, podem ser utilizadas para prever rótulos de classe categórica, classificar os dados com base no conjunto de treinamento e nos rótulos de classe, e classificar dados desconhecidos. O termo pode abranger qualquer contexto em que alguma decisão ou previsão é feita com base em informações previamente disponíveis. O procedimento de classificação é o método reconhecido para fazer tais decisões repetidamente em novas situações (NIKAM, 2015).

A criação de um procedimento de classificação a partir de um conjunto de dados para os quais as classes exatas são conhecidas antecipadamente é denominado reconhecimento de padrões ou aprendizado supervisionado. Três principais linhas históricas de pesquisa podem ser identificadas: a estatística, o aprendizado de máquina e a rede neural. Esses três grupos têm como objetivos em comum tentar desenvolver procedimentos que possam trabalhar com uma ampla variedade de problemas, serem extremamente genéricos e usados em ambientes práticos com eficácia (NIKAM, 2015).

- **Procedimentos estatísticos**

Duas fases principais de trabalho na classificação podem ser identificadas na análise estatística. A primeira fase "clássica" concentrou-se na extensão do trabalho inicial de Fisher sobre a discriminação linear. A segunda fase, "moderna", concentrou-se em classes de modelos mais flexíveis, muitas das quais tentam fornecer uma estimativa da distribuição conjunta das características dentro de cada classe, o que pode, por sua vez, fornecer uma regra de classificação. Os procedimentos estatísticos são geralmente caracterizados por ter um modelo preciso de probabilidade fundamental que fornece uma probabilidade de estar em cada classe, em vez de apenas uma classificação. Geralmente, as técnicas serão empregadas por profissionais estatísticos e, portanto, requerem algum envolvimento humano com relação à seleção e transformação das variáveis, e à estruturação geral do problema (MICHIE *et al*., 1994). São exemplos de metodologias tradicionais a regressão linear, a regressão logística e a regressão por mínimos quadrados parciais, ou PLS (*Partial Least Squares*).

A análise de regressão estuda a relação entre a variável dependente e outras independentes, sendo a relação entre elas representada por um modelo matemático, o qual será designado por modelo de regressão linear simples, se definir uma relação linear entre a variável dependente e uma variável independente, ou múltipla, se forem incorporadas várias variáveis independentes. Com base nos dados obtidos, constrói-se um diagrama de dispersão, que deve apresentar uma tendência linear para que este método seja empregado, sendo o coeficiente de correlação linear uma medida do grau de dependência entre as variáveis. Já a regressão logística permite a predição de valores tomados por uma variável dependente categórica, de natureza dicotômica, frequentemente binária. Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente, sendo os coeficientes estimados pelo método da máxima verossimilhança (HASTIE *et al*., 2017).

A regressão parcial por mínimos quadrados (PLS) é um modelo que diminui os preditores a um conjunto reduzido de componentes não correlacionados, e executa neste conjunto a regressão de mínimos quadrados, ao invés dos dados originais. A regressão PLS é aplicada principalmente quando os preditores são colineares, ou quando existem

mais preditores do que observações, e os coeficientes gerados apresentam altos erros padrões. Além disso, a PLS não assume preditores fixos, ao contrário da regressão múltipla, podendo medi-los com erro e tornando a PLS mais robusta à incerteza da medição (HASTIE *et al*., 2017).

Muitos trabalhos são descritos na literatura utilizando esta metodologia tradicional para a avaliação da qualidade e da autenticidade do leite e produtos derivados. No estudo de Santos e colaboradores (2013), a aplicação de microespectroscopia no infravermelho médio atenuado (MIR - *microspectroscopy*) foi avaliada como um método rápido para detecção e quantificação da adulteração do leite. A análise de reconhecimento de padrões foi realizada por *Soft Independent Modeling of Class Analogy* (SIMCA) e PLS.

No trabalho de Botelho *et al*. (2015), foi proposto um método de triagem para a detecção simultânea de cinco adulterantes em leite, usando espectroscopia de infravermelho médio atenuada (*Attenuated Total Reflection* - ATR) e classificação supervisionada por *Partial Least Squares Discrimination Analysis* - PLSDA, juntamente com uma validação qualitativa multivariada.

Coitinho e colaboradores trabalharam com os dados obtidos por um equipamento compacto (*MilkoScan* FT1), que adota uma metodologia de espectroscopia de infravermelho por transformada de Fourier, para calibrar e validar o monitoramento da adulteração em leite, através da análise de componentes principais (*Principal Component Analysis* - PCA). Já Gondim *et al*. (2017) propuseram uma estratégia sequencial para detectar adulterantes em leite por espectroscopia de infravermelho médio, utilizando PCA e SIMCA.

A metodologia PLS também foi utilizada para a análise de leite ou produtos derivados nos estudos de Kasemsumran *et al*. (2007), Nicolaou *et al*. (2010), Santos *et al*. (2012), Jawaid *et al*. (2013), Yang *et al*. (2013), Zou *et al*. (2014), Capuano *et al*. (2015) Carvalho *et al*. (2015) e Luna *et al*. (2016). Além dessas, também foram descritas, nesta mesma linha de trabalho, as metodologias de regressão linear (OLIVEIRA *et al*., 2012; SANTOS *et al*., 2012), *Support Vector Machine* – SVM (BORIN *et al*., 2006; ZOU *et a*l., 2014), *Hierarchical Cluster Analysis* – HCA (SOUZA *et al*., 2011) e *Simplified K Nearest Neighbours* – IS-KNN (ZHANG *et al*., 2014).

- **Aprendizado de máquina**

O Aprendizado de Máquina geralmente envolve procedimentos automáticos de computação, baseados em operações lógicas ou binárias, que aprendem uma tarefa a partir de uma série de exemplos. Neste caso, objetiva-se a classificação fundamentada nas abordagens de árvore de decisão, nas quais a classificação resulta de uma sequência de etapas lógicas. Estes resultados de classificação são capazes de representar o problema mais complexo, desde que sejam fornecidos dados suficientes. A aprendizagem automática tem como objetivo gerar expressões classificatórias simples o suficiente para ser compreendida facilmente, e deve mimetizar o raciocínio humano sobre o processo de decisão. O conhecimento prévio pode ser usado no desenvolvimento assim como nas abordagens estatísticas tradicionais, mas a operação é realizada sem a interferência humana (MICHIE *et al*., 1994; NIKAM, 2015). *Gradient Boosting Decision Trees* e *Random Forests* são exemplos de métodos de aprendizado de máquina que utilizam conjuntos de árvores de decisão.

- ***Gradient Boosting Decision Tree***

Gradient Boosting Decision Tree (GBDT) é um algoritmo de aprendizado de máquina amplamente utilizado devido à sua eficiência, precisão e interpretabilidade. O GBDT alcança performances de ponta em muitas tarefas de aprendizado de máquina, como classificação de várias classes e aprendizado de classificação.

O GBDT é um modelo conjunto de árvores de decisão que são treinadas em sequência. Em cada iteração, o GBDT aprende as árvores de decisão ajustando os gradientes negativos, também conhecidos como erros residuais (KE *et al*., 2017).

- ***Random Forests***

A técnica de *bagging* ou *bootstrap aggregation* é utilizada para reduzir a variação de uma função de previsão estimada. O *bagging* se aplica especialmente bem para procedimentos de alta variância e baixa polarização, como as árvores de decisão. Para a classificação, a mesma árvore é ajustada várias vezes para inicializar as versões amostradas dos dados de treinamento e calcular a média do resultado.

A técnica de R*andom Forests* (RF) é uma modificação substancial de *bagging* que constrói uma coleção de árvores não correlacionadas e, em seguida, calcula a média

entre elas. Em muitos problemas, o desempenho de *random forests* é muito semelhante ao *boosting*, e são mais simples de treinar e ajustar. Como consequência, as *random forests* são populares e implementadas em vários pacotes computacionais (HASTIE *et al*., 2017). Todavia, não foram encontrados estudos prévios utilizando esta metodologia para a identificação de adulteração no leite cru.

- **Redes neurais artificiais**

As redes neurais artificiais (RNAs) podem ser definidas como estruturas compostas por elementos de processamento simples, adaptativos, densamente interconectados (chamados de neurônios artificiais ou nós), que são capazes de realizar cálculos paralelos para processamento de dados e representação de conhecimento (BASHEER & HAJMEER, 2000).

As RNAs são métodos computacionais com arquiteturas inspiradas em redes neurais biológicas (sistemas nervosos do cérebro) e são usadas para aproximar funções que podem depender de um grande número de entradas, geralmente desconhecidas. São apresentadas como sistemas de "neurônios" interconectados que podem calcular valores a partir de entradas, sendo capazes de aprendizado de máquina, assim como reconhecimento de padrões devido à sua natureza adaptativa (NIKAM, 2015).

A atratividade das RNAs vem das notáveis características de processamento de informações do sistema biológico, como não linearidade, alto paralelismo, robustez, tolerância a falhas, aprendizado, capacidade de lidar com informações imprecisas e difusas, e sua capacidade de generalizar. Os modelos artificiais que possuam tais características são desejáveis porque (i) a não linearidade permite melhor ajuste aos dados, (ii) a insensibilidade a ruídos fornece predição precisa na presença de dados incertos e erros de medição, (iii) o alto paralelismo implica em processamento rápido e tolerância a falhas de hardware, (iv) o aprendizado e a adaptabilidade permitem que o sistema atualize (modifique) sua estrutura interna em resposta a mudanças de ambiente, e (v) a generalização permite a aplicação do modelo a dados desconhecidos (BASHEER & HAJMEER, 2000). Por esses motivos, as redes neurais têm sido amplamente utilizadas em quimiometria para substituir os métodos tradicionais de calibração multivariada com base em modelos de regressão linear múltipla (ALVES DA ROCHA

et al., 2015). Basicamente, a regressão pode ser considerada uma camada da rede neural, que é capaz de modelar problemas não lineares e mais complexos.

O principal objetivo da computação baseada em RNA (neurocomputação) é desenvolver algoritmos matemáticos que permitam que as RNAs aprendam mimetizando o processamento de informações e a aquisição de conhecimento, como no cérebro humano. A analogia grosseira entre o neurônio artificial e o neurônio biológico se baseia na representação das conexões entre os nós pelos axônios e dendritos, enquanto que os pesos de conexão representam as sinapses, e o limiar se relaciona à atividade na soma de sinais. Como um neurônio tem um grande número de dendritos/sinapses, ele pode receber e transferir muitos sinais simultaneamente, que podem excitar ou inibir o disparo do neurônio. Esse mecanismo simplificado de transferência de sinal constituiu a etapa fundamental do desenvolvimento precoce da neurocomputação e da operação da unidade de construção das RNAs (BASHEER & HAJMEER, 2000).

A Figura 4 ilustra n neurônios biológicos com vários sinais de intensidade x e a força sináptica se alimentando em um neurônio com um limiar de b, e o equivalente sistema de neurônios artificiais. Tanto a rede biológica quanto a RNA aprendem ajustando incrementalmente as magnitudes dos pesos ou das forças das sinapses. *Perceptron* é uma rede neural de camada única, sendo que um *perceptron* de várias camadas é denominado rede neural artificial.

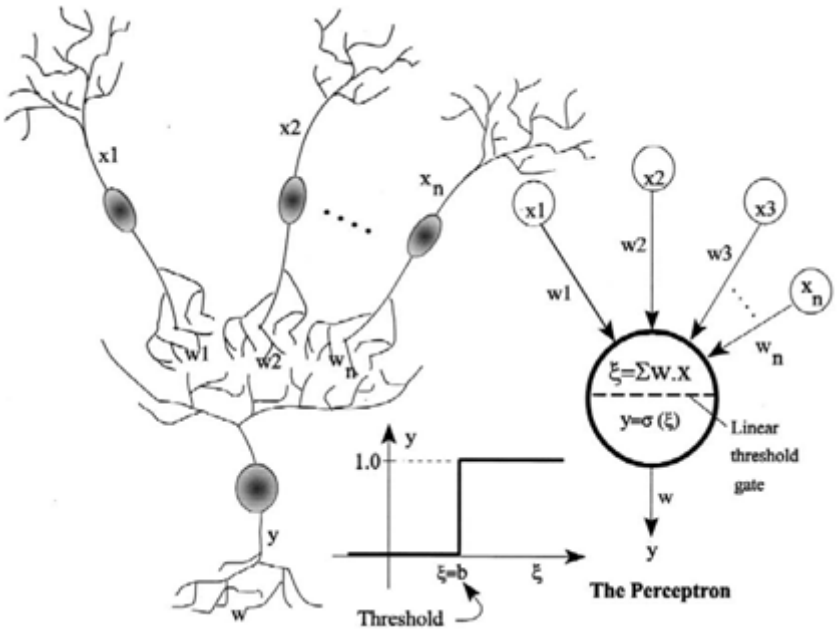


Figura 4: Interação de sinal de n neurônios e analogia à soma de sinal em um neurônio artificial compreendendo o perceptron de camada única. Adaptado de BASHEER & HAJMEER, 2000.

Uma rede neural artificial opera criando conexões entre muitos diferentes elementos de processamento, cada um correspondendo a um único neurônio em um cérebro biológico. Esses neurônios podem ser construídos ou simulados por um sistema digital de computador. Cada neurônio recebe muitos sinais de entrada, em seguida, com base em uma ponderação interna, produz um único sinal de saída, que é enviado como entrada para outro neurônio. Os neurônios estão fortemente interconectados e organizados em diferentes camadas. A camada de entrada recebe os sinais iniciais, e a camada de saída produz o resultado final. Em geral, uma ou mais camadas ocultas são inseridas entre as duas, tornando impossível prever ou conhecer o fluxo exato de dados (NIKAM, 2015).

A rede p*erceptron* de camada única consiste em uma única camada de nós de saída; as entradas são alimentadas diretamente às saídas através de uma série de pesos. A soma dos produtos dos pesos e das entradas é calculada em cada nó, e se o valor estiver acima de algum limiar (normalmente 0), o neurônio dispara e assume o valor ativado (normalmente 1). Caso contrário, leva o valor desativado (normalmente -1). Os neurônios com esse tipo de função de ativação também são chamados de neurônios artificiais ou unidades lineares limítrofes. Já as redes multicamadas consistem em várias camadas de unidades computacionais, geralmente interconectadas de maneira direta. As redes multicamadas usam uma variedade de técnicas de aprendizado, sendo a retropropagação (*Backpropagation*) uma das mais populares. Nesta última, os valores de saída são comparados com a resposta correta para calcular o valor de alguma função de erro pré-definida. Após retornar o erro para a rede, o algoritmo ajusta os pesos de cada conexão para reduzir o valor da função de erro. Depois de repetir este processo por vários ciclos de treinamento seguidos, a rede converge para um estado no qual o erro dos cálculos se torna pequeno (GOYAL & GOYAL, 2012). O padrão de treinamento de entrada e saída da RNA é ilustrado na Figura 5.

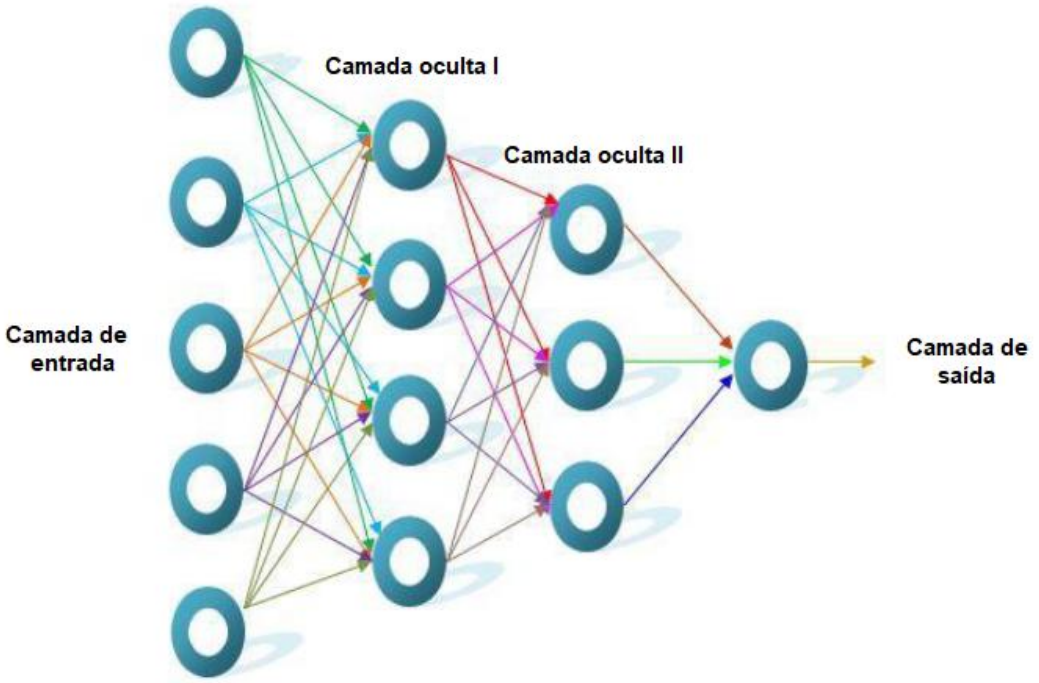


Figura 5: Modelo de camadas de entrada, ocultas e de saída das RNAs. Adaptado de GOYAL & GOYAL, 2012.

Na literatura, são descritos alguns estudos utilizando esta metodologia na análise do leite e de derivados lácteos. No trabalho de Cevoli e colaboradores (2011), por exemplo, as redes neurais artificiais foram utilizadas para classificar queijos Pecorinos de acordo com seu tempo de maturação e técnicas de fabricação, através de sensores de gás. Em Cossignani *et al*. (2011), os resultados de um método analítico químico-enzimático-cromatográfico foram elaborados por análise discriminante linear (LDA) e redes neurais artificiais (RNA), com o objetivo de caracterizar amostras de leite asinino e misturas com o leite de vaca. Em Valente *et al*. (2014), as redes neurais artificiais foram empregadas para classificar amostras de leite adulteradas com a adição de soro de queijo, a partir de análises de rotina de um laticínio. No estudo de Alves da Rocha e colaboradores (2015), foi utilizada a microscopia confocal Raman associada à rede neural artificial para avaliar e quantificar a adulteração de leite fluido pela adição de soro de leite. Enquanto que na pesquisa realizada por Conceição *et al*. (2019), foram utilizados os métodos de regressão linear múltipla e redes neurais artificiais nos dados obtidos por FTIR – ATR para identificar adulterações em leite cru. Já no estudo de Hernández-Ramos *et al*. (2019), as redes neurais artificiais foram empregadas para prever a contagem de células somáticas do leite de ovelha utilizado na produção de queijo.

- **Redes neurais convolucionais**

A Rede Neural Convolucional (ConvNet, *Convolutional Neural Network* ou CNN) é um algoritmo de aprendizagem profunda capaz de captar dados de entrada, atribuir importância, pesos e vieses que podem ser aprendidos, e diferenciá-los. O pré-processamento exigido pela CNN é muito menor em relação a outros algoritmos de classificação. Enquanto nos métodos anteriores os filtros são construídos manualmente, as CNN são capazes de aprendê-los com o treinamento suficiente.

A CNN é uma variação das redes de p*erceptrons* de múltiplas camadas, tendo sido inspirada no processo biológico de processamentos de dados visuais. Este tipo de rede vem sendo amplamente utilizado em diversas áreas (ZEILER & FERGUS, 2014; VARGAS *et al*., 2016; ZHANG *et al*., 2016; GUZHVA *et al*., 2018), principalmente nas aplicações de classificação, detecção e reconhecimento em imagens e vídeos. Entretanto, não foram encontrados na literatura trabalhos anteriores utilizando a CNN para a detecção de adulterantes em leite.

A CNN consiste em múltiplas partes com funções diferentes, na qual é comum aplicar camadas de convolução sobre o dado de entrada. Cada camada de convolução é composta por diversos neurônios, responsáveis pela aplicação de filtros e pesos. A combinação das entradas de um neurônio, utilizando os pesos respectivos de cada uma de suas conexões, produz uma saída passada para a camada seguinte. Os pesos atribuídos às conexões de um neurônio podem ser interpretados como uma matriz, que representa o filtro de uma convolução no domínio espacial, também conhecido como *kernel* ou máscara (VARGAS *et al*., 2016).

- **Algoritmos de classificação**

A classificação é uma das técnicas da mineração de dados utilizada principalmente para analisar um determinado conjunto de dados, de modo que o erro de classificação seja menor. É usada para extrair modelos que definem com precisão classes de dados importantes dentro do conjunto de dados determinado, e constitui em um processo de duas etapas. Durante a primeira etapa, o modelo é criado aplicando o algoritmo de

classificação no conjunto de dados de treinamento e, na segunda etapa, o modelo extraído é testado em relação a um conjunto de dados de teste predefinido para medir o desempenho e a precisão do modelo (NIKAM, 2015).

- ***Cross-validation***

O método mais simples e mais amplamente utilizado para estimar erros de previsão é a validação cruzada, ou *cross-validation*, que funciona dividindo os dados de treinamento aleatoriamente em *k* partes iguais (*k-fold cross validation*). O método de aprendizado é adequado a parte dos dados, e o erro de predição é computado na parte restante. Isso é feito para cada parte dos dados por vez, calculando-se a média das *k* estimativas de erro de previsão. Em seguida, obtêm-se uma curva de erro de previsão estimada, como uma função do parâmetro de complexidade (HASTIE *et al*., 2017).

Outra técnica utilizada é o *hold-out cross validation*, também conhecido como *test sample estimation*, que define proporções fixas de dados (como 90/10, 80/20 ou 70/30, por exemplo), para a execução do conjunto de treinamento e do conjunto teste (KOHAVI, 1995).

A validação cruzada é aplicada ao conjunto de treinamento, uma vez que selecionar o parâmetro de contração é parte deste processo. O objetivo do conjunto de testes é julgar o desempenho do modelo selecionado (HASTIE *et al*., 2017).

A figura 6 ilustra um exemplo visual de separação dos conjuntos de dados em treinamento e seleção por validação cruzada.

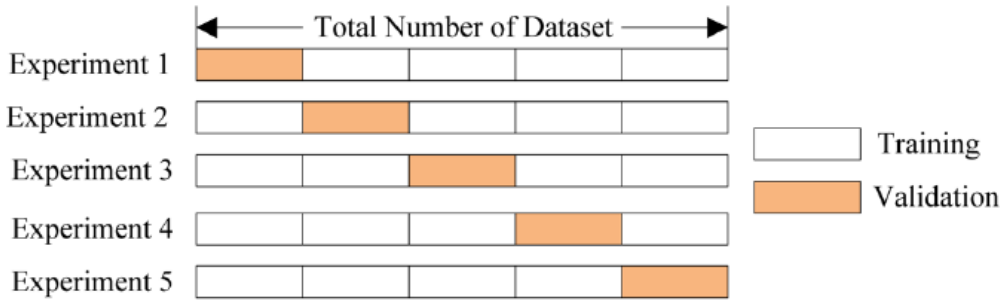


Figura 6: Exemplo visual de separação baseada em cross-validation. Fonte:

https://www.kaggle.com/dansbecker/cross-validation.

4. MATERIAL E MÉTODOS

4.1 Local de realização do experimento

Todas as análises foram realizadas no Laboratório de Análise da Qualidade do Leite (LabUFMG) da Escola de Veterinária da Universidade Federal de Minas Gerais (EV-UFMG), credenciado pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA) e certificado pelas normas ABNT NBR ISO/IEC 17025:2017.

4.2 Preparo das amostras e análises laboratoriais

O leite cru foi coletado da ordenha completa da primeira linha de ordenha, do primeiro lote de vacas, às 6:00h da manhã, da Fazenda Experimental da Escola de Veterinária da UFMG, localizada no município de Igarapé – MG.

O leite foi obtido do tanque de expansão, após agitação, sem mistura de leite de ordenhas anteriores. O volume de leite coletado foi de 30 litros por semana, durante 8 semanas, totalizando 240 litros. O leite foi transportado imediatamente após a coleta ao Laboratório de Análise da Qualidade do Leite – LabUFMG, em galões previamente higienizados e sanitizados com solução de hipoclorito de sódio a 200 ppm, seguido de enxágue.

Após a chegada ao laboratório, foi realizada a pesagem dos adulterantes sólidos em balança de precisão previamente calibrada, e mensuração dos adulterantes líquidos em micropipetas para a realização das adulterações.

As amostras foram adulteradas com sacarose, amido, bicarbonato de sódio, peróxido de hidrogênio e formol (reagentes P. A. - *Pro Analyse*), em três diferentes concentrações dos adulterantes: amido (0,1%, 0,5% e 1%); sacarose (0,1%, 0,5% e 1%); bicarbonato

de sódio (0,03%, 0,05%, 0,1%); peróxido de hidrogênio 29% (100ppm, 500ppm e 1.000ppm); e formol 37% (25ppm, 50ppm e 100ppm).

As concentrações dos reagentes foram calculadas com base na observação técnica das fraudes mais comumente praticadas no país, e em estudos anteriores (SANTOS *et al*., 2013; BOTELHO *et al*., 2015).

Os adulterantes foram diluídos no leite de forma padronizada, com mensuração do volume de leite em balão volumétrico calibrado, pesagem dos reagentes sólidos em balança de precisão e aspiração dos líquidos com micropipetas, conforme descrito anteriormente. As amostras foram homogeneizadas em béquer de vidro, com auxílio de bastão de vidro, em ambiente controlado.

Após a preparação das soluções, as amostras foram acondicionadas em frascos plásticos de 40 mL, adicionados de comprimidos conservantes Bronopol®, para análise de composição e crioscopia, e de Azidiol®, para análises complementares de CBT (BactoScan™ FC/*Foss)*. Foram montadas as *racks* para a leitura no aparelho em ordem crescente de concentração dos adulterantes, de forma padronizada, com uma amostra controle (branco) de leite sem nenhuma adição, no início de cada *rack*. Todos os frascos foram devidamente identificados com etiquetas à prova de água. Cada *rack* foi etiquetada com identificação, data e hora da execução das análises.

As amostras foram armazenadas nas temperaturas de 7ºC e 25°C, durante 0, 3, 24, 48, 72 horas e 07 dias, para cada adulterante utilizado. As amostras armazenadas a 7ºC foram mantidas na câmara fria do LabUFMG até o momento determinado de suas análises. As amostras armazenadas a 25°C foram mantidas em ambiente controlado nas dependências do laboratório. O esquema das adulterações é representado na figura 7.

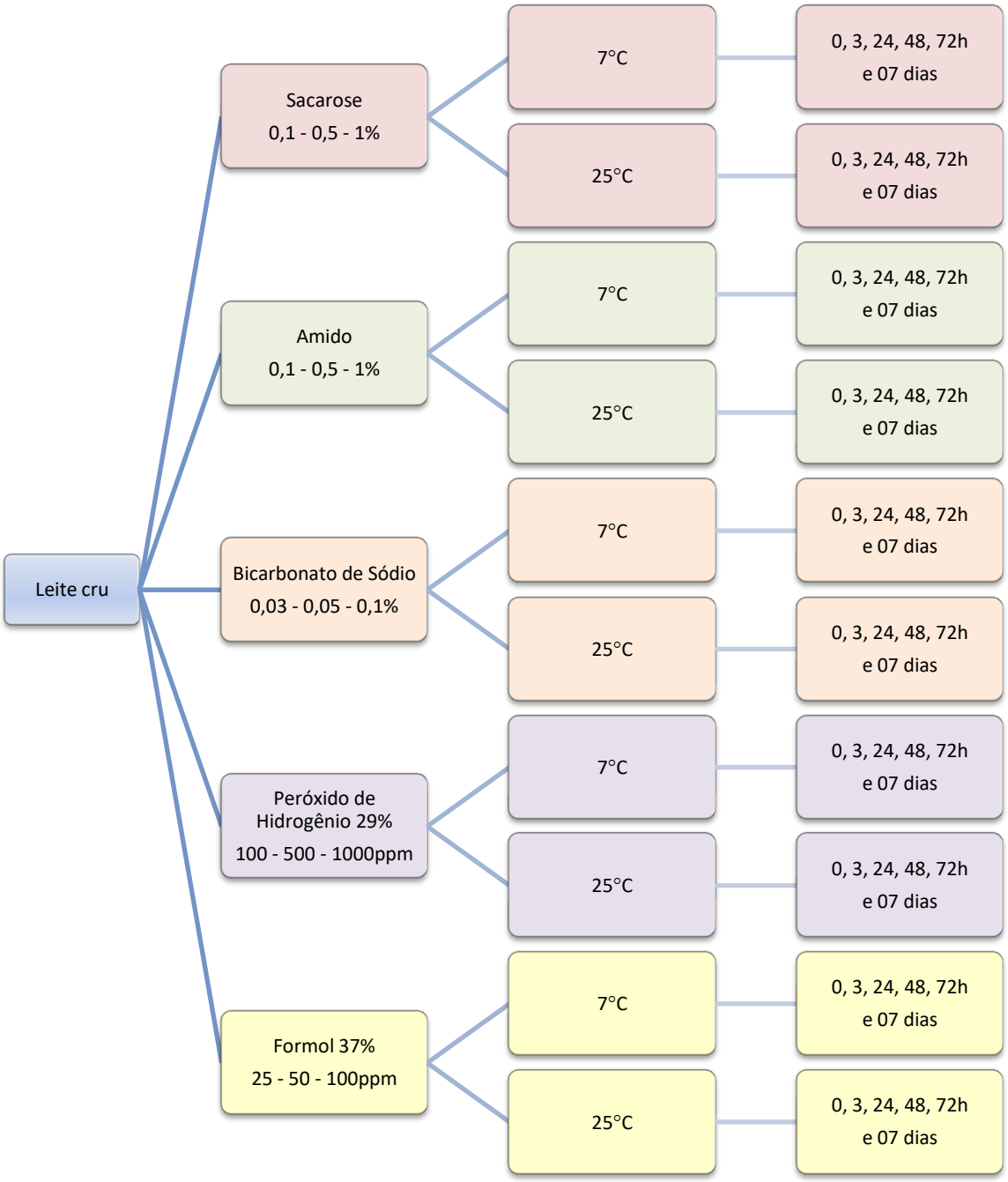


Figura 7: Esquema do preparo das amostras adulteradas.

Foram realizadas 8 repetições do esquema anterior, gerando 2.376 amostras adulteradas. Além dessas, foram utilizadas 7.412 amostras como controle, compostas por brancos (esquema anterior) e *checks* da própria rotina laboratorial, totalizando 9.788 amostras para fins de mineração computacional de dados.

Para a análise de composição, Contagem de Células Somáticas (CCS) e crioscopia, foi utilizado o equipamento de espectrofotometria no infravermelho por FTIR (*CombiScopeTM FTIR 400, Delta Instruments*, Drachten/Holanda), além de análises complementares de CBT (Contagem Bacteriana Total) no equipamento BactoScan™ FC/*Foss*, para estudos posteriores.

Para a realização das análises de composição e crioscopia no equipamento Delta FTIR, cada *rack* previamente preparada foi colocada no banho-maria à temperatura de 40 (±2) ºC, pelo período de 20 minutos, e homogeneizada por inversão 10 vezes com o auxílio de régua metálica.

A calibração do equipamento *CombiScope*™ FTIR (*Delta Instruments*) foi realizada por meio de 14 amostras padrão de leite cru, com diferentes faixas de composição para gordura, proteína, lactose e sólidos totais, obtidas do laboratório VALACTA® (*Dairy Production Centre of Expertise*, Quebec, Canadá), juntamente com os laudos contendo os resultados das análises realizadas no laboratório de origem.

4.3 Aquisição e mineração dos dados

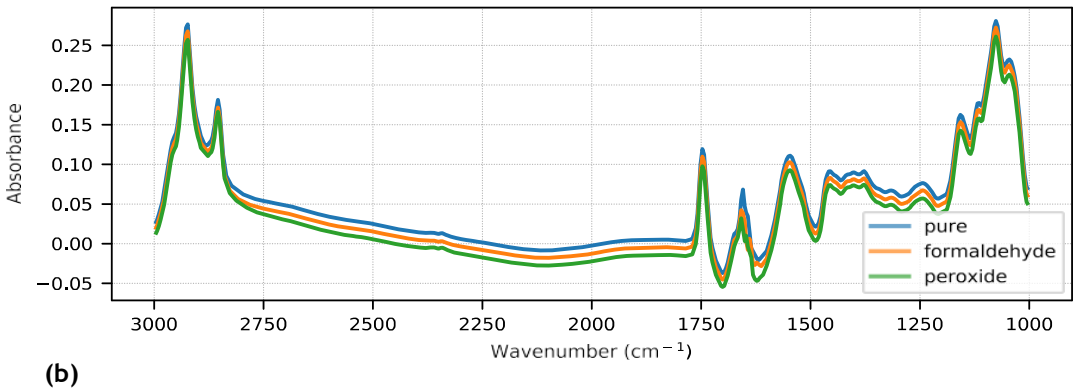
A caracterização do leite bovino foi realizada através de técnicas de aprendizado de máquina para detectar a presença de adulterantes no leite ou identificar o adulterante específico. Foram utilizados métodos de classificação por redes neurais e conjunto de árvores de decisão para determinar as categorias de amostras de leite, por meio de treinamento e teste em amostras reais. Os modelos classificadores foram treinados em amostras de leite adulteradas e controladas, a fim de reconhecer os padrões que identificam as características das adulterações.

Foram consideradas duas versões de classificação: binária e multiclasse. Na classificação binária, as respostas possíveis para uma classificação de amostra são a presença ou a ausência de um adulterante. Na multiclasse, as classes são um dos cinco

adulterantes específicos adicionados ao leite ou o leite normal (puro), quando a amostra não tem adulterante adicionado. Nesta metodologia, todas as concentrações e tempos de armazenamento utilizados foram agrupados em uma só classe por tipo de adulterante adicionado.

A espectroscopia FTIR foi aplicada a todas as amostras de leite coletadas para obter espectros de infravermelho. O equipamento FTIR produz duas informações para cada amostra analisada: um arquivo com dados espectrais (formato SPC), que contém coordenadas para o espectro infravermelho; e um arquivo de componentes (formato CSV), que contém variáveis numéricas, calculadas através de fórmulas pelo equipamento a partir do espectro. Na Figura 8, são demonstrados os espectros e os componentes numéricos extraídos de amostras aleatórias.

(a)



(b)

Cells	QValue	Fat	Protein	Lactose	Solids	FrzPoint	SNF	MUN	Casein	Ingredient
59	95.10	3.42	3.38	4.73	12.50	0.578	9.14	11.1	2.62	pure
58	90.74	3.42	3.39	4.77	12.55	0.582	9.20	13.0	2.63	formaldehyde
46	91.76	3.37	3.35	4.74	12.44	0.578	9.12	27.8	2.56	peroxide

Figura 8: (a) Espectros de infravermelho para amostras puras e adulteradas com formaldeído e peróxido, selecionadas aleatoriamente e analisadas diretamente por CNN. (b) Componentes numéricos para as mesmas amostras, gerados pelo equipamento FTIR em formato CSV, e analisados pelos algoritmos *Random Forests* e *Gradient Boosting Machine*.

No conjunto de dados, cada amostra é representada tanto pelos componentes numéricos, quanto pelos dados espectrais. No entanto, cada um dos dois tipos de dados foi trabalhado de forma diferente.

A estrutura de dados das variáveis numéricas é ideal para a aplicação de um

classificador de árvore de decisão, pois cada variável representa fortemente características específicas da composição do leite. Neste caso, foram aplicados os métodos de *Random Forest* (RF) e *Gradient Boosting Machine* (GBM).

Já os dados espectrais são compostos por coordenadas espectrais completas, que podem ser interpretadas como imagens para o reconhecimento das redes neurais. Neste caso, foram aplicadas as Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs), que são capazes de detectar características específicas do espectro, sem a necessidade de qualquer pré-processamento.

A título de comparação com estudos anteriores, também foram realizadas as análises de regressão linear, logística e PLS (*Partial Least Squares*).

5. RESULTADOS E DISCUSSÃO

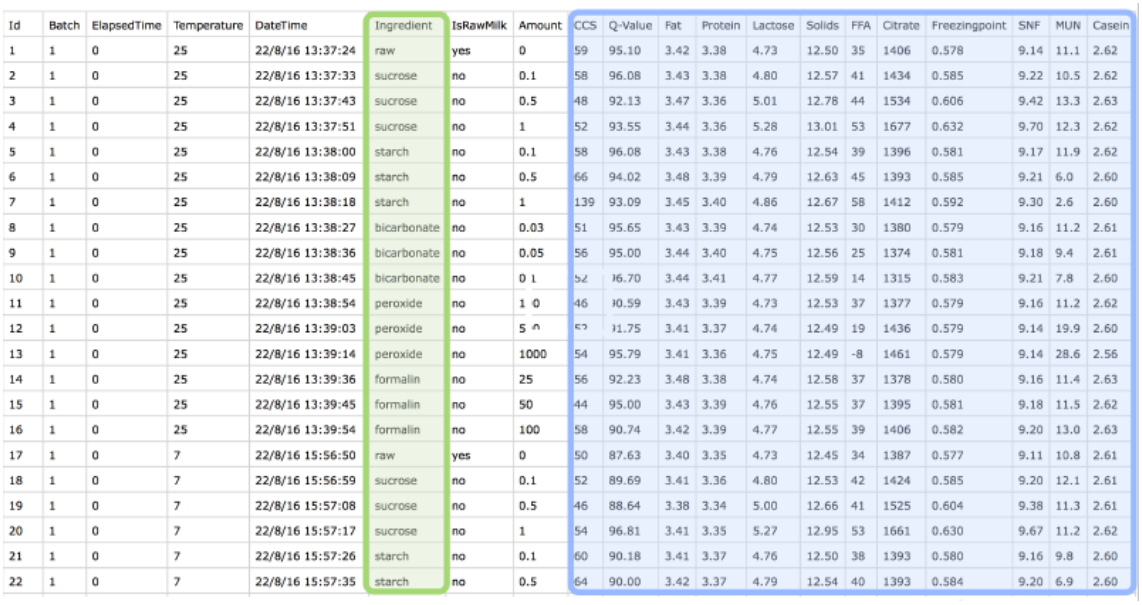
Os dados espectrais produzidos a partir da composição das amostras de leite foram utilizados para o treinamento de diferentes algoritmos de aprendizado de máquina, como Aprendizagem Profunda (*Deep Learners*) e Em Conjunto (*Ensemble Learners*). O método proposto é utilizado para predizer a presença de adulterantes em um problema de classificação binária e, também, a identificação de qual dos cinco adulterantes (amido, sacarose, bicarbonato de sódio, peróxido de hidrogênio e formaldeído) foi encontrado através da classificação multiclasse. Na Aprendizagem Profunda (*Deep Learning*), foi proposta uma arquitetura de Rede Neural Convolucional (*Convolutional Neural Network* - CNN) que dispensa o pré-processamento de dados espectrais. Na aprendizagem Em Conjunto (*Ensemble Learners*), foram implementados dois classificadores distintos: *Random Forest* (RF) e *Gradient Boosting Machine* (GBM), que apresentam melhorias em relação às técnicas de árvores de decisão independentes (HASTIE, 2017).

5.1 Análise de componentes numéricos e métodos preditores conjuntos

Durante o processo de leitura do espectro infravermelho, o equipamento FTIR executa uma série de cálculos gerados como valores numéricos para diferentes componentes do leite. De acordo com a documentação do equipamento, os cálculos são baseados em um modelo de Regressão Linear Múltipla (*Multiple Linear Regression* - MLR) que considera a absorbância da energia da luz pela amostra para regiões de comprimento de onda específicas. A informação extraída depende das calibrações do equipamento para a concentração dos componentes do leite (leitura do percentual de gordura, proteína, lactose, sólidos totais, sólidos não gordurosos (*Solids Non Fat* - SNF), caseína e nitrogênio ureico do leite (*Milk Urea Nitrogen* - MUN)). Outros três valores extras também estão incluídos na leitura do equipamento: a contagem de células somáticas, o valor do ponto de congelamento (índice crioscópico) e um valor de controle de qualidade da amostra (*Q-Value*), conforme ilustrado nas Figuras 9 e 10.

Cells	QValue	Fat	Protein	Lactose	Solids	FrzPoint	SNF	MUN	Casein	Ingredient
59	95.10	3.42	3.38	4.73	12.50	0.578	9.14	11.1	2.62	pure
58	90.74	3.42	3.39	4.77	12.55	0.582	9.20	13.0	2.63	formaldehyde
52	93.55	3.44	3.36	5.28	13.01	0.632	9.70	12.3	2.62	sucrose
58	96.08	3.43	3.38	4.76	12.54	0.581	9.17	11.9	2.62	starch
51	95.65	3.43	3.39	4.74	12.53	0.579	9.16	11.2	2.61	bicarbonate
46	91.76	3.37	3.35	4.74	12.44	0.578	9.12	27.8	2.56	peroxide

Figura 9: Componentes numéricos gerados pelo equipamento FTIR em formato CSV de amostras aleatórias.



Id	Batch	ElapsedTime	Temperature	DateTime	Ingredient	IsRawMilk	Amount	CCS	Q-Value	Fat	Protein	Lactose	Solids	FFA	Citrate	Freezingpoint	SNF	MUN	Casein
1	1	0	25	22/8/16 13:37:24	raw	yes	0	59	95.10	3.42	3.38	4.73	12.50	35	1406	0.578	9.14	11.1	2.62
2	1	0	25	22/8/16 13:37:33	sucrose	no	0.1	58	96.08	3.43	3.38	4.80	12.57	41	1434	0.585	9.22	10.5	2.62
3	1	0	25	22/8/16 13:37:43	sucrose	no	0.5	48	92.13	3.47	3.36	5.01	12.78	44	1534	0.606	9.42	13.3	2.63
4	1	0	25	22/8/16 13:37:51	sucrose	no	1	52	93.55	3.44	3.36	5.28	13.01	53	1677	0.632	9.70	12.3	2.62
5	1	0	25	22/8/16 13:38:00	starch	no	0.1	58	96.08	3.43	3.38	4.76	12.54	39	1396	0.581	9.17	11.9	2.62
6	1	0	25	22/8/16 13:38:09	starch	no	0.5	66	94.02	3.48	3.39	4.79	12.63	45	1393	0.585	9.21	6.0	2.60
7	1	0	25	22/8/16 13:38:18	starch	no	1	139	93.09	3.45	3.40	4.86	12.67	58	1412	0.592	9.30	2.6	2.60
8	1	0	25	22/8/16 13:38:27	bicarbonate	no	0.03	51	95.65	3.43	3.39	4.74	12.53	30	1380	0.579	9.16	11.2	2.61
9	1	0	25	22/8/16 13:38:36	bicarbonate	no	0.05	56	95.00	3.44	3.40	4.75	12.56	25	1374	0.581	9.18	9.4	2.61
10	1	0	25	22/8/16 13:38:45	bicarbonate	no	0.1	52	96.70	3.44	3.41	4.77	12.59	14	1315	0.583	9.21	7.8	2.60
11	1	0	25	22/8/16 13:38:54	peroxide	no	1.0	46	90.59	3.43	3.39	4.73	12.53	37	1377	0.579	9.16	11.2	2.62
12	1	0	25	22/8/16 13:39:03	peroxide	no	5.0	52	91.75	3.41	3.37	4.74	12.49	19	1436	0.579	9.14	19.9	2.60
13	1	0	25	22/8/16 13:39:14	peroxide	no	1000	54	95.79	3.41	3.36	4.75	12.49	-8	1461	0.579	9.14	28.6	2.56
14	1	0	25	22/8/16 13:39:36	formalin	no	25	56	92.23	3.48	3.38	4.74	12.58	37	1378	0.580	9.16	11.4	2.63
15	1	0	25	22/8/16 13:39:45	formalin	no	50	44	95.00	3.43	3.39	4.76	12.55	37	1395	0.581	9.18	11.5	2.62
16	1	0	25	22/8/16 13:39:54	formalin	no	100	58	90.74	3.42	3.39	4.77	12.55	39	1406	0.582	9.20	13.0	2.63
17	1	0	7	22/8/16 15:56:50	raw	yes	0	50	87.63	3.40	3.35	4.73	12.45	34	1387	0.577	9.11	10.8	2.61
18	1	0	7	22/8/16 15:56:59	sucrose	no	0.1	52	89.69	3.41	3.36	4.80	12.53	42	1424	0.585	9.20	12.1	2.61
19	1	0	7	22/8/16 15:57:08	sucrose	no	0.5	46	88.64	3.38	3.34	5.00	12.66	41	1525	0.604	9.38	11.3	2.61
20	1	0	7	22/8/16 15:57:17	sucrose	no	1	54	96.81	3.41	3.35	5.27	12.95	53	1661	0.630	9.67	11.2	2.62
21	1	0	7	22/8/16 15:57:26	starch	no	0.1	60	90.18	3.41	3.37	4.76	12.50	38	1393	0.580	9.16	9.8	2.60
22	1	0	7	22/8/16 15:57:35	starch	no	0.5	64	90.00	3.42	3.37	4.79	12.54	40	1393	0.584	9.20	6.9	2.60

Figura 10: Exemplo de aquisição de dados convertidos em arquivo formato XLSX com todos os componentes numéricos obtidos de amostras aleatórias.

A correlação completa das variáveis que representam os componentes numéricos é apresentada na Figura 11. As relações entre essas variáveis com correlações pareadas demonstraram que a proteína e a caseína são altamente correlacionadas (0,96) no conjunto de dados. Como a caseína é uma proteína específica do leite, essa correlação faz sentido. Correlações significativas também foram encontradas entre sólidos totais e gordura (0,85), lactose com ponto de congelamento (0,77), e lactose com SNF (0,81). As outras variáveis encontradas não foram correlacionadas significativamente.

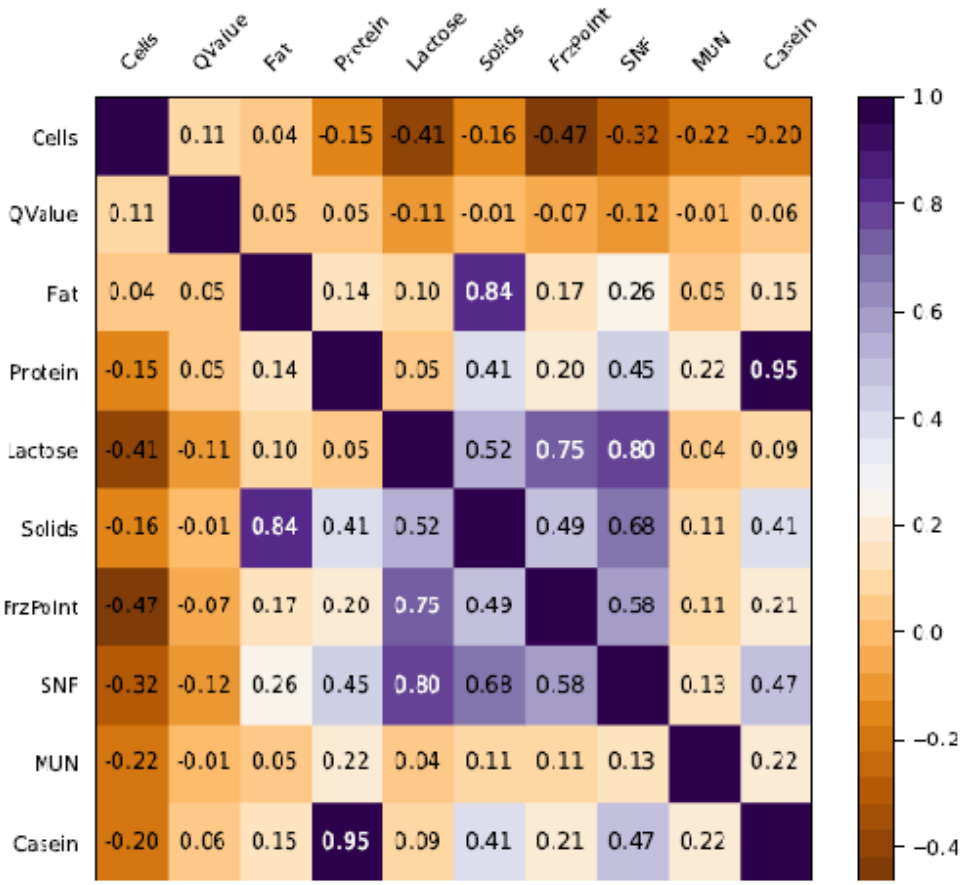


Figura 11: Matriz de correlação dos componentes numéricos do leite. Os valores indicam que a caseína e a proteína são altamente correlacionadas, uma vez que a caseína é a proteína no leite. Sólidos e gorduras também são correlacionados, já que a gordura é um dos componentes dos sólidos totais. A lactose está correlacionada com o ponto de congelamento e sólidos não gordurosos (SNF). Outras variáveis não são significativamente correlacionadas.

Todas as variáveis são lidas a partir de um arquivo CSV gerado pelo equipamento e são utilizadas em métodos preditores de Conjuntos de Árvores de Decisão. As amostras com adulteração conhecida foram consideradas como marcadores de classe e utilizadas para treinar os classificadores. A Figura 12 mostra um gráfico *boxplot* considerando a escala e a variação de todos os componentes numéricos.

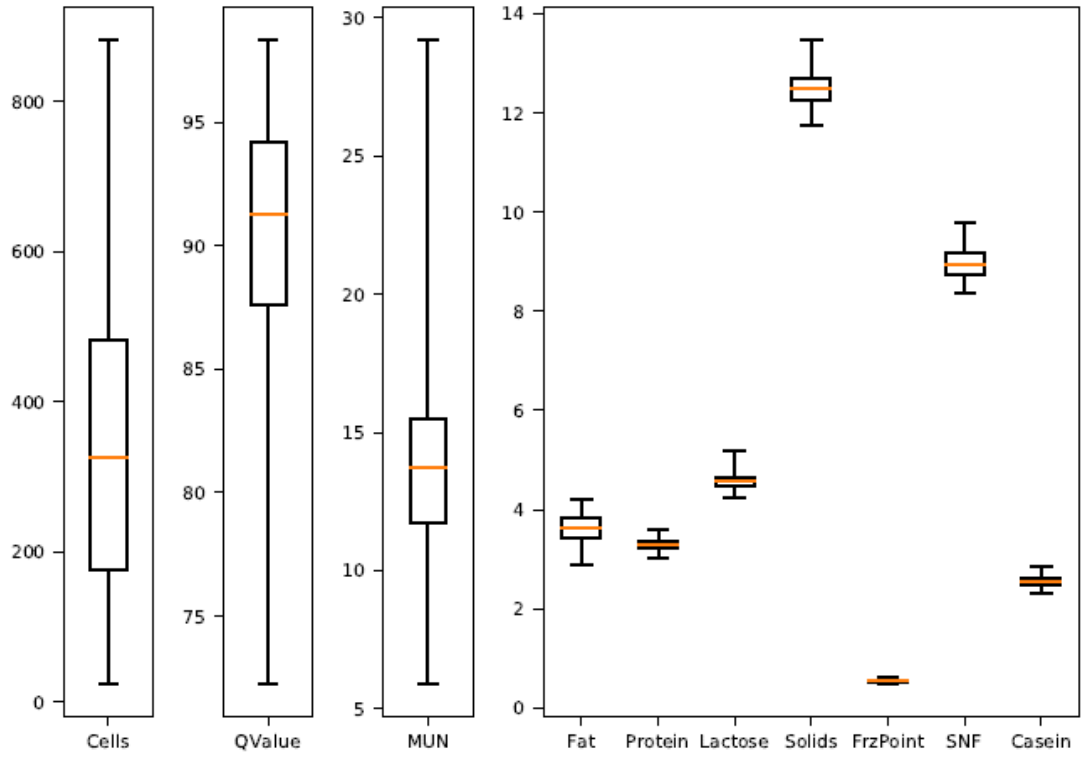


Figura 12: Boxplot dos componentes numéricos. CCS, Q-Value e MUN têm escalas significativamente diferentes e foram plotadas separadamente das outras variáveis.

As variáveis foram analisadas utilizando aprendizagem em conjuntos de árvores de decisão. Os métodos de aprendizagem em conjunto são compostos de múltiplas árvores de decisão com duas técnicas populares: *Bagging* e *Boosting*. O *Bagging* (*Bootstrap Aggregating*) treina vários modelos de árvore de decisão de forma independente, com subconjuntos de dados escolhidos aleatoriamente, e usa a votação por maioria para agregar as saídas dos preditores de base. O *Boosting* também treina classificadores usando diferentes conjuntos de treinamento, mas eles são aprendidos sequencialmente, com cada árvore tentando minimizar o erro da árvore anterior. A figura 13 ilustra frações de execução de uma das árvores de decisão utilizadas no modelo.

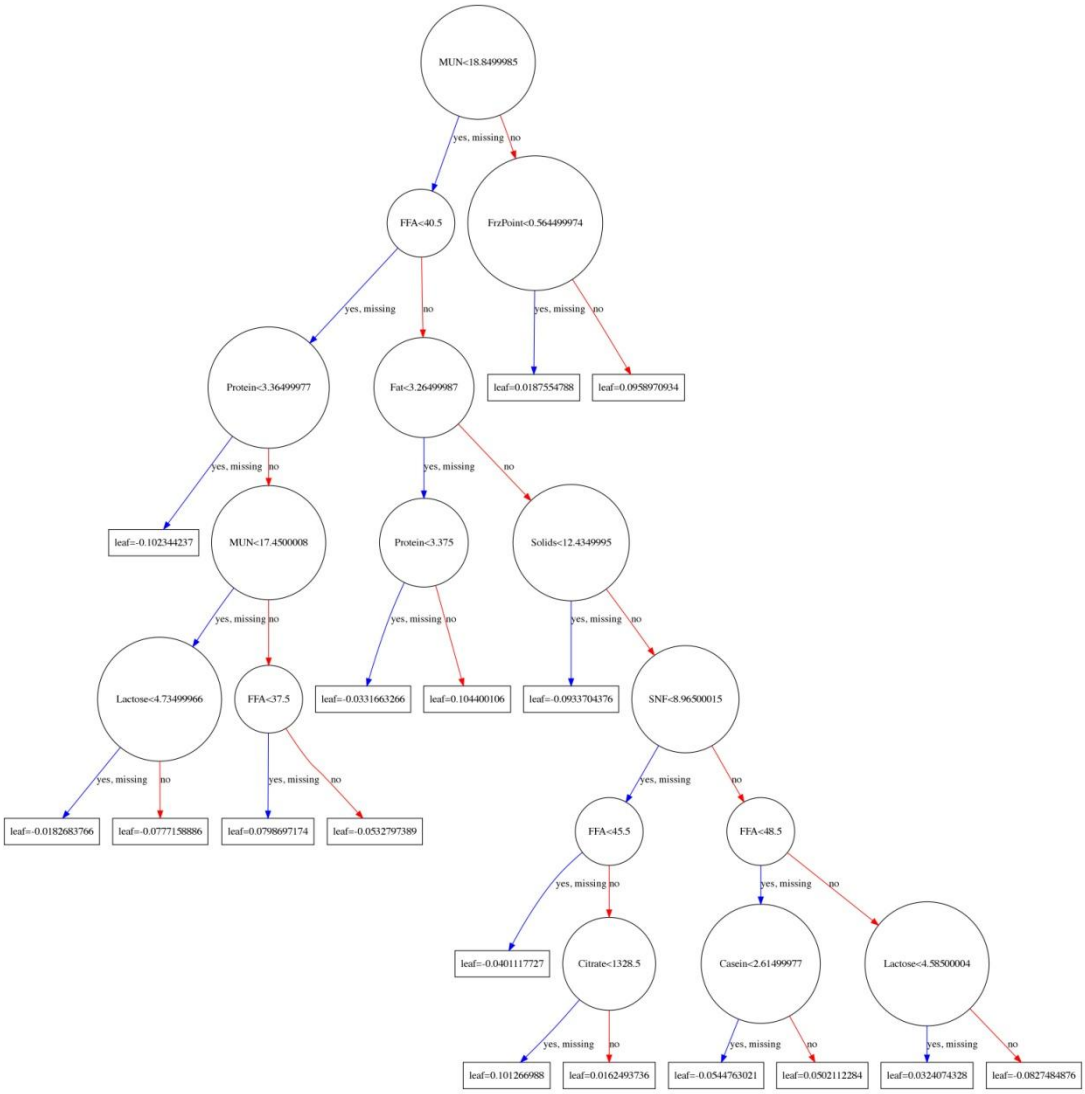


Figura 13: Frações da árvore de decisão gerada pelo algoritmo de classificação: execução da árvore de decisão (Gradient Boosted Tree - GB).

A combinação de preditores individualmente fracos cria um modelo de melhor desempenho (HASTIE *et al*., 2017). *Random Forest* (RF) e *Gradient Boosting Machine* (GBM) são exemplos de técnicas de *Bagging* e *Boosting*, respectivamente. As implementações padrão dos classificadores RF e GBM utilizadas estão disponíveis na biblioteca *Scikit-learn* do *Python* (PEDREGOSA *et al*., 2011), que fornece implementações de códigos de diversos métodos algorítmicos de aprendizagem de máquina. O número de preditores é controlado pelo parâmetro *n_estimators* e foi definido como 200 para cada classificador. Os modelos de ambos os métodos foram avaliados para cada treinamento disponível e os conjuntos de testes, usando características dos componentes presentes nas amostras. As classificações binária e multiclasse foram realizadas considerando os mesmos conjuntos de dados. As figuras 14 e 15 ilustram as etapas iniciais das classificações binária e multiclasse empregadas, respectivamente.

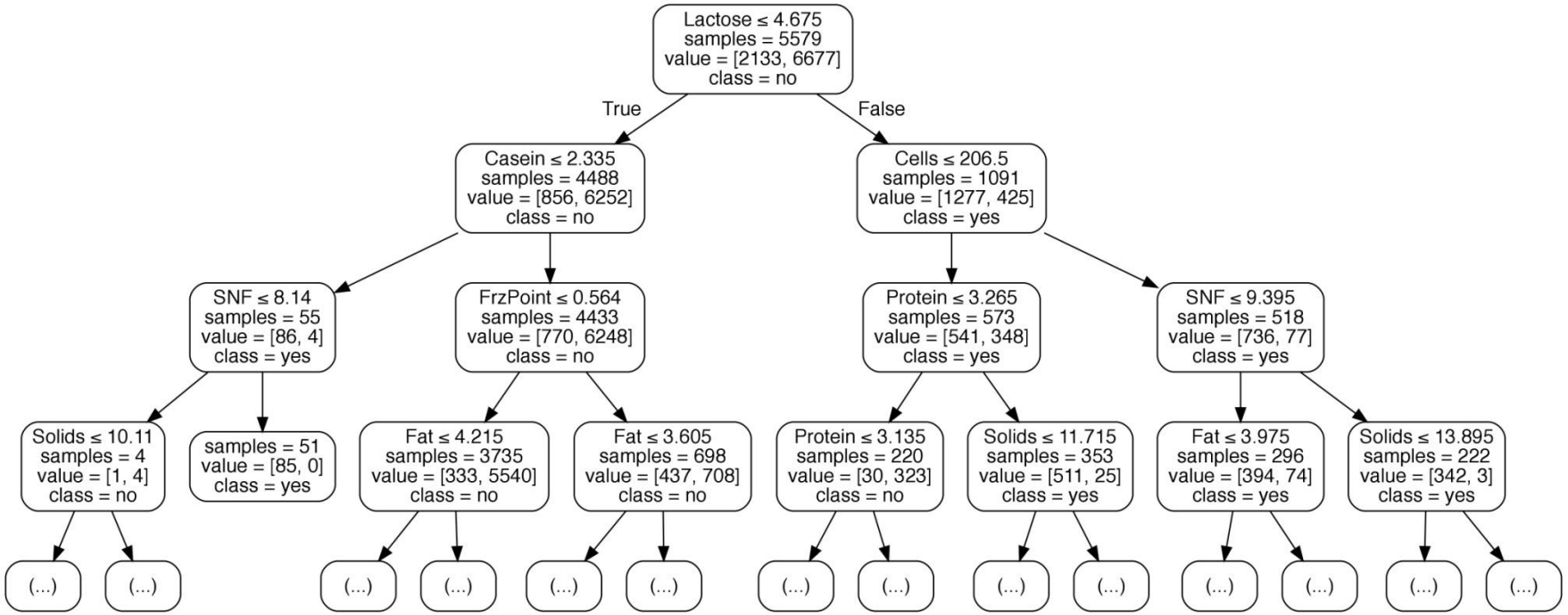


Figura 14: Etapas iniciais das tomadas de decisão para a classificação binária.

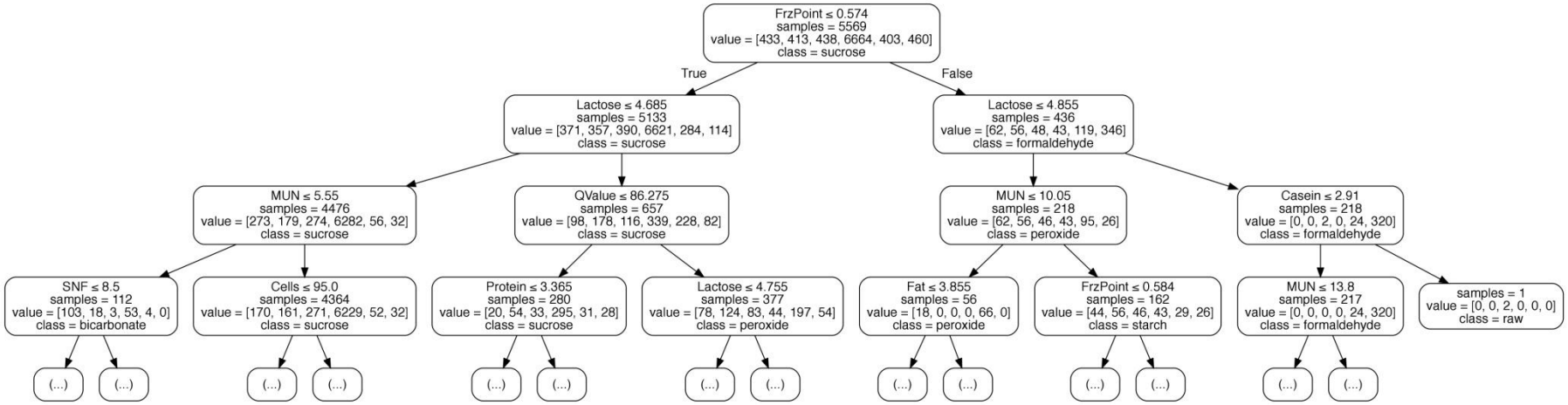


Figura 15: Etapas iniciais das tomadas de decisão para a classificação multiclasse.

5.2 Análise dos espectros de infravermelho por aprendizagem profunda

Os espectros de infravermelho produzidos pela técnica de FTIR são formados por 518 pontos ou varreduras, com resolução de 4 cm^{-1}, medidos em números de onda variando de 3000 cm^{-1} a 1000 cm^{-1}, que são usados como entrada para o classificador da CNN. Cada espectro é seguido pelo rótulo de classe (substância adulterante) no conjunto de dados, o que permite o processo de treinamento da rede. Nas figuras 16 a 20, são demonstrados os espectros das amostras adulteradas com amido, sacarose, bicarbonato de sódio, peróxido de hidrogênio e formaldeído, concomitantemente ao espectro do leite normal (puro), revelando as alterações provocadas no mesmo.

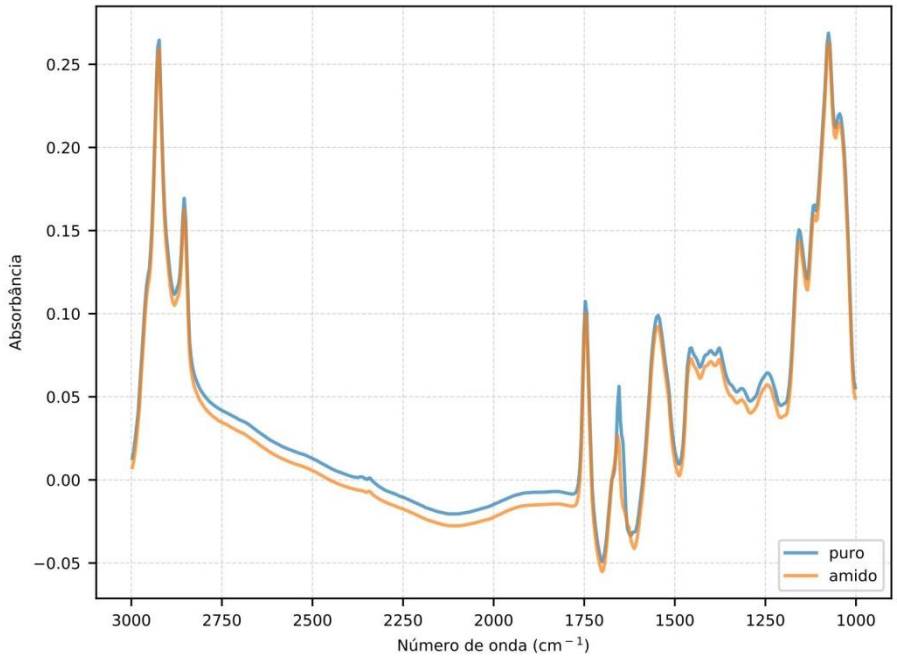


Figura 16: Alterações provocadas no espectro de amostra adulterada com amido, em relação ao espectro do leite puro.

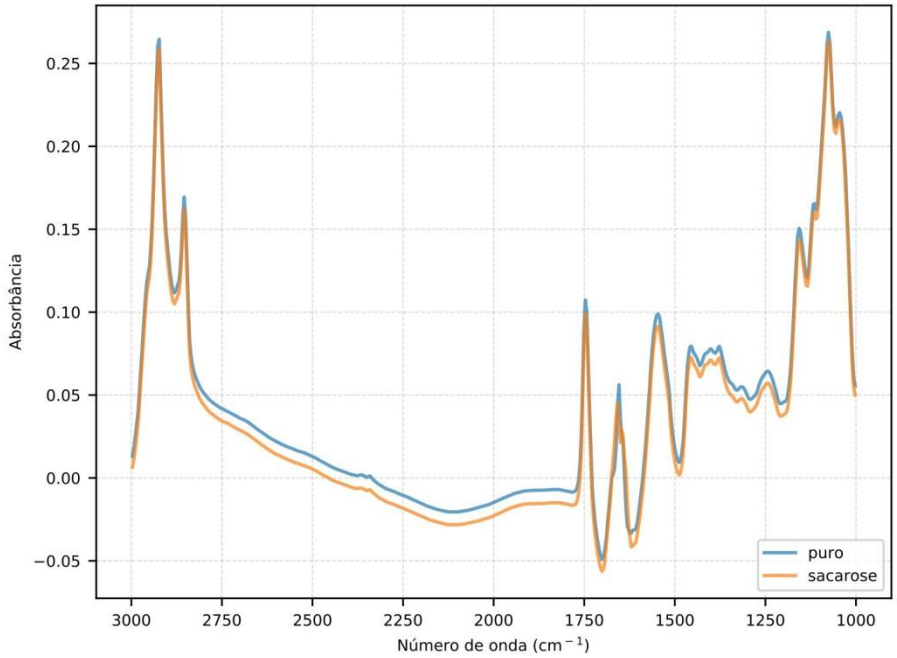


Figura 17: Alterações provocadas no espectro de amostra adulterada com sacarose, em relação ao espectro do leite puro.

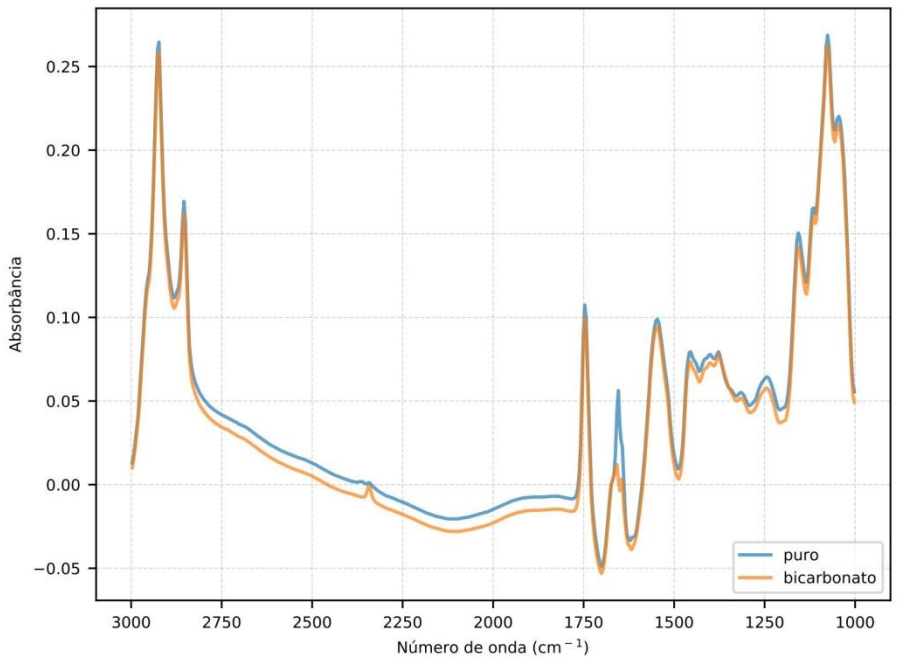


Figura 18: Alterações provocadas no espectro de amostra adulterada com bicarbonato de sódio, em relação ao espectro do leite puro.

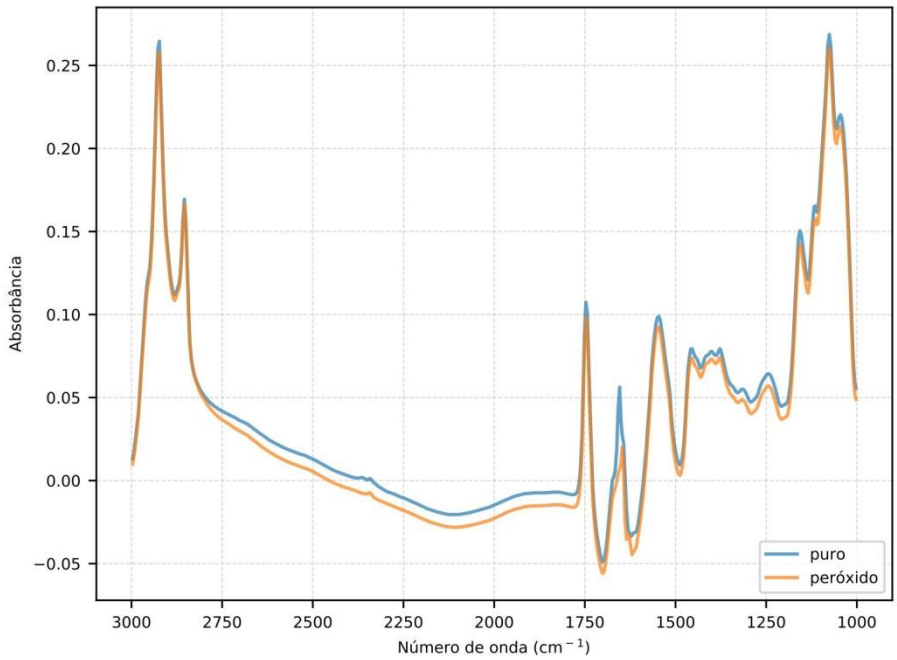


Figura 19: Alterações provocadas no espectro de amostra adulterada com peróxido de hidrogênio, em relação ao espectro do leite puro.

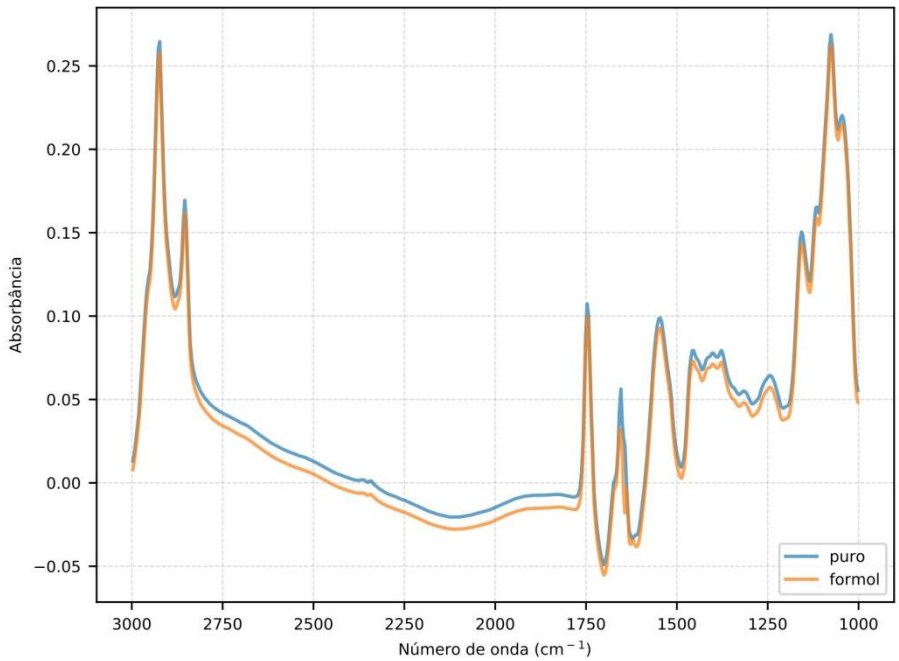


Figura 20: Alterações provocadas no espectro de amostra adulterada com formaldeído, em relação ao espectro do leite puro.

Quando comparado a redes neurais regulares, camadas adicionais (camadas convolucionais) são usadas em CNNs. No processo de treinamento, essas camadas são usadas como filtros que reconhecem regiões espectrais com características específicas. Por esse motivo, as CNNs podem receber dados espectrais brutos como entrada, sem a necessidade de qualquer etapa de pré-processamento, e podem manipular a extração de recursos importantes dos dados, sem interação manual (SCHMIDHUBER, 2015).

Neste trabalho, foi proposta uma arquitetura CNN que possui uma camada convolucional unidimensional, que aprende 32 filtros de tamanho de *kernel* 5, capazes de extrair características diretamente dos espectros de infravermelho. Os filtros são concatenados e seguidos por uma camada densa (totalmente conectada) de 1024 neurônios. A estrutura de rede proposta foi baseada no trabalho de Liu *et al*. (2017), porém muito mais simples, com menos camadas e filtros. Na Figura 21, é demonstrada a arquitetura proposta da CNN.

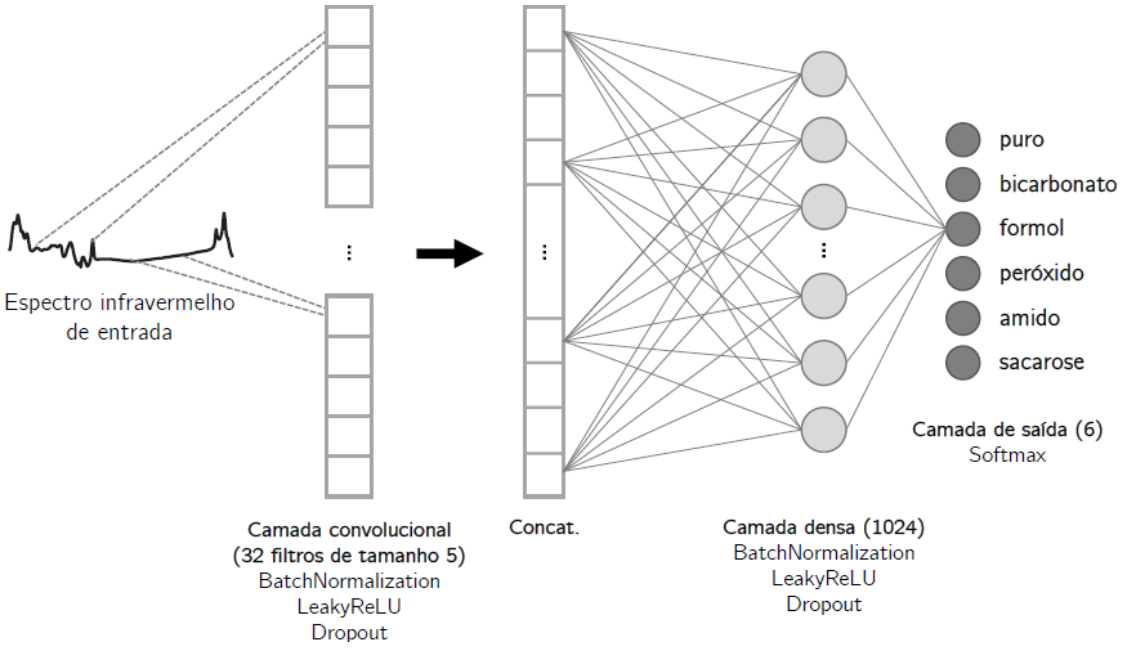


Figura 21: A arquitetura CNN proposta para classificação multiclasse consiste em uma camada convolucional, uma camada totalmente conectada e a camada de saída. Informações adicionais são descritas em cada camada.

Para as classificações binária e multiclasse, foi treinada uma CNN que difere apenas no número de neurônios na camada de saída. Como essa camada produz a classificação, o número de neurônios deve ser exatamente o número de classes que se queira classificar

os dados. Assim, a CNN para a classificação binária apresentou uma camada de saída de um neurônio com saída binária, ativada pela função sigmoide. Já a CNN para a classificação multiclasse teve uma camada de saída com seis neurônios, ativados pela função *softmax* (ZEILER & FERGUS, 2014). O modelo binário classifica as amostras com a presença ou a ausência de um adulterante, e a classificação multiclasse classifica as amostras como leite puro ou com uma das cinco substâncias adulterantes conhecidas.

O treinamento da CNN foi feito utilizando *Adam (Adaptive Moment Estimation) Optimizer* (KINGMA & BA, 2015) com 100 iterações (épocas ou fórmulas) para as classificações binária e multiclasse. Cada execução da CNN considerou 20% do conjunto de treinamento como o conjunto de validação interno à rede. A Figura 22 mostra os gráficos da precisão e da perda do modelo em conjuntos de treinamento e validação, revelando o aprendizado da rede neural. À medida que se aumentam as épocas, o aprendizado da rede evolui (acurácia) e diminuem-se os erros (perdas).

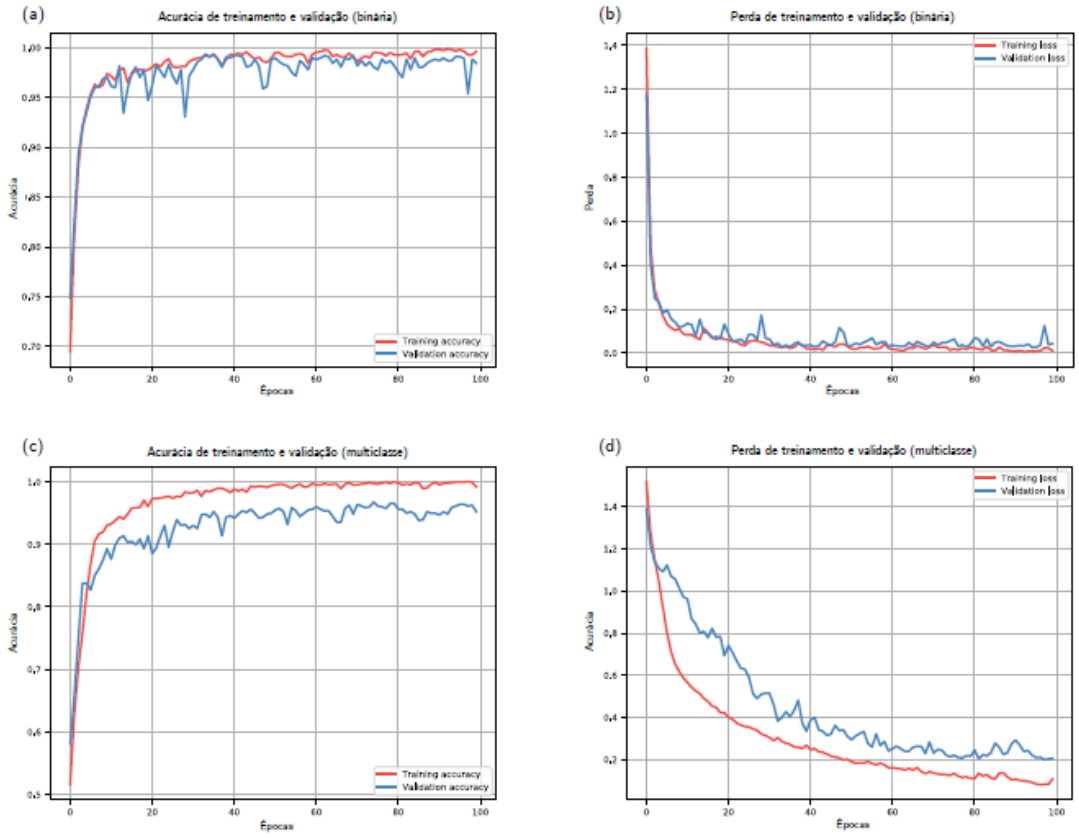


Figura 22: Figura 22 Gráfico da precisão e da perda do modelo da CNN nas etapas de treinamento e validação, considerando a divisão do conjunto de dados de 80% / 20%. O modelo foi treinado por 100 épocas. (a) Precisão do treinamento e validação considerando o problema binário. (b) Perda de treinamento e validação considerando o problema multiclasse.

A arquitetura da CNN foi implementada utilizando as bibliotecas *Keras* (CHOLLET *et al*., 2015) e *TensorFlow* (ABADI *et al*., 2016) em *Python*.

5.3 Acurácias das classificações

Para a avaliação da qualidade dos classificadores, a técnica de validação cruzada de *hold-out* (KOHAVI, 1995; ARLOT & CELISSE, 2010) foi realizada com três pares de subconjuntos de treinamento/teste nas proporções: 80/20%, 70/30% e 60/40% (Tabela 02). Essa estratégia de divisão apresentou os melhores resultados após o teste com classificadores diferentes, incluindo aprendizagem profunda, que geralmente divide os conjuntos de dados em conjuntos de treinamento, validação e teste.

Tabela 2: Precisão dos classificadores avaliados para classificações binárias e multiclasse. Todos os classificadores foram avaliados com 3 pares de conjuntos de dados de treinamento e teste selecionados aleatoriamente das amostras de leite, identificados por sua proporção de treinamento e amostras de teste.

Conjunto de dados	Classificação	RF	GBM	CNN
80%/20%	Binária	0.9872	0.9826	0.9908
	Multiclasse	0.9367	0.9346	0.9816
70%/30%	Binária	0.9854	0.9782	0.9758
	Multiclasse	0.9431	0.9346	0.9833
60%/40%	Binária	0.9867	0.9821	0.9934
	Multiclasse	0.9415	0.9318	0.9849

Para os métodos em conjunto, as precisões de classificação variaram de 93,18% (método GBM na classificação multiclasse) a 98,72% (método RF na classificação binária), sendo considerados resultados excelentes. Esses métodos trabalham apenas com os dados numéricos em formato CSV gerados pelo próprio equipamento através de fórmulas (Figura 9), não utilizando diretamente os espectros para as classificações. Não foram encontrados na literatura trabalhos anteriores que utilizassem esta metodologia, embora os resultados sejam surpreendentes e de rápida execução (aproximadamente 16 minutos para treinamento e classificação em computador pessoal).

No entanto, ao trabalhar diretamente com os dados espectrais através da CNN proposta,

foram gerados valores de até 99,34% de acurácia para a classificação binária, e de até 98,49% de acurácia para a classificação multiclasse, ou seja, valores ainda superiores aos obtidos pelos métodos em conjunto. De acordo com a arquitetura CNN proposta, pode-se predizer, com >99% de segurança que a amostra analisada está ou não adulterada (método de triagem) e, ainda mais, em caso positivo, identificar qual o adulterante adicionado no modelo treinado (>98%).

As precisões médias para RF, GBM e CNN foram 96,34%, 95,73% e 98,49%, respectivamente. De forma geral, a classificação binária tende a ser um problema mais simples e pode levar a melhores resultados, o que é observado nos resultados de RF e GBM, onde as precisões de classificações binárias são até 10% mais altas do que as precisões de multiclasse. No entanto, os resultados da CNN mostram que o método é mais robusto nas classificações multiclasse, sendo particularmente mais adequadas para esta classificação.

Em nenhum dos estudos anteriores pesquisados, que avaliaram a identificação de adulterantes no leite cru por espectroscopia FTIR, foi utilizada metodologia de *data mining* similar, ou obtidos resultados superiores aos deste trabalho. Além disso, a realização das análises em diversos dias é de grande importância para mimetizar a realidade da rotina laboratorial, onde as amostras serão analisadas no prazo médio de 72 horas.

O estudo de Conceição e colaboradores (2019) teve como objetivo utilizar os dados espectrais de FTIR – ATR (*Attenuated Total Reflection*) combinados à análise multivariada para identificar adulterações no leite cru. Um total de 620 amostras foram adulteradas com bicarbonato de sódio, hidróxido de sódio, peróxido de hidrogênio, amido, sacarose e ureia. Foram empregados os métodos de análise de regressão linear múltipla e rede neural artificial, que apresentaram média de correlação de apenas 76% para detecção dos adulterantes.

No trabalho realizado por Gondim e colaboradores (2017), foi aplicada a técnica de classificação *Soft Independent Modelling of Class Analogy* – SIMCA em dados espectrais de infravermelho médio para detectar adulterantes em leite. Foram adulteradas 360 amostras com formaldeído, peróxido de hidrogênio, citrato de sódio, carbonato de sódio e amido, as quais obtiveram 82% de classificações corretas, 17%

inconclusivas e 1% de erros por essa metodologia.

No estudo de Coitinho *et al*. (2017), foram adulteradas 1650 amostras com amido, bicarbonato de sódio, citrato de sódio, formaldeído, sacarose, soro e água para a calibração de um equipamento compacto de espectroscopia FTIR, utilizando a análise de componentes principais (*Principal Component Analysis* – PCA). Para a avaliação das 12 calibrações obtidas, foram determinadas a sensibilidade e a especificidade de cada uma delas. Os melhores resultados demonstraram 84% de sensibilidade para todos os adulterantes simultaneamente.

Já o trabalho de Botelho e colaboradores (2015) propôs um método de triagem para a detecção simultânea dos adulterantes amido, citrato de sódio, formaldeído, sacarose e água em apenas 171 amostras, usando espectroscopia de infravermelho médio com Refletância Total Atenuada (*Attenuated Total Reflectance* - ATR) e classificação supervisionada multivariada, baseada em Análise de Discriminação Parcial de Mínimos Quadrados (*Partial Least Squares Discrimination Analysis* – PLSDA). Foram estimadas as taxas de falso positivo, falso negativo, seletividade, especificidade, eficiência, acordância e concordância. A média das taxas de sensibilidade obtidas no conjunto de teste dos cinco adulterantes foi de 92,94%.

Em Santos *et al*. (2013), foi avaliada a aplicação de microespectroscopia no infravermelho médio (ATR MIR - *Microspectroscopy*) em 370 amostras para detecção e quantificação dos adulterantes: soro lácteo, peróxido de hidrogênio, urina sintética, ureia e leite sintético (vegetal). Foram realizadas análises de reconhecimento de padrões por *Soft Independent Modeling de Class Analogy* (SIMCA) e Regressão Parcial dos Mínimos Quadrados (*Partial Least Squares Regression* - PLSR), com resultados de coeficientes de correção próximos a 95%.

Comparativamente, os estudos anteriores utilizaram uma amostragem muito inferior, foram baseados em análises estatísticas multivariadas mais tradicionais e mostraram taxas de classificação menores do que as do método proposto.

Cada subconjunto foi obtido aleatoriamente a partir do conjunto de dados original, no total de 9.788 amostras, sendo 2.376 adulteradas e 7.412 amostras puras. A distribuição dos conjuntos de dados das amostras é descrita na Tabela 3. As distribuições de classes detalhadas para cada divisão de conjunto de dados de treinamento e teste são

apresentadas na Tabela 4.

Tabela 3: Distribuição das amostras binárias e multiclasses do conjunto de dados coletados, no total de 9.788 amostras.

Multiclasse	Nº de amostras	Binária	Nº de amostras2
Puro	7412	Puro	7412
Sacarose	486	Adulterado	2376
Formol	485		
Amido	480		
Peroxide	465		
Bicarbonate	460		

Tabela 4: Distribuição de classes para amostras em cada conjunto de treinamento e teste em versão multiclasse. Na versão binária, as cinco classes de substâncias adulterantes são resumidas como uma classe única.

Conjunto de dados	Classes	Nº de amostras de treinamento	Nº de amostras de teste
80%/20%	Puro	5906	1506
	Formol	395	109
	Amido	391	99
	Sacarose	387	85
	Peróxido	376	85
	Bicarbonato	375	74
70%/30%	Puro	5224	2188
	Formol	350	169
	Amido	327	169
	Sacarose	323	159
	Peróxido	316	142
	Bicarbonato	311	110
60%/40%	Puro	4451	2961
	Formol	315	200
	Amido	293	199
	Sacarose	285	198
	Peróxido	267	187
	Bicarbonato	261	171

Com o objetivo de comparar estes resultados com os métodos tradicionais mais comumente utilizados nos trabalhos anteriores, foram realizadas também as análises de

regressão linear, regressão logística e PLS. As versões de classificação desses métodos foram avaliadas para cada conjunto de dados de treinamento e teste, com os espectros completos. Como resultado, as acurácias da regressão linear variaram de 69,49% a 71,71% no modelo multiclasse, e de 85,73% a 86,63% na classificação binária. Para a regressão logística, as precisões variaram de 74,09% a 76,81% na classificação multiclasse, e 79,21% a 83,04% na binária. Por fim, para a análise PLS, as precisões variaram de 72,45% a 74,97% na multiclasse, e 83,96% a 85,04% na binária. Embora essas análises tenham apresentado resultados razoáveis de acurácia, corroborando com os dados encontrados anteriormente na literatura, ainda assim são muito inferiores aos valores obtidos pelos classificadores de RF, GBM e CNN neste estudo. A Tabela 5 mostra todos os valores de precisão dos modelos de regressão linear, logística e PLS para as classificações binária e multiclasse.

Tabela 5: Tabela 5 Acurácias de regressão linear, regressão logística e PLS para as classificações multiclasse e binária, nos três pares de conjuntos de dados de treinamento e teste.

Conjunto de dados	Classificação	Regressão Linear	Regressão Logística	PLS
80%/20%	Binária	0,8662	0,8304	0,8504
	Multiclasse	0,7068	0,7681	0,7497
70%/30%	Binária	0,8577	0,7921	0,8421
	Multiclasse	0,6949	0,7409	0,7245
60%/40%	Binária	0,8573	0,8207	0,8396
	Multiclasse	0,7171	0,7559	0,7377

A estratégia proposta mostra-se bastante eficiente para uma abordagem de triagem, a fim de determinar se uma amostra está adulterada ou não. Caso o modelo esteja treinado, a exemplo dos cinco adulterantes utilizados neste estudo (amido, sacarose, bicarbonato de sódio, peróxido de hidrogênio e formaldeído), permite-se também a identificação da substância fraudulenta adicionada, com acurácia próxima a 100%.

No trabalho de Conceição *et al*. (2019), a rede neural artificial proposta foi utilizada para identificar possíveis adulterações em 249 amostras desconhecidas de produtores, nas quais foram acusadas 3 amostras adulteradas com hidróxido de sódio, 2 amostras com amido e 1 amostra com ureia, correspondendo a 2,4% do total.

No nosso estudo, apenas no mês de janeiro de 2017, foram coletadas 12.927 amostras desconhecidas da rotina laboratorial. Após a classificação pela rede neural convolucional proposta, foram identificadas 12.491 amostras puras (96,63%), 310 amostras adulteradas com formol (2,40%), 44 amostras com bicarbonato de sódio (0,34%), 43 amostras com peróxido de hidrogênio (0,33%), 35 amostras com amido (0,27%) e 4 amostras com sacarose (0,03%), totalizando 3,37% de amostras com adulterações.

No mês de novembro de 2017, foram analisadas mais 23.428 amostras, das quais 23.073 foram classificadas como puras (98,48%), 206 amostras adulteradas com formol (0,88%), 74 amostras com bicarbonato de sódio (0,32%), 61 amostras com peróxido de hidrogênio (0,26%), 8 amostras com sacarose (0,03%) e 6 amostras com amido (0,03%), totalizando 1,52% de amostras adulteradas.

Após a identificação dessas amostras pelo equipamento como análise de triagem, será necessário segregá-las e encaminhá-las para as análises confirmatórias oficiais estabelecidas pela legislação (IN Nº30/2018 – MAPA). A implementação desta metodologia na rotina laboratorial é capaz de reduzir significativamente o número de amostras submetidas a estas análises, as quais são baseadas em técnicas de bancada, além de permitir um controle mais eficiente. Atualmente, os métodos de detecção de adulterantes no leite cru previstos pela legislação brasileira apresentam limitações em relação à sua sensibilidade analítica, são análises demoradas, apresentam alto custo, requerem mão de obra qualificada, utilizam grandes volumes de reagentes químicos e geram resíduos tóxicos poluentes. Dessa forma, reduzir-se-ia tempo, custos, resíduos e erros na detecção dessas substâncias (BRASIL, 2018).

Neste contexto, a adoção de técnicas laboratoriais instrumentais mais rápidas e eficientes torna-se necessária ao melhor desempenho técnico e econômico da gestão laboratorial. Além disso, amplia-se o leque de possibilidades para a identificação de qualquer substância estranha adicionada ao leite, com vistas à expansão e ao compartilhamento da metodologia para a Rede Brasileira de Laboratórios de Controle da Qualidade do Leite (RBQL), contribuindo sobremaneira para a fiscalização agropecuária nacional, visando à garantia de autenticidade, qualidade e de saúde pública.

6. CONCLUSÕES

A aplicação das técnicas de aprendizagem de máquina junto à espectroscopia FTIR, por meio da extração e interpretação de informações de conjuntos de dados complexos, apresenta uma grande vantagem para a detecção de adulteração nos produtos lácteos.

Neste trabalho, foram utilizados métodos de classificação para reconhecer padrões nos dados espectrais do leite analisado pelo equipamento FTIR, sendo possível detectar e identificar as adulterações nas amostras analisadas, classificando-as através de métodos de aprendizagem profunda e conjunto de árvores, com acurácias próximas a 100%.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. ABADI, M. *et al.* **TensorFlow: A System for Large-scale Machine Learning**. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. OSDI'16, p. 265–283. USENIX Association, Savannah, GA, USA, 2016.

2. ABNT - Associação Brasileira de Normas Técnicas. ISO - International Standard Organization. **ABNT NBR ISO/IEC 17025**. Requisitos gerais para a competência de laboratório de ensaio e calibração. Rio de Janeiro: ABNT; 2017.

3. ALVES DA ROCHA, R.; PAIVA, I. M.; ANJOS, V.; FURTADO, M. A. M.; BELL, M. J. V. Quantification of whey in fluid milk using confocal raman microscopy and artificial neural network. **Journal of Dairy Science**, v.98, n.6, p.3559–3567, 2015.

4. ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v.4, p.40–79, 2010.

5. AZAD, T.; AHMED, S. Common milk adulteration and their detection techniques. **International Journal of Food Contamination**, v.2, n.22, p.1-9, 2016.

6. BARBOSA, L. C. A. **Espectroscopia no infravermelho na caracterização de compostos orgânicos**. Editora UFV, 1a ed, 189p., Viçosa, 2013.

7. BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. **Journal of Microbiological Methods**, v.43, p.3-31, 2000.

8. BERSOT, L. S.; DAGUER, H.; MAZIERO, M. T.; PINTO, J. P. A. N.; BARCELLOS, V. C.; GALVÃO, J. A. Raw milk trade: profile of the consumers and microbiological and physicochemical characterization of the product in Pelotina – PR region. **Revista Instituto de Laticínios "Cândido Tostes"**, v.65, n.373, p.3-8, 2010.

9. BIGGS, D. A.; JOHNSSON, G.; SJAUNJA, L. O. Analysis of fat, protein, lactose and total solids by infra-red absorption. In: Monograph on rapid indirect methods for measurement of the major components of milk. **Bulletin of the International Dairy Federation**, n.208, 1987.

10. BORIN, A.; FERRÃO, M. F.; MELLO, C.; MARETTO, D. A.; POPPI, R. J. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. **Analytica Chimica Acta**, v.579, n.1, p.25-32, 2006.

11. BOTELHO, B. G.; REIS, N.; OLIVEIRA, L. S.; SENA, M. M. Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. **Food Chemistry**, v.181, p.31-37, 2015.

12. BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Instrução Normativa N° 30, de 26 de junho de 2018. Ficam estabelecidos como oficiais os métodos constantes do Manual de Métodos Oficiais para Análise de Alimentos de Origem Animal. **Diário Oficial da República Federativa do Brasil,** Brasília, seção 1, n. 134, p.9, 13/07/2018.

13. BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Instrução Normativa N° 76, de 26 de novembro de 2018. Ficam aprovados os Regulamentos Técnicos que fixam a identidade e as características de qualidade que devem apresentar o leite cru refrigerado, o leite pasteurizado e o leite pasteurizado tipo A. **Diário Oficial da República Federativa do Brasil,** Brasília, seção 1, n. 230, p.9, 30/11/2018.

14. BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Instrução Normativa N° 77, de 26 de novembro de 2018. Ficam estabelecidos os critérios e procedimentos para a produção, acondicionamento, conservação, transporte, seleção e recepção do leite cru em estabelecimentos registrados no serviço de inspeção oficial. **Diário Oficial da República Federativa do Brasil,** Brasília, seção 1, n. 230, p.10, 30/11/2018.

15. BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA).

Decreto N° 9.013 de 29 de março de 2017. Regulamento da Inspeção Industrial e Sanitária de Produtos de Origem Animal - R.I.I.S.P.O.A. **Diário Oficial da República Federativa do Brasil,** Brasília, 29/03/2017. Alterado pelo Decreto 9.069 de 31 de maio de 2017.

16. CAPUANO, E.; EENLING, R. B.; KOOT, A.; VAN RUTH, S. M. Target and untarget detection of skim milk powder adulteration by near-infrared spectroscopy. **Food Analytical Methods**, v.8, p.2125-2134, 2015.

17. CARVALHO, B. M. A.; CARVALHO, L. M..; COIMBRA, J. S. R; MININ, L. A.; BARCELLOS, E. S.; JUNIOR, W. F. S.; DETMANN, E.; CARVALHO, G. G. P. Rapid detection of whey in milk powder samples by spectrophotometric and multivariate calibration. **Food Chemistry**, v.174, p.1-7, 2015.

18. CEVOLI, C.; CERRETANI, L.; GORI, A.; CABONI, M.F.; TOSCHI, T. G.; FABBRI, A. Classification of Pecorino cheeses using electronic nose combined with artificial neural network and comparison with GC–MS analysis of volatile compounds. **Food Chemistry**, v.129, p.1315–1319, 2011.

19. CHOLLET, F., et al.: **Keras**. Available at https://keras.io (2015)

20. COATES, J. Vibrational Spectroscopy: Instrumentation for infrared and Raman spectroscopy. **Applied Spectroscopy Review**, v.33, p.267-425, 1998.

21. COITINHO, T. B.; CASSOLI, L. D.; CERQUEIRA, P. H. R.; SILVA, H. K.; COITINHO, J. B.; MACHADO, P. F. Adulteration identification in raw milk using fourier transform infrared spectroscopy. **Journal of Food Science and Technology**, v.54, n.8, p.2394-2402, 2017.

22. CONCEIÇÃO, D. G.; GONÇALVES, B. R. F.; da HORA, F. F.; FALEIRO, A. S.; SANTOS, L. S.; FERRÃO, S. P. B. Use of FTIR-ATR spectroscopy combined with multivariate analysis as a screnning tool to identify adulterants in raw milk. **Journal of the Brazilian Chemical Society**, v.30, n.4, p.780-785, 2019.

23. COSSIGNANI, L.; BLASI, F.; BOSI, A.; D'ARCO, G.; MAURELLI, S.; SIMONETTI, M. S.; DAMIANI, P. Detection of cow milk in donkey milk by

chemometric procedures on triacylglycerol stereospecific analysis results. **Journal of Dairy Research**, v.78, p.335–342, 2011.

24. COSTA FILHO, P. A.; POPPI, R. J. Application of genetic algorithms in the variable selection in mid infrared spectroscopy: simultaneous determination of glucose, maltose and fructose. **Química Nova**, v.25, n.1, p.46-52, 2002.

25. DAS, S.; SIVARAMAKRISHNA, M.; BISWAS, K.; GOSWAMI, B. A low cost instrumentation system to analyze different types of milk adulteration. **ISA Transactions**, v.56, p.268-275, 2015.

26. DURIG, J. R.; SULLIVAN, J. F. Vibrational spectroscopy, Fourier transforms and analytical chemistry. **Trends in Analytical Chemistry**, v.9, n.4, p.104-106, 1990.

27. EIKREM, L. O. Process Fourier transform infrared spectroscopy. **Trends in Analytical Chemistry**, v.9, n.4, p.107-109, 1990.

28. FERRÃO, M. F.; CARVALHO, C. W.; MULLER, E. I. *et al*. Determinação simultânea dos teores de cinza e proteína em farinha de trigo empregando NIRR-PLS e DRIFT- PLS. **Ciência e Tecnologia de Alimentos**, v.24, n.3, p.333-340, 2004.

29. FIRMINO, F. C.; TALMA, S. V.; MARTINS, M. L.; LEITE, M. O.; MARTINS, A. D. O. Detecção de fraudes em leite cru dos tanques de expansão da região de rio Pomba, Minas Gerais. **Revista do Instituto de Laticínios Cândido Tostes,** v.65, n.376, p.5-11, 2010.

30. FRANCO, B. D. G. M.; LANDGRAF, M. **Microbiologia dos alimentos.** Atheneu, 182p., São Paulo, 2008.

31. FREITAS FILHO, J. R.; SOUZA FILHO, J. S.; GONÇALVES, T. M.; SOUZA, J. J. F.; SILVA, A. H. I.; OLIVEIRA, H. B.; BEZERRA, J. D. C. Caracterização físico-química e microbiológica do leite *in natura* comercializado informalmente no município de Garanhuns/PE. **Revista Brasileira de Tecnologia Agroindustrial,** v.03, n.02, p.38-46, 2009.

32. GONDIM, C. S.; JUNQUEIRA, R. G.; SOUZA, S. V. C.; RUISÁNCHEZ, I.; CALLAO, M. P. Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies. **Food Chemistry**, v.230, p.68-75, 2017.

33. GOYAL, S.; GOYAL, G. K. Artificial neural networks for dairy industry: A review. **Journal of Advanced Computer Science and Technology**, v.1, n.3, p.101-115, 2012.

34. GUZHVA, O.; ARDÖ, H.; NILSSON, M.; HERLIN, A.; TUFVESSON, L. Now You See Me: Convolutional Neural Network Based Tracker for Dairy Cows. **Frontiers in Robotics and AI**, v.5, n.107, 2018.

35. HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2nd ed., 745p., New York, 2017.

36. HELFER, G. A.; FERRÃO, M. F.; FERREIRA, C. V. Aplicação de métodos de análise multivariada no controle qualitativo de essências alimentícias empregando espectroscopia no infravermelho médio. **Ciência e Tecnologia de Alimentos**, v.26, n.4, p.779-786, 2006.

37. HERNÁNDEZ-RAMOS, P. A.; VIVAR-QUINTANA, A. M.; REVILLA, I. Estimation of somatic cell count levels of hard cheeses using physicochemical composition and artificial neural networks. **Journal of Dairy Science**, v.102, n.2, p.1014-1024, 2019.

38. HOLLER, F. JAMES; SKOOG, DOUGLAS A; CROUCH, STANLEY R. **Princípios de análise instrumental.** Tradução para o português: CELIO PASQUINI *et al* [Coord.]. 6a ed. Porto alegre, 2009.

39. JAWAID, S.; TALPUR, F. N.; SHERAZI, S. T. H.; NIZAMANI, S. M.; KHASKHELI, A. A. Rapid detection of melamine adulteration in dairy milk by SB-ATR-Fourier transform infrared spectroscopy. **Food Chemistry**, v.141, p.3066-3071, 2013.

40. KALASINSKY, K. S. Industrial applications of vibrational spectroscopy.

Trends in Analytical Chemistry, v.9, n.3, p.83-89, 1990.

41. KAMAL, M.; KAROUI, R. Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: A review. **Trends in Food Science & Technology**, v.46, p.27-48, 2015.

42. KAROUI, R.; BAERDEMAEKER, J. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. **Food Chemistry**, v.102, p.621-640, 2007.

43. KARTHEEK, M.; A.; SMITH, A.; KOTTAI MUTHU, A.; MANAVALAN, R. Determination of Adulterants in Food: A Review. **Journal of Chemical and Pharmaceutical Research,** v.3, n.2, p.629-636, 2011.

44. KASEMSUMRAN, S.; THANAPASE, W.; KIATSOONTHON, A. Feasibility of near-infrared spectroscopy to detect and to quantify adulterants in cow milk. **Analytical Sciences**, v.23, p.907-910, 2007.

45. KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: **31st Conference on Neural Information Processing Systems** (NIPS 2017), Long Beach, CA, USA, 2017.

46. KESAVARAJ, G.; SUKUMARAN,S. A Study on Classification Techniques in Data Mining. **IEEE International Conference on Computational Intelligence and Computing Research**. Tiruchengode, India, july, 2013.

47. KINGMA, D. P.; BA, J. ADAM: A method for stochastic optimization. CoRR **abs/1412.6980**, 2014.

48. KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. Appears in the International Joint Conference on Artificial Intelligence (IJCAI), 1995.

49. LIU, J.; OSADCHY, M.; ASHTON, L.; FOSTER, M.; SOLOMON, C. J.; GIBSON, S. J. Deep convolutional neural networks for Raman spectrum recognition: A unified solution. **Analyst**, v.142, n.21, p.4067–4074, 2017.

50. LUNA, A. S.; PINHO, J. S. A.; MACHADO, L. C. Discrimination of adulterants in UHT milk samples by NIRS coupled with supervision discrimination techniques. **Analytical Methods**, v.8, p. 7204-7208, 2016.

51. MENDES, C. G.; SAKAMOTO, S. M.; SILVA, J. B. A.; JÁCOME, C. G. M.; LEITE, A. I. Análises físico-químicas e pesquisa de fraude no leite informal comercializado no município de Mossoró, RN. **Ciência Animal Brasileira**, v.11, n.2, p.349-356, 2010.

52. MICHIE, D.; SPIEGELHALTER, D.; TAYLOR, C. **Machine Learning, Neural and Statistical Classification**. Ellis Horwood Series in Artificial Intelligence, Ed. Ellis Horwood, 1994.

53. MONTANHINI, M. T. M.; HEIN, K. K. Qualidade do leite cru comercializado informalmente no município de Piraí do Sul, estado do Paraná, Brasil. **Revista Instituto de Laticínios "Cândido Tostes"**, v.68, n.393, p.10-14, 2013.

54. MORGANO, M. A.; FARIA, C. G.; FERRÃO, M. F. *et al.* Determination of protein in raw coffee for NIR spectroscopy and regression PLS. **Ciência e Tecnologia de Alimentos**, v.25, n.1, p.25-31, 2005.

55. NICOLAOU, N.; XU, Y.; GOODACRE, R. Fourier transform infrared spectroscopy and multivariate analysis for the detection and quantification of different milk species. **Journal of Dairy Science**, v.93, n.12, p.5651–5660, 2010.

56. NIKAM, S. S. A comparative study of classification techniques in data mining algorithms. **Oriental Journal of Computer Science & Technology**, v.8, n.1, p.13-19, 2015.

57. OLIVEIRA, M. C. P. P.; SILVA, N. M. A.; BASTOS, L. P. F.; FONSECA, L. M.; CERQUEIRA, M. M. O. P.; LEITE, M. O.; CONRRADO, R. S. Fourier transform infrared spectroscopy (FTIR) for MUN analysis in normal and adulterated milk. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.64, n.5, p.1360-1366, 2012.

58. PAVIA, D. L.; LAMPMAN, G. M.; KRIZ, G. S.; VYVYAN, J. R. **Introdução**

à espectroscopia. Cengage Learning, Trad. Pedro Barros. São Paulo, 2012.

59. PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v.12, n.2825–2830, 2011.

60. POONIA, A.; JHA, A.; SHARMA, R.; SINGH, H. B.; RAI, A. K.; SHARMA, N. Detection of adulteration in milk: A review. **International Journal of Dairy Technology**, v.70, n.1, p.23-42, 2017.

61. RIBEIRO, D. C. S. Z.; TAVARES, W. L. F.; LEITE, M. O.; CERQUEIRA, M. M. O. P.; LIMA, J. S.; FERREIRA, L. F.; FEIJO, F. A. C.; HADDAD, J. P.; FONSECA, L. M. Adulterants interference on Fourier-transform infrared analysis of raw milk. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.70, p.1649-1654, 2018.

62. SALA, O. **Fundamentos da espectroscopia Raman e no Infravermelho**, 2a ed. Unesp: São Paulo, 2008.

63. SALIBA, E. O. S.; GONTIJO NETO, M.M; RODRIGUES, N.M. *et al*. Predição da composição química do sorgo pela técnica de espectroscopia de refletância no infravermelho próximo. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.55, n.3, 2003.

64. SANDA, A. C. M. M.; SILVA, T. L.; PIVA, K. P.; SANDA, R. T.; ORSINE, J. V. C. Características do leite cru consumido pela população de Pires do Rio – GO. **Revista HCPA**, v.33, n.2, p.127-134, 2013.

65. SANTOS, M. V.; FONSECA, L. F. L. **Estratégias para controle de mastite e melhoria da qualidade do leite.** São Paulo: Manole, 314p., 2007.

66. SANTOS, N. A. F.; LACERDA, L. M.; RIBEIRO, A. C.; LIMA, M. F. V. Avaliação da composição e qualidade físico-química do leite pasteurizado padronizado comercializado na Cidade de São Luiz, MA. **Arquivos do Instituto**

Biológico, v.78, n.1, p.109-113, jan./mar., 2011.

67. SANTOS, P. M.; PEREIRA-FILHO, E. R.; RODRIGUEZ-SAONA, L. E. Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis. **Food Chemistry**, v.138, p.19-24, 2013.

68. SANTOS, P. M.; WENTZELL, P. D.; PEREIRA-FILHO, E. R. Scanner digital images combined with color parameters: a case study to detect adulterations in liquid cow´s milk. **Food Analytical Methods**, v.5, p.89-95, 2012.

69. SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v.61, p.85–117, 2015.

70. SHARMA, K.; PARADAKAR, M. The melamine adulteration scandal. **Food Security,** v.2, p.97-107, 2010.

71. SILVA, L. C. C.; TAMANINI, R.; PEREIRA, J. R.; RIOS, E. A.; RIBEIRO JUNIOR, J. C.; BELOTI, V. Preservatives and neutralizing substances in milk: analytical sensitivity of official specific and nonspecific tests, microbial inhibition effect, and residual persistence in milk. **Ciência Rural**, v.45, n.9, p.1613-1618, 2015.

72. SILVEIRA, T. M. L.; FONSECA, L. M.; CANÇADO, S. V. *et al*. Comparação entre os métodos de referência e a análise eletrônica na determinação da composição do leite bovino. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.56, n.6, p.782-787, 2004.

73. SILVERSTEIN, R. M.; BASSLER, G. C.; MORRIL, T. C. **Identificação Espectrométrica de Compostos Orgânicos**. Guanabara Koogan, 5a ed., Rio de Janeiro, 1994.

74. SKOOG, D. A.; HOLLER, F. J.; NIEMAN, T. A. **Princípios de análise instrumental**. Editora Bookman, 5a.ed, 836p., Porto Alegre, 2002.

75. SOUZA, S. S.; CRUZ, A. G.; WALTER, E. H. M.; FARIA, J. A. F.; CELEGHINI, R. M. S.; FERREIRA, M. M. C.; GRANATO, D.; SANTANA, A.

S. Monitoring the authenticity of Brazilian UHT milk: A chemometric approach. **Food Chemistry,** v.124, p.692-695, 2011.

76. TRONCO, V. M. **Manual para inspeção da qualidade do leite.** UFSM, 3a ed., 206p., Santa Maria, 2008.

77. VALENTE, G. F. S.; GUIMARÃES, D. C.; GASPARDI, A. L. A.; OLIVEIRA, L. A. Aplicação de redes neurais artificiais como teste de detecção de fraude de leite por adição de soro de queijo. **Revista do Instituto de Laticínios Cândido Tostes**, v.69, n.6, p.425-432, 2014.

78. VARGAS, A. C. G.; CARVALHO, A. M. P.; VASCONCELOS, C. N. Um estudo sobre Redes Neurais Convolucionais e sua aplicação em detecção de pedestres. **In: Proceedings of the XXIX Conference on Graphics, Patterns and Images**, 4p., 2016.

79. WEBSTER, L.; SIMPSON, P.; SHANKS, A. M.; *et al*. The authentication of olive oil on the basis of hydrocarbon concentration and composition. **Analyst London**, v.125, p. 97-104, 2000.

80. WILLIAMS, P. C. Commercial Near-Infrared Reflectance Analyzers. In: WILLIAMS, P. C.; NORRIS, K. H. (Ed) **Near-infrared technology in the agricultural and food industries.** Saint Paul: American Association of Cereal Chemists, p.1074-142, 1987.

81. YANG, R.; LIU, R.; XU, K. Detection of adulterated milk using two-dimensional correlation spectroscopy combined with multi-way partial least squares. **Food Bioscience**, v.2, p.61-67, 2013.

82. ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: FLEET, D.; PAJDLA, T.; SCHIELE, B.; TUYTELAARS, T. (eds.) Computer Vision – ECCV 2014. **Lecture Notes in Computer Science**, v.8689, p.818–833, Springer Cham, 2014.

83. ZHANG, L. G.; ZHANG, X.; NI, L. J.; XUE, Z. B.; GU, X.; HUANG, S. X. Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy. **Food Chemistry**, v.145, p.342–

348, 2014.

84. ZHANG, X. J.; LU, Y. F.; ZHANG, S. H. Multi-task learning for food identification and analysis with deep convolutional neural networks. **Journal of Computer Science and Technology**, v.31, n.3, p.489–500, 2016.

85. ZOU, H.; ZHANG, W.; FENG, Y.; LIANG, B. Simultaneous determination of melamine and dicyandiamide in milk by UV spectroscopy coupled with chemometrics. **Analytical Methods**, v.6, p.5865-5871, 2014.

8. ANEXOS

Produção científica relacionada à tese:

- Artigo 01:

RIBEIRO, D. C. S. Z.; TAVARES, W. L. F.; LEITE, M. O.; CERQUEIRA, M. M. O. P.; LIMA, J. S.; FERREIRA, L. F.; FEIJO, F. A. C.; HADDAD, J. P.; FONSECA, L. M. Adulterants interference on Fourier-transform infrared analysis of raw milk. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.70, p.1649-1654, 2018.

- Artigo 02:

NETO, H. A.; TAVARES, W. L. F.; RIBEIRO, D. C. S. Z.; ALVES, R. C. O.; FONSECA, L. M.; CAMPOS, S. V. A. On the utilization of deep and ensemble learning to detect milk adulteration. **BioData Mining**, v.12, n. 13, p.1-13, 2019.

Adicionalmente, temos 2 artigos em fase final de revisão para envio para publicação, não constantes no anexo:

- Artigo 03:

TAVARES, W. L. F.; RIBEIRO, D. C. S. Z.; NETO, H. A.; LIMA, J. S.; LEITE, M. O.; CERQUEIRA, M. M. O. P.; FONSECA, L. M. Use of FTIR spectroscopy and data mining for detection of adulterants in raw milk. 2019.

- Artigo 04:

TAVARES, W. L. F.; RIBEIRO, D. C. S. Z.; LIMA, J. S.; LEITE, M. O.; CERQUEIRA, M. M. O. P.; HADDAD, J. P.; FONSECA, L. M. Quimiometria aplicada à detecção de adulterantes no leite cru por meio da espectrofotometria no infravermelho com transformada de fourier (FTIR). 2019.

Arq. Bras. Med. Vet. Zootec., v.70, n.5, p.1649-1654, 2018

Communication

[*Comunicação*]

Adulterants interference on Fourier Transform Infrared analysis of raw milk

[*Interferência de adulterantes na análise do leite cru por espectroscopia pela transformada de Fourier*]

D.C.S.Z. *Ribeiro*, W.L.F. *Tavares*, M.O. *Leite*, M.O.P. *Cerqueira*, J.S. *Lima*, L.F. *Ferreira*, F.A.C. *Feijó*, J.P. *Haddad*, L.M. *Fonseca**

Universidade Federal de Minas Gerais - Belo Horizonte, MG

Milk is one of the most complete foods in nature, with high nutritive value and characteristic physical-chemical, sensorial, and microbiological properties. The physical-chemical analysis is a tool to evaluate its value or the industrial yield as well as to detect possible milk frauds (Silva, 2013).

Usual major milk composition is of 87.3% water, 4.6% lactose, 3.25% protein, 3.9% fat, 12.7% total solids, and non-fat 8.8% solids. Knowledge of milk components and characteristics is the basis for tests that are designed to investigate modified milk composition. Substantial reduction of solid components could, for example, indicate the fraudulent addition of water to the milk. Altogether, adulteration and variations due to factors such as lactation stage, breed, feeding, environmental temperature, management, milking interval, production and mammary gland health may affect milk composition (Brito *et al*., 2016). Optimal composition is a preponderant factor for the dairy industry, in processes such as cheese, butter, powdered milk, milk cream, cream cheese and other products (Leite, 2006).

Milk components, and particularly fat and protein are important for quality and integrity determination. Additional analyses for somatic cell count (SCC) and individual bacteria counting are complementary for this quality measurement (Brandão *et al*., 2010; Oliveira *et al.*, 2012).

Factors such as age of sample, transport conditions and use of preservatives may interfere in the quality evaluation. For example, bronopol is a preservative added to the raw milk sample to avoid quality loss; usually samples may be kept ten days at temperatures up to 10^{o}C (Leite, 2006). However, other authors have reported decrease in milk fat concentration after seven days, and slight modifications in protein and total solids concentration after storage of six days (Ribas *et al*., 2004).

In Brazil, new variables were legalized to ensure a minimum standard to be met by all milk producers (Brasil, 2002; 2011; 2016). Several solid components, IBC, SCC, and drug residues were included in this new regulation for at least monthly monitoring of producers linked to dairies under Federal Inspection. A national net of laboratories was also created to respond to the growing analytical demand (Rede Brasileira de Laboratórios de Análise da Qualidade do Leite; RBQL). These labs are configured with modern electronic equipment for quick and precise analysis, to provide milk quality information in the adequate time (Silveira *et al*., 2004).

Raw milk must present the original composition, and any component withdrawal, or addition of water, preservatives, contaminants, or any other substance to the raw milk is illegal. The consequence of fraud is consumer misleading and risks including health issues, illegal competition, economical and marketing unbalance. One outstanding example took place in China, where a widespread fraud with melamine addition to raw milk happened in 2008. The outcome was the death of six and more than 300,000 people were sickened.

Recebido em 23 de janeiro de 2017
Aceito em 16 de maio de 2018
E-mail: leorges@ufmg.br

Melamine was added with the intent to artificially increase the protein levels to disguise water illegally added to raw milk (Silva, 2013).

Similar incidents have come to light in the last years in Brazil, with several cases of systematic adulteration being detected in inspected dairy industries. Common aspects of these cases included water addition to the raw milk together with substances to counterbalance properties modification of the milk due to the extraneous water. For example, to restore density and freezing point, salt, sugar, and other substances are used. Impairment of the milk nutritional value is a direct consequence of these frauds. Additionally, preservative is illegally added to decrease microbial growth in low quality raw milk (Rosa-Campos *et al*., 2011; Santos *et al*., 2013; Silva, 2013; Abrantes *et al*., 2014).

Although several analytical methods have been used for a long time to investigate raw milk adulteration in the dairy industry, usually they are time consuming and expensive. However, equipments with relatively low analytical cost have been developed in recent decades. Mid infrared equipment, for example, has expanded the analytical capacity for compositional analysis, while flow cytometers have been used for somatic cell and bacterial counting in raw milk (Oliveira *et al.*, 2012).

The infrared equipment is based in differential IR absorption in specific wavelengths of chemical groups present in fat, protein, and lactose in milk (Biggs *et al.*, 1987). For milk, the mid infrared spectrum (MIR) represents great advantages when compared to previous techniques, as the simultaneous reading of several milk components, without any previous sample treatment, with high output and possibility of process automation (Rodriguez-Otero *et al.*, 1997). The IR technique matched the reference methods (Silveira *et al*., 2004), with major milk components being simultaneously analyzed.

The automated routine analysis of raw milk with electronic equipment usually includes the main milk components total solids, fat, protein, casein, ureic nitrogen, and lactose among others through IR analysis, and somatic cells and bacterial counting by flow cytometry. Nowadays, old IR equipment based on filters are being replaced by FTIR based equipment, a new and promising technology (Oliveira *et al.*, 2012).

Although the infrared methods have been extensively evaluated for milk quality analysis, not much has advanced in the study of extraneous substances used as adulterant in milk. One of the concerns is related to the potential interference of these substances in the analysis of milk components because of superimposed absorptive wavelengths. Moreover, since the current methods used to detect adulterants in raw milk are relatively expensive, highly dependent on labor and time-consuming, innovative screening methods to simultaneously detect extraneous substances added to milk would represent an expressive advance in the dairy sector.

The objective of the current work was to evaluate the potential interference of extraneous substances, added to milk as adulterants, in the component analysis based on Fourier Transform Infrared technique.

Six batches of bulk tank raw milk were obtained from August to October 2016, from healthy cows, under veterinary supervision. To each batch sucrose was added (0.1g/100mL; 0.5g/100mL and 1g/100mL) or starch (0.1g/100mL; 0.5g/100mL and 1g/100mL) and aliquoted in 50mL vials. Samples for compositional analysis and somatic cell counts were added with bronopol, and for bacterial counting with azidiol.

Samples were stored during 3, 24, 48, 72, and 168 hours at 7°C±2°C and 25±2°C. The number of samples were 1008 (six batches x two adulterants x three concentrations x six storage time x two storage temperatures x two preservatives; and 144 samples as control) (Ribas *et al*., 2004; Leite, 2006; Rosa-Campos *et al*., 2011).

Compositional analyses were done by FTIR 400® (LactoScope FTIR, Delta Instruments, Drachten, Netherlands) and some spectra are shown in Figure 1. Somatic cell count by flow cytometry in FTIR 400® (SomaScope, Delta Instruments, Drachten, Netherlands) and total bacterial counting by flow cytometry in Bactoscan FC® (Foss, Hillerod, Denmark). Multiple Linear Regression was used for

statistical analysis with a confidence interval of 95%, using the program Stata version 12.0 (Stata..., 2011).

Milk analyzed for SCC and TBC were monitored to ensure IR readings deviations for components were not due to quality issues. However, starch and sucrose additon affected both values. Starch addition resulted in increasing SCC and TBC, while sucrose addition resulted in reducing SCC readings (Table 1).

Storage time was an influential factor, with SCC reduction after 24h, but this effect was more pronounced with storage at 25±2°C and starch addition. Previous report by Leite (2006), in raw milk without adulterants, has found SCC reduction storage at 30ºC during five days, while TBC increased. In the current work, TBC increased was noticed in the seventh day of storage at 25ºC.

Every experimental repetition presented extremely low p-values for the all dependent variables, except for TBC, which was significant only in the last repetition. This indicates the strict control of the milk used in the experiment, once the variables that could adversely affect the compositional results outcome, such as lactation stage, calving order, milking management and hygiene, feeding, among others were controlled in the experimental repetitions (Brito *et al.*, 2016). Similarly, high *Adjusted* R^2 value was found for all dependent variables, except for TBC in the milk added with sucrose treatment. This indicates that the multiple linear regression was well adjusted to the sample data (Triola, 2008).

Composition results for milk added with starch deviated from the control with increase of parameters fat, protein, lactose, total solids, solids non fat and reduction of casein and MUN using FTIR ($P< 0.05$). However, the addition of sucrose to milk modified FTIR results for lactose, TS, SNF and MUN with increasing these values and decreasing protein and casein (Table 2).

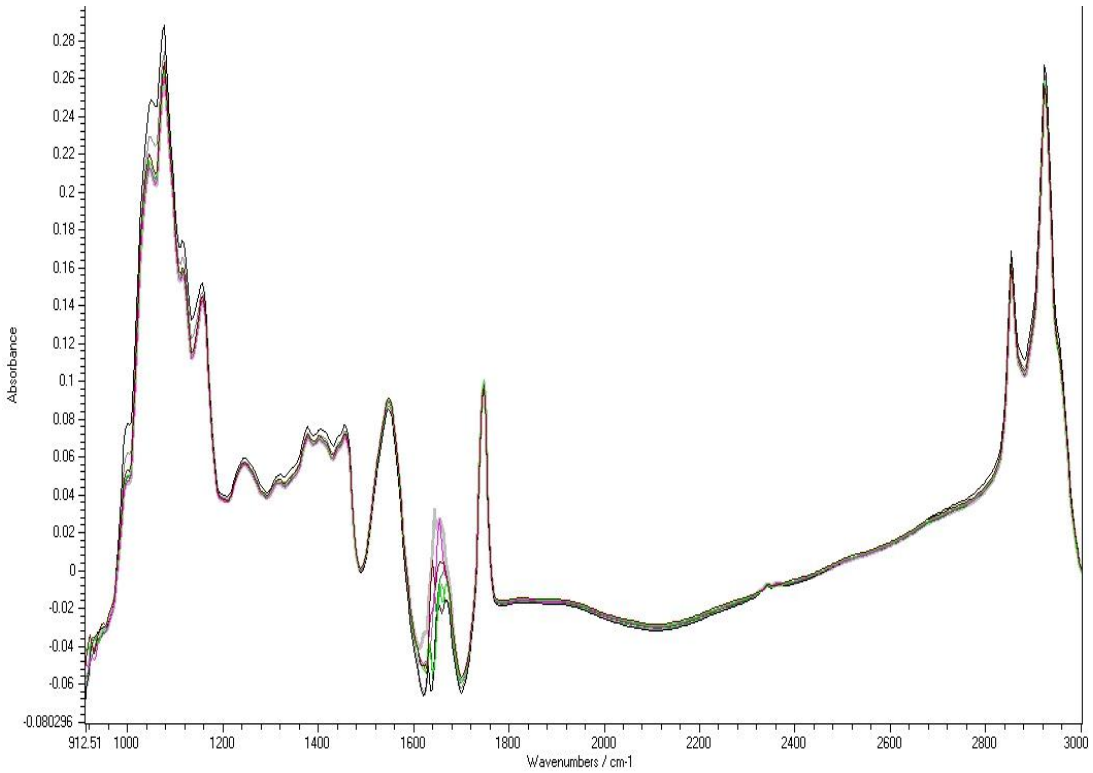


Figure 1. FTIR spectra of milk control and added starch and sucrose, in three different concentrations.

Table 1. Mean and standard deviation (sd) of the SCC and TBC in raw milk added with starch and sucrose and analyzed by flow cytometry

	Control		Starch						Sucrose					
Temp.(°C)			0,1%		0,5%		1%		0,1%		0,5%		1%	
Time(h)	SCC	TBC	SCC	TBC	SCC	TBC	SCC	TBC	SCC	TBC	SCC	TBC	SCC	TBC
7°C	71 (54)	5 (7)	76 (59)	151 (16)	99 (51)	683 (43)	132 (54)	1199 (171)	67 (53)	6 (7)	66 (52)	6 (9)	64 (52)	7 (11)
0h	68 (58)	7 (12)	78 (71)	150 (22)	95 (51)	710 (67)	145 (49)	1297 (166)	70 (57)	7 (10)	61 (51)	7 (13)	68 (63)	9 (150
3h	73 (52)	6 (9)	80 (58)	161 (26)	98 (41)	668 (36)	134 (46)	1255 (10)	68 (53)	9 (10)	67 (49)	9 (13)	66 (510	13 (18)
24h	73 (50)	4 (6)	74 (51)	148 (15)	97 (55)	676 (35)	130 (70)	1127 (119)	63 (51)	6 (9)	62 (51)	6 (8)	66 (58)	6 (10)
48h	60 (54)	4 (4)	69 (56)	147 (10)	99 (48)	660 (35)	131 (47)	1086 (244)	71 (52)	4 (5)	64 (55)	7 (9)	62 (55)	6 (8)
72h	67 (62)	4 (4)	71 (66)	151 (6)	99 (57)	702 (39)	116 (59)	1249 (119)	63 (57)	5 (5)	62 (54)	5 (5)	59 (52)	6 (9)
7d	84 (69)	3 (2)	86 (72)	149 (6)	108 (73)	680 (26)	138 (71)	1181 (195)	69 (71)	4 (4)	78 (74)	4 (4)	65 (59)	3 (3)
25°C	68 (57)	134 (397)	71 (55)	368 (527)	91 (52)	859 (467)	118 (56)	1373 (432)	64 (54)	200 (503)	63 (54)	150 (361)	62 (55)	148 (399)
0h	76 (62)	7 (10)	76 (57)	157 (26)	98 (59)	698 (34)	143 (59)	1335 (99)	68 (60)	7 (11)	69 (56)	9 (15)	73 (68)	10 (19)
3h	67 (53)	6 (9)	77 (57)	156 (20)	104 (53)	680 (63)	126 (45)	1216 (165)	71 (59)	7 (10)	68 (54)	8 (12)	67 (53)	9 (16)
24h	84 (65)	6 (6)	74 (61)	146 (27)	98 (51)	633 (74)	130 (66)	1151 (290)	66 (55)	7 (9)	66 (59)	7 (9)	68 (59)	12 (18)
48h	67 (63)	11 (13)	71 (52)	142 (9)	93 (55)	653 (78)	119 (55)	1175 (173)	63 (50)	12 (20)	63 (57)	21 (37)	60 (58)	21 (27)
72h	62 (61)	50 (70)	67 (59)	408 (542)	87 (53)	878 (538)	114 (61)	1297 (222)	61 (59)	107 (157)	61 (57)	96 (141)	55 (52)	99 (215)
7d	54 (59)	725 (772)	63 (70)	1197 (770)	68 (53)	1614 (592)	78 (52)	2063 (626)	55 (64)	1064 (815)	54 (65)	757 (623)	51 (61)	735 (748)

Temp.: temperature SCC – (x1000 SC/mL) TBC – (x1000 CFU/mL) n=1008.

Table 2. Mean and standard deviation (sd) of the physical chemical parameters of raw milk added with starch and sucrose analyzed by FTIR

		Control	Starch			Sucrose		
			0,1%	0,5%	1%	0,1%	0,5%	1%
Fat (g/100g)	7°C	3,35 (0,32)	3,38 (0,33)	3,41 (0,34)	3,44 (0,36)	3,37 (0,32)	3,35 (0,31)	3,35 (0,32)
	25°C	3,37 (0,32)	3,40 (0,34)	3,43 (0,34)	3,44 (0,36)	3,37 (0,32)	3,37 (0,32)	3,36 (0,33)
Protein (g/100g)	7°C	3,35 (0,03)	3,36 (0,03)	3,36 (0,03)	3,37 (0,03)	3,35 (0,03)	3,34 (0,03)	3,33 (0,03)
	25°C	3,35 (0,04)	3,36 (0,03)	3,36 (0,03)	3,37 (0,03)	3,35 (0,03)	3,34 (0,03)	3,33 (0,03)
Lactose (g/100g)	7°C	4,65 (0,06)	4,67 (0,06)	4,71 (0,06)	4,77 (0,06)	4,70 (0,06)	4,92 (0,06)	5,19 (0,05)
	25°C	4,66 (0,06)	4,67 (0,06)	4,72 (0,06)	4,78 (0,06)	4,71 (0,06)	4,92 (0,06)	5,19 (0,06)
TS (g/100g)	7°C	12,32 (0,30)	12,37 (0,32)	12,45 (0,33)	12,54 (0,34)	12,39 (0,30)	12,56 (0,30)	12,80 (0,31)
	25°C	12,35 (0,31)	12,40 (0,33)	12,48 (0,33)	12,54 (0,35)	12,40 (0,31)	12,59 (0,31)	12,82 (0,32)
SNF (g/100g)	7°C	9,01 (0,07)	9,03 (0,08)	9,09 (0,08)	9,15 (0,08)	9,06 (0,08)	9,28 (0,08)	9,55 (0,08)
	25°C	9,01 (0,08)	9,04 (0,08)	9,09 (0,08)	9,16 (0,08)	9,07 (0,08)	9,28 (0,08)	9,55 (0,08)
MUN (mg/dL)	7°C	14,53 (1,94)	14,36 (2,07)	10,55 (2,04)	5,14 (2,92)	14,80 (2,00)	15,04 (1,97)	14,92 (1,69)
	25°C	14,97 (2,07)	15,10 (1,86)	11,07 (2,27)	7,05 (2,58)	15,73 (1,97)	15,75 (2,04)	15,12 (1,95)
Casein (g/100g)	7°C	2,63 (0,02)	2,62 (0,02)	2,61 (0,02)	2,60 (0,03)	2,62 (0,02)	2,62 (0,02)	2,62 (0,02)
	25°C	2,63 (0,02)	2,62 (0,03)	2,61 (0,03)	2,61 (0,03)	2,63 (0,02)	2,63 (0,03)	2,62 (0,02)
FP (°C)	7°C	-0,530 (0,01)	-0,532 (0,01)	-0,537 (0,01)	-0,543 (0,01)	-0,535 (0,01)	-0,555 (0,01)	-0,580 (0,01)
	25°C	-0,531 (0,01)	-0,533 (0,01)	-0,538 (0,01)	-0,543 (0,01)	-0,536 (0,01)	-0,556 (0,01)	-0,580 (0,01)

n=504; FTIR: Fourier Transform Infrared; TS: total solids; SNF: solids nonfat; MUN: milk urea nitrogen; FP: freezing point.

Lactose IR readings were slightly higher in milk added with starch, however, after sucrose addition, lactose content readings reached values as high as 5.19% (Table 2).

Freezing point of milk decreased with increasing sucrose concentration, and this is due to the increased amount of soluble substance. As a colligative property, the freezing point is correlated to the vapour pressure at the solution according to Raoult's Law. On the other hand, since starch molecules are longer and have higher molecular weight than sucrose, it is expected that starch addition to the milk will have lesser effect on freezing point.

The use of sucrose and starch as milk adulterants is a very common fraud. For example, Rosa-Campos *et al*. (2011) analyzed 72 pasteurized milk samples and in some cases found sucrose addition in 100% of some brands samples.

MUN values obtained by IR were lower ($P< 0.05$) after starch addition, even during storage at 7ºC and 25ºC. On the other hand, MUN values slightly increased in the control group during storage. Casein results were lower in milk added with starch or sucrose than in control milk.

In conclusion, infrared analysis of raw milk added with extraneous substances used as adulterants presented deviating results for component analysis due to superimposed absorptive wavelengths.

Starch addition changed expected compositional results of raw milk in a different direction and intensity when compared with sucrose, with interference in the major components, fat and protein. Simultaneously, increasing individual bacterial count by flow cytometry were obtained after starch addition.

Keywords: *raw milk, FTIR, startch, sucrose, composition analysis*

ACKNOWLEDGEMENTS

Fapemig – APQ1179-14; CNPq PQ309801/2014-1

RESUMO

O objetivo deste trabalho foi analisar as leituras de composição do leite cru por meio de espectrofotometria FTIR, utilizando-se curva de regressão PLS, bem como as contagens de células somáticas e bacteriana total por citometria de fluxo, após adição de amido e sacarose. O leite cru foi adulterado com três concentrações de amido e sacarose (0,1%, 0,5% e 1%), colocado em frascos contendo bronopol ou azidiol, os quais foram armazenados em duas temperaturas (7±2°C e 25±2°C). As análises foram realizadas após zero, três, 24, 48, 72 e 168 horas de armazenamento. O modelo de regressão linear múltipla foi utilizado para análise estatística. A adição de amido e sacarose resultou em mudança significativa (P<0,05) para todas as variáveis dependentes. O leite adulterado com amido resultou em aumento nas leituras de gordura, proteína, lactose, sólidos totais (ST), sólidos não gordurosos (SNG), CCS e CBT e em diminuição das leituras de caseína, do nitrogênio ureico do leite (NUL) e do ponto de congelamento. O leite adulterado com sacarose resultou no aumento das leituras da lactose, ST, SNG e NUL, enquanto as leituras de proteína, caseína, ponto de congelamento e CCS diminuíram. Este trabalho evidencia a importância do monitoramento de adulterantes reconstituintes no leite por afetarem os resultados analíticos da qualidade do leite, obtidos por métodos eletrônicos.

Palavras-chave: leite cru, FTIR, amido, sacarose, análise composicional

REFERENCES

ABRANTES, M.R.; CAMPÊLO, C.S.; SILVA, J.B.A. Fraude em leite: métodos de detecção e implicações para o consumidor. *Rev. Inst. Adolfo Lutz*, v.73, p.244-251, 2014.

BIGGS, D.A.; JOHNSSON, G.; SJAUNJA, L.O. Analysis of fat, protein, lactose and total solids by infra-red absorption. In: monograph on rapid indirect methods for measurement of the major components of milk. *Bull. Int. Dairy Fed.*, n.208, p.21-29, 1987.

BRANDÃO, M.C.M.P.; CARMO, A.P.; BELL, M.J.V. *et al*. Characterization of milk by infrared spectroscopy. *Rev. Inst. Latic. Cândido Tostes*, v.65, p.30-33, 2010.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n.51, de 18/09/2002. Aprova os regulamentos técnicos de produção, identidade e qualidade do leite tipo A, do leite tipo B, do leite tipo C, do leite pasteurizado e do leite cru refrigerado e o regulamento técnico da coleta de leite cru. *Diário Oficial da União*, Brasília, 20 set. 2002. Seção I, p.13.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n.62, de 29/12/2011. Aprova o Regulamento Técnico de Identidade e Qualidade do leite tipo A, o Regulamento Técnico de Identidade e Qualidade do leite cru refrigerado, o Regulamento Técnico de Identidade e Qualidade do leite pasteurizado e o Regulamento Técnico da Coleta de Leite Cru Refrigerado e seu transporte a granel. *Diário Oficial da União*, Brasília, 30 dez. 2011b. Seção I, p.6.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n.7, de 03/05/2016. Altera o Regulamento Técnico de Identidade e Qualidade do leite tipo A, o Regulamento Técnico de Identidade e Qualidade do leite cru refrigerado, o Regulamento Técnico de Identidade e Qualidade do leite pasteurizado e o Regulamento Técnico da Coleta de Leite Cru Refrigerado e seu transporte a granel. *Diário Oficial da União*, Brasília, 04 mai. 2016. Seção I, p.11.

BRITO, M.A.; BRITO, J.R.; ARCURI, E. *et al.* 2016. *Composição do leite*. Disponível em: <http://www.agencia.cnptia.embrapa.br/Agencia8/AG01/arvore/AG01_128_21720039243.html>. Acessado em: 02 dez. 2016.

LEITE, M.O. *Fatores interferentes na análise eletrônica da qualidade do leite cru conservado com azidiol líquido, azidiol comprimido e bronopol*. 2006. 63f. Tese (Doutorado em Medicina Veterinária) – Escola de Veterinária, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

OLIVEIRA, M.C.P.P.; SILVA, N.M.A.; BASTOS, L.P.F. *et al*. Fourier Transform Infrared Spectroscopy (FTIR) for MUN analysis in normal and adulterated milk. *Arq. Bras. Med. Vet. Zootec.*, v.64, p.1360-1366, 2012.

RIBAS, N.P.; HARTMANN. W.; MONARDES, H.G. *et al*. Sólidos totais do leite em amostras de tanque nos Estados do Paraná, Santa Catarina e São Paulo. *Rev. Bras. Zootec*., v.33, Supl.3, p.2343-2350, 2004.

RODRIGUEZ-OTERO, J.L.; HERMIDA, M.; CENTENO, J. Analysis of dairy products by nearinfrared spectroscopy: a review. *J. Agrar. Food Chem.*, v.45, p.2815-2818, 1997.

ROSA-CAMPOS, A.A.; ROCHA J.E.S.; BORGO, L.A.; MENDONÇA, M.A. Avaliação físico-química e pesquisa de fraudes em leite pasteurizado integral tipo "c" produzido na região de Brasília, Distrito Federal. *Rev. Inst. Latic. Cândido Tostes*, v.66, p.30-34, 2011.

SANTOS, P.M.; PEREIRA-FILHO, E.R.; RODRIGUEZ-SAONA, L.E. Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis. *Food Chem.*, v.138, p.19-24, 2013.

SILVA, L.C.C. *Capacidade de detecção de adulterações e suficiência das provas oficiais para assegurar a qualidade do leite pasteurizado*. 2013. 99f. Tese (Doutorado em Ciência Animal), Universidade Estadual de Londrina, Londrina, PR.

SILVEIRA, T.M.L.; FONSECA, L.M.; CANÇADO, S.V. *et al*. Comparação entre os métodos de referência e a análise eletrônica na determinação da composição do leite bovino. *Arq. Bras. Med. Vet. Zootec.*, v.56, p.782-787, 2004.

STATA statistical software: release 12. College Station, Texas: StataCorp LP. 2011.

TRIOLA, M.F. *Introdução à estatística*. 10.ed. Rio de Janeiro: LTC, 2008. 696p.

Asseiss Neto *et al. BioData Mining* (2019) 12:13
https://doi.org/10.1186/s13040-019-0200-5

BioData Mining

METHODOLOGY Open Access

On the utilization of deep and ensemble learning to detect milk adulteration

Habib Asseiss Neto[1,2*], Wanessa L.F. Tavares[3], Daniela C.S.Z. Ribeiro[3], Ronnie C.O. Alves[4,5], Leorges M. Fonseca[3] and Sérgio V.A. Campos[2]

*Correspondence:
habib.asseiss@ifms.edu.br
[1]Federal Institute of Mato Grosso do Sul, Rua Ângelo Melão, 790, 79641-162 Três Lagoas, MS, Brazil
[2]Department of Computer Science, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, 31270-901 Belo Horizonte, MG, Brazil
Full list of author information is available at the end of the article

Abstract

Background: Fraudulent milk adulteration is a dangerous practice in the dairy industry that is harmful to consumers since milk is one of the most consumed food products. Milk quality can be assessed by Fourier Transformed Infrared Spectroscopy (FTIR), a simple and fast method for obtaining its compositional information. The spectral data produced by this technique can be explored using machine learning methods, such as neural networks and decision trees, in order to create models that represent the characteristics of pure and adulterated milk samples.

Results: Thousands of milk samples were collected, some of them were manually adulterated with five different substances and subjected to infrared spectroscopy. This technique produced spectral data from the milk samples composition, which were used for training different machine learning algorithms, such as deep and ensemble decision tree learners. The proposed method is used to predict the presence of adulterants in a binary classification problem and also the specific assessment of which of five adulterants was found through multiclass classification. In deep learning, we propose a Convolutional Neural Network architecture that needs no preprocessing on spectral data. Classifiers evaluated show promising results, with classification accuracies up to 98.76%, outperforming commonly used classical learning methods.

Conclusions: The proposed methodology uses machine learning techniques on milk spectral data. It is able to predict common adulterations that occur in the dairy industry. Both deep and ensemble tree learners were evaluated considering binary and multiclass classifications and the results were compared. The proposed neural network architecture is able to outperform the composition recognition made by the FTIR equipment and by commonly used methods in the dairy industry.

Keywords: Classification, Machine learning, Deep learning, Ensemble learning, Infrared spectroscopy, Milk, Adulteration

Background

Milk fraudulent adulteration consists of adding foreign substances to the milk. This is a common practice in Brazil and several countries worldwide [12], with the objective of increasing the product volume, disguising poor quality parameters and profiting with illegal actions [2, 7, 22]. Different substances can be added to milk with specific purposes. For instance, sucrose and starch are often used to modify density and freezing point after

extra water added to milk. Sodium bicarbonate can be added to reduce high acidity levels related to high bacteria contamination and bad manufacturing practices. Hydrogen peroxide and formaldehyde can preserve microbial count related to poor milk quality [9].

Fourier Transformed Infrared spectroscopy (FTIR) is one of the most commonly used techniques to read the composition of a sample in the food industry [9]. FTIR is a fast, nondestructive, and simple method that can be applied for milk composition analysis and it generates spectral data that can be computationally explored [22]. Machine learning techniques provide ways of understanding spectral data and producing useful knowledge regarding milk composition quality for consumers and regulatory agencies.

These techniques have been widely used in several areas and classification is a common machine learning task capable of understanding and categorizing data. In supervised learning, the classification task involves a training process with labeled data in order to generate a computational model that learns with that data [14]. Once the model is trained, it can be used to predict the label of new, unseen data. In the testing process, the model can be applied to a dataset and predicted labels can be compared to actual labels. Then, classification accuracy is used to evaluate the predictive capabilities of the model [10]. Deep and ensemble learners are two well-known methods with different characteristics that have shown excellent performance in several machine learning applications.

Ensemble learners are methods that combine many models' predictions. Bagging (Bootstrap Aggregating) is a technique that trains several machine learning models independently with randomly chosen subsets of data, and it uses majority voting for aggregating the outputs of base learners [14]. Boosting also trains classifiers using different training sets, but they are learned sequentially, with each model trying to minimize the error from the previous one. The combination of individually weak learners creates a better performing model [10]. Random Forest (RF) and Gradient Boosting Machine (GBM) are examples of bagging and boosting techniques, respectively. Since ensemble methods rely on the combination of models, they build smooth decision boundaries capable of finding the optimal feature and model combination to the classification problem [21].

In the area of *deep learning*, Convolutional Neural Networks (CNNs) are gaining great attention due to their high accuracy in pattern recognition and it has been successfully applied in a diversity of classification problems [19]. When compared to regular neural networks, additional layers (convolutional layers) are used in CNNs in order to filter input data and learn specific features from the data with different levels of abstraction [17].

Machine learning classifiers have been applied successfully in many applications, including image recognition, speech detection, and signal processing. Considering spectral data classification, decision trees have been used for classification of landscapes using satellite spectral data [8]. CNNs have been applied to electrocardiogram signals (ECG), significantly outperforming other ECG classification methods [16]. Mineral spectrum classification using CNN has achieved interesting results and has been compared to other machine learning methods [17]. CNNs also have been applied to audio spectral data for detecting sound events with human-level accuracy [15].

Milk adulteration analysis has been done with more traditional statistical methods, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) regression, that have been applied to infrared spectroscopy data in order to obtain adulteration estimates of whey, synthetic milk, hydrogen peroxide and others [22]. Milk adulteration by whey has also been studied by measuring specific proteins using PCA from spectral data [7].

Different milk adulterants have been analyzed with infrared spectroscopy using PCA multivariate analysis [9]. The work from [2] has similarities with our study by also using neural networks for milk adulteration detection. However, the authors used a regression model for quantifying the adulteration by a single ingredient (whey).

The objective of this work was to perform experiments with classification methods to recognize patterns in infrared milk composition in order to predict possible adulterations by foreign substances.

Methods

In this work, the characterization of bovine milk was made using machine learning techniques to detect the presence of milk adulterants or to assert which adulterant was found. In order to accomplish this, classification methods were used to determine milk sample adulteration. Classical statistical learning strategies such as Logistic Regression, Linear Regression, and PLS, usually employed by the industry [6, 7], were explored as benchmark models. Ensemble and deep learning classifiers were trained and tested on real, manually adulterated, milk samples in order to recognize patterns that identify adulteration characteristics.

Two versions of the classification problem were considered: binary and multiclass classifications. In the binary problem, the possible classes for a sample classification were either the presence or absence of an adulterant. In the multiclass problem, the classes were either one of the specific adulterant added to the milk or the "raw" class, when the sample has no adulterant added.

Data acquisition and sample preparation

Milk samples were acquired from the experimental farm at the Federal University of Minas Gerais, Brazil, and from the Laboratory for Milk Quality Analysis (Accredited ISO/IEC 17025) at the same university using commercial milk samples from the laboratory routine processes. A total of 4846 milk samples were collected, whereas 2376 were adulterated for the purpose of this study. The adulterated milk samples were added with one of five different substances (all of analytical grade): sucrose, soluble starch (amylose and amylopectin), sodium bicarbonate, hydrogen peroxide, and formaldehyde (Synth, Brazil). Although multiple adulterants can be found at once in a fraudulent milk sample [4, 24], in this work we aimed to analyze the effects of each adulterant individually, in order to describe how it affects pure milk composition.

FTIR spectroscopy was applied to all the collected milk samples in order to obtain infrared spectra, using the FTIR equipment (LactoScope™ FTIR 400, Delta Instruments, Drachten, The Netherlands), which outputs two pieces of information for each analyzed sample: an infrared spectrum file (SPC format) that contains coordinates for the infrared spectrum and a components file (CSV format), which contains numerical variables, called component features, that the equipment calculates from the infrared spectrum Additional file 2.

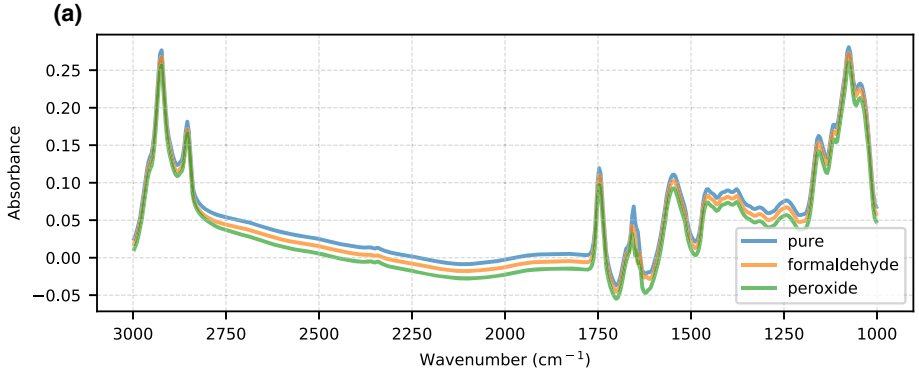
In our milk dataset, each sample is represented by both the component features and the spectral data. However, we used each of the two types of data differently. The component features data structure is ideal for the application of a decision tree classifier because each feature strongly represents some known characteristics in the milk composition. Since the combination of several classifiers may reduce the risk of an unfortunate selection of

a poorly performing classifier [21], ensemble tree learners were chosen for this task. On the other hand, the spectral data are composed by the full spectral coordinates, which can be interpreted as "images" for neural network recognition. For the latter, we used CNNs that are capable of detecting specific features from spectra without any required preprocessing. In Fig. 1 we show some spectra and some extracted component features.

For the purpose of estimating the quality of our classifiers, the hold-out cross-validation technique [3] was performed with three pairs of training/testing subsets with proportions: 90/10%, 75/25%, and 50/50%. This was the preferred split strategy since the same subsets needed to be tested with different classifiers, including deep learning, that usually splits datasets into training/validation/test sets. Each subset was obtained randomly from the original dataset (4846 samples) and the class distribution remained: ≈50% for raw milk and ≈10% for each of the five adulterant classes. Dataset samples distribution is described in Table 1. Detailed class distributions for each training and test dataset split are presented in Table 2.

Analysis of component features using ensemble learners

During the process of reading the infrared spectrum, the FTIR equipment performs a series of calculations that determines numerical values for different milk components. According to the equipment documentation, calculations are based on a Multiple Linear Regression (MLR) model that considers the absorbance of light energy by the sample for specific wavelength regions. The extracted information depends on the equipment



(b)

Cells (no. cells/µL)	QValue (%)	Fat (g/100g)	Protein (g/100g)	Lactose (g/100g)	Solids (g/100g)	FrzPoint (°H)	SNF (g/100g)	MUN (mg/dL)	Casein (g/100g)	Ingredient
59	95.10	3.42	3.38	4.73	12.50	0.578	9.14	11.1	2.62	pure
58	90.74	3.42	3.39	4.77	12.55	0.582	9.20	13.0	2.63	formaldehyde
46	91.76	3.37	3.35	4.74	12.44	0.578	9.12	27.8	2.56	peroxide

Fig. 1 **a** Plot of the infrared spectra for three randomly selected samples of the classes pure, formaldehyde and peroxide, acquired by the FTIR equipment. The raw spectra were analyzed directly by the proposed Convolutional Neural Network. Each spectrum was plotted with subtle shifts for viewing purpose. **b** Component features for the same samples, generated by the FTIR equipment and stored in CSV format. Each column quantifies an important milk composition information. The columns for fat, protein, lactose, solids, solids non-fat (SNF), casein and milk urea nitrogen (MUN) represent each component concentration in the sample. The Cells column represents the somatic cells counting, FrzPoint represents the freezing point of the sample, with values given in degrees Hortvet (°H), and QValue is a calculation of the sample quality by the equipment. These numerical features were analyzed by ensemble methods

Table 1 Sample distribution for the binary and multiclass versions of the collected dataset

Multiclass	# Samples	Binary	# Samples
Raw	2470	Raw	2470
Sucrose	486	Adulterated	2376
Formaldehyde	485		
Starch	480		
Peroxide	465		
Bicarbonate	460		

The full dataset has 4846 samples. The raw class is ≈50% of the dataset. Each adulterant class is ≈10% on multiclass, while the binary considers these samples as one class (≈50%)

calibrations for milk components concentration (fat, protein, lactose, total solids, solids non-fat/SNF, casein and milk urea nitrogen/MUN). Other three extra values are also included: the somatic cells counting, the freezing point value and a quality control value (Q-Value).

Pairwise correlations were calculated on standardized variables from the dataset. The relationship among these variables demonstrated that protein and casein are highly correlated (0.96). Since casein is a specific milk protein, the correlation makes sense. Correlations were also found with solids and fat (0.85), lactose with freezing point (0.77), and lactose with SNF (0.81). Other variables were found to be not expressively correlated. The complete feature correlation is presented in Additional file 1: Figure S1. All variables were read from equipment generated CSV file and were used as features in ensemble decision tree learners. The adulterants added to each sample were considered class labels for the samples and were used for training the classifiers. Figure 2 shows a boxplot considering scale and variation of all component features.

The component features were analyzed with Random Forest and Gradient Boosting Machine ensemble learners using the default implementations available in Scikit-learn [20]. The number of learners is controlled by the parameter `n_estimators` and it was

Table 2 Class distribution for samples in each split of training and test set in multiclass version

Dataset split	Classes	# Training samples	# Test samples
90/10%	Raw	2213	257
	Bicarbonate	419	41
	Formaldehyde	442	43
	Peroxide	417	48
	Starch	439	41
	Sucrose	431	55
75/25%	Raw	1846	624
	Bicarbonate	338	122
	Formaldehyde	347	138
	Peroxide	364	101
	Starch	359	121
	Sucrose	380	106
50/50%	Raw	1239	1231
	Bicarbonate	219	241
	Formaldehyde	223	262
	Peroxide	242	223
	Starch	253	227
	Sucrose	247	239

In binary version, the five classes of adulterant substances are summed up as one "adulterated" class

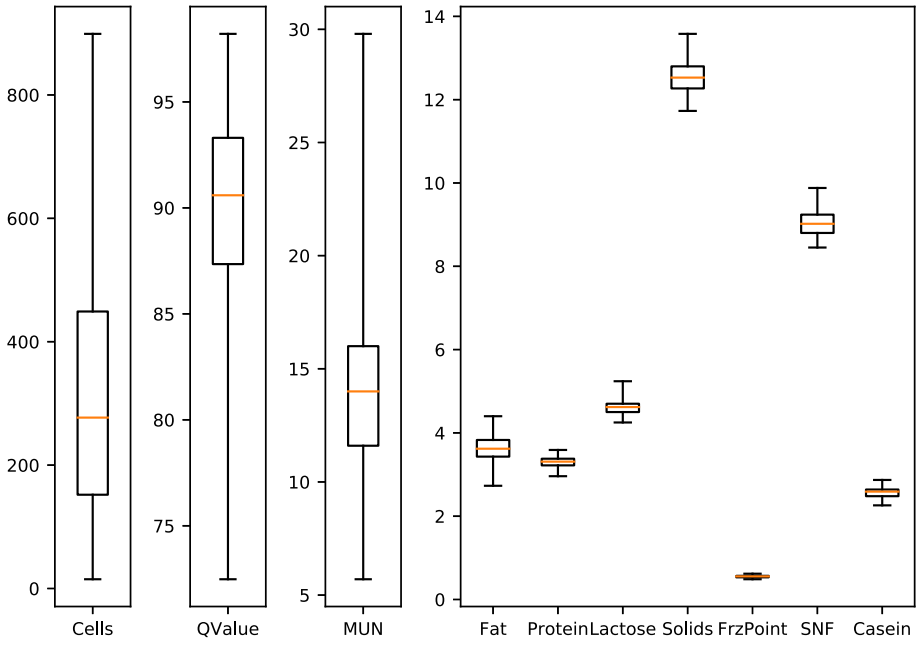


Fig. 2 Boxplot of the component features from the dataset, analyzed by ensemble learners. The plots show scales and variation of each feature. Cells, QValue and MUN have significantly different scales and were plotted separately from the other variables. The Cells plot show that most samples has somatic cells counting from ≈150 to ≈450. The QValue plot shows that most samples on the dataset were considered of high quality (≈87 to ≈93). The FrzPoint plot shows that all samples has a freezing point at just above zero. All the other plots consider specific component concentrations and they show low variation

set as 200 for each classifier. Models from both methods were evaluated for each available training and test sets using component features present in the samples. Binary and multiclass classifications were performed considering the same datasets.

Analysis of infrared spectra using deep learning

The infrared spectra used as input to the CNN classifier were produced by the FTIR technique. They are formed by 518 points measured in wavenumbers ranging from 3000 cm^{-1} to 1000 cm^{-1}. In the dataset, each spectrum is followed by the class label (adulteration substance), which allows the network to be trained. During the training process, the convolutional layers are used as filters that recognize specific features within spectral regions. For that reason, CNNs are able to receive raw spectral data as input, without the need of any preprocessing step, and they can handle important feature extraction from the data with no manual interaction [23].

We propose a CNN architecture that has one 1-dimensional convolutional layer that learns 32 filters of kernel size 5, which are capable of extracting features directly from the infrared spectra. Filters are concatenated and followed by one dense (fully connected) layer of 1024 neurons. At each layer, LeakyReLU [18] activation is used to add non-linearity to the model. Batch normalization [11] and dropout operations [25] are also performed at each layer so that the model avoids overfitting to the training data. The proposed network structure was based on the work from [17] but our structure is much simpler, with fewer layers and filters. In Fig. 3 we show the proposed CNN architecture.

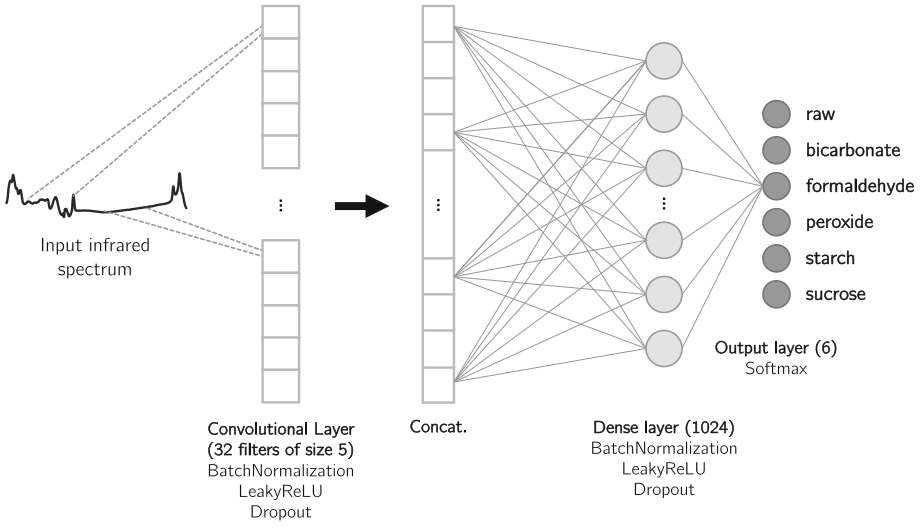


Fig. 3 The proposed Convolutional Neural Network for multiclass classification of whole infrared spectra. The architecture consists of one convolutional layer that learns 32 filters of kernel size 5, which is capable of recognizing features directly from the raw infrared spectra. The output of the convolutional layer is concatenated then passed as input to a dense (fully-connected) layer, consisting of 1024 neurons. BatchNormalization, LeakyReLU and Dropout operations are performed in both convolutional and dense layers. Finally, the output layer of size 6 (the number of classes in multiclass problem) is activated by the Softmax function

For binary and multiclass classifications, we trained a CNN that differed only at the number of neurons in the output layer. Since this layer outputs the classification, the number of neurons must be exactly the number of classes we want to classify our data. So, the CNN for the binary classification has an output layer of one neuron with binary output, activated by the sigmoid function and the CNN for the multiclass classification has an output layer of six neurons, activated by the softmax function [26]. The binary model classifies the samples with the presence or absence of an adulterant and the multiclass classification classifies samples as raw milk or one of five known adulterant substances.

The CNN training was made using Adam optimizer [13] for 100 epochs for both binary and multiclass problems. Every CNN execution considered 20% of the training set as the validation set. Figure 4 shows plots for the model's accuracy and loss of training and validation sets. The plots show that validation of the network achieved better results in the binary problem when compared to the multiclass problem, which is expected because the binary is considered a simpler problem. The CNN architecture was implemented in Keras [5] and TensorFlow [1] in Python. All CNN processing was made on a personal laptop computer. The model training takes up to 16 min, while the classification for all samples in the test dataset takes at most 270 ms.

Results and discussion

In order to determine that the chosen techniques and machine learning models were adequate for our experiments, we conducted a test that compared the performance of methods that are simpler and more commonly used in the dairy industry: Logistic Regression, Linear Regression and PLS [6, 7]. Classification versions of these methods were evaluated for each dataset split with whole spectra using the default implementations

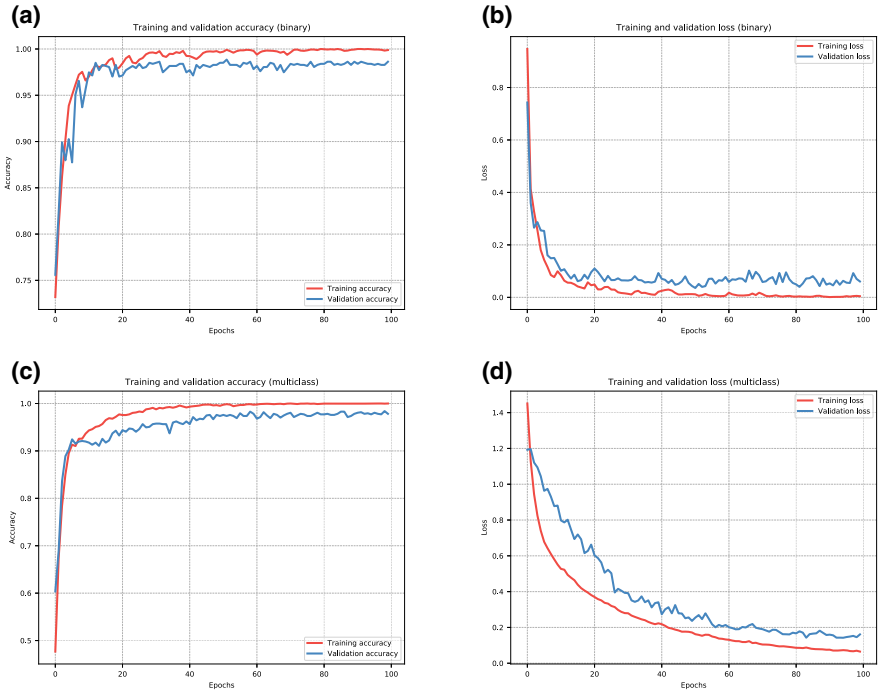


Fig. 4 Plot of the CNN model's accuracy and loss on training and validation steps considering the dataset split 80%/20%. The model was trained for 100 epochs. **a** Accuracy of training and validation considering the binary problem. **b** Loss of training and validation considering the binary problem. **c** Accuracy of training and validation considering the multiclass problem. **d** Loss of training and validation considering the multiclass problem. Each plotted curve is obtained from the history of the Keras model, which calculates both accuracy and loss for each epoch performed by the network. Accuracy is calculated by comparing the predicted class to the actual class. Loss is calculated by the cross entropy value between the predicted class and the actual class

available in Scikit-learn. Accuracies for Logistic Regression ranged from 55.92% to 58.76% in multiclass and from 71.40% to 76.49% in binary classification. For Linear Regression, accuracies ranged from 31.55% to 33.50% in multiclass and 79.20% to 79.62% in binary classification. Finally, for PLS, accuracies ranged from 32.56% to 35.26% in multiclass and 76.91% to 77.39% in the binary problem. Although all methods had relatively good performances in the binary problem, accuracies were not satisfactory in multiclass classifications. Therefore, these values serve as a comparative basis for our ensemble and deep learners. Table 3 shows all accuracy values from Linear Regression, Logistic Regression, and PLS models.

Table 3 Accuracy from simpler classifiers (Logistic Regression, Linear Regression, and Partial Least Squares) for binary and multiclass classifications that serve as baseline for our deep and ensemble learners

Dataset	Classification	Logistic Regression	Linear Regression	Partial Least Squares
90/10%	Multiclass	0.5876	0.3155	0.3526
	Binary	0.7649	0.7959	0.7691
75/25%	Multiclass	0.5693	0.3350	0.3267
	Binary	0.7583	0.7962	0.7739
50/50%	Multiclass	0.5592	0.3281	0.3256
	Binary	0.7140	0.7920	0.7714

All classifiers were evaluated with 3 pairs of training and test datasets randomly selected from our milk samples, identified by their proportion of training and test samples

Both ensemble and deep learners were evaluated for adulterant detection on milk samples. The dataset has 4846 samples labeled as one of six possible classes: raw, sucrose, starch, bicarbonate, peroxide, and formaldehyde. For the multiclass version of the problem, all six classes were used. For the binary version, classes for each adulterant were considered as one class: adulterant present, while class raw was considered as the second class. Binary and multiclass classifications were evaluated with the selected subsets of training and testing described earlier for GBM, RF, and CNN classifiers. For the ensemble methods, classification accuracies ranged from 86.09% to 98.56%. The proposed CNN produced accuracies up to 98.76%. The mean accuracies for RF, GBM and CNN were 93.23%, 92.25% and 96.76%, respectively. Accuracy values show that all classifiers have better performance on binary classifications. However, CNN has shown to be a more robust classifier, since it has very close accuracy levels with both binary and multiclass problems. All accuracy results from our models are shown in Table 4. We also detail the accuracy results per class in multiclass classifications for RF, GBM and CNN in Table 5. These values show that all classifiers have better performance on 'raw' classification, and that CNN has a best overall performance with every class. Values also show that increasing the training set (i.e., 90%) not always leads to better predictive performance in all classes. Finally, the CNN classifier is generally more robust when there is a decrease in training test size (50%).

The area under the ROC (Receiver Operating Characteristic) curve was evaluated for five repetitions in all classifiers, which yielded the AUC score. We then performed a pairwise t-test (t-value) comparing the difference in average AUC score across classifiers for binary and multiclass classifications. The greater the magnitude of t, the greater the evidence against the null hypothesis. This means there is greater evidence that there is a significant difference. The closer t is to 0, the more likely there isn't a significant difference. The larger the absolute value of the t-value, the smaller the p-value, and the greater the evidence against the null hypothesis. Statistical significance tests show that the CNN classifiers are more robust, having significant differences in performance when compared to ensemble ones, as shown in Fig. 5. The ROC curves are presented in Fig. 6, where it is shown that all ROC curves from binary classification (continuous lines) show good performance and predictive power, while the multiclass ROC curves (dotted lines) show that the CNN model has better predictive performance.

Intuitively, the binary classification tends to be a simpler problem and can lead to better results, which is observed on the RF and GBM results, where binary classifications accuracies are at most 10% higher than multiclass accuracies. However, the CNN results

Table 4 Accuracy from evaluated classifiers (RF, GBM, and CNN) for binary and multiclass classifications

Dataset	Classification	RF	GBM	CNN
90/10%	Multiclass	0.9093	0.8907	0.9608
	Binary	0.9856	0.9711	0.9794
75/25%	Multiclass	0.8812	0.8787	0.9695
	Binary	0.9744	0.9686	0.9876
50/50%	Multiclass	0.8700	0.8609	0.9538
	Binary	0.9736	0.9653	0.9546

All classifiers were evaluated with 3 pairs of training and test datasets randomly selected from our milk samples, identified by their proportion of training and test samples

Table 5 Accuracies for each individual class (bicarbonate, formaldehyde, peroxide, raw, starch, and sucrose) for multiclass classifications considering RF, GBM and CNN classifiers, in each of the selected training and test datasets: 90/10%, 75/25%, and 50/50%

Classifier	Dataset	Bicarbonate	Formaldehyde	Peroxide	Raw	Starch	Sucrose
RF	90/10%	0.7804	0.7209	0.7916	0.9883	0.8048	0.9636
	75/25%	0.7540	0.6884	0.7524	0.9839	0.8099	0.8773
	50/50%	0.7634	0.6259	0.7847	0.9805	0.7444	0.8953
GBM	90/10%	0.7804	0.7441	0.7291	0.9766	0.7560	0.9272
	75/25%	0.8032	0.6739	0.7623	0.9759	0.7768	0.8867
	50/50%	0.7551	0.6259	0.7309	0.9780	0.7356	0.8870
CNN	90/10%	0.9756	0.9302	0.8958	0.9844	0.9024	0.9636
	75/25%	0.9918	0.9057	0.9108	0.9887	0.9421	1.0000
	50/50%	0.9958	0.9236	0.8340	0.9861	0.8854	0.9539

show that the method is more robust on the multiclass classifications, with accuracies slightly lower than binary versions. We conclude that CNNs are particularly better suited for multiclass classification in this problem.

It is important to notice that the number of adulterated milk samples in our dataset is roughly half the total samples, which in terms of binary classification leads to balanced class distribution. On the other hand, when it comes to multiclass classification, we have six different classes and the majority of samples are of type raw, which leads to imbalanced class distribution. However, our method showed the capability to handle this situation without any issues, as shown in Table 4.

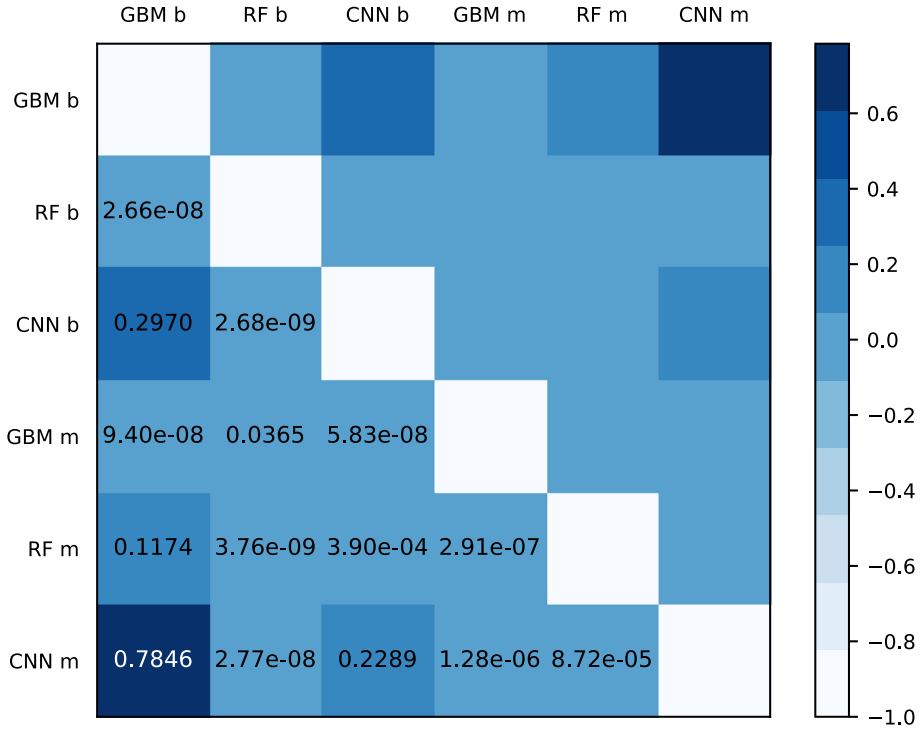


Fig. 5 A t-test over pairwise differences in average AUC score for binary (**b**) and multiclass (**m**) versions of GBM, RF and CNN classifiers

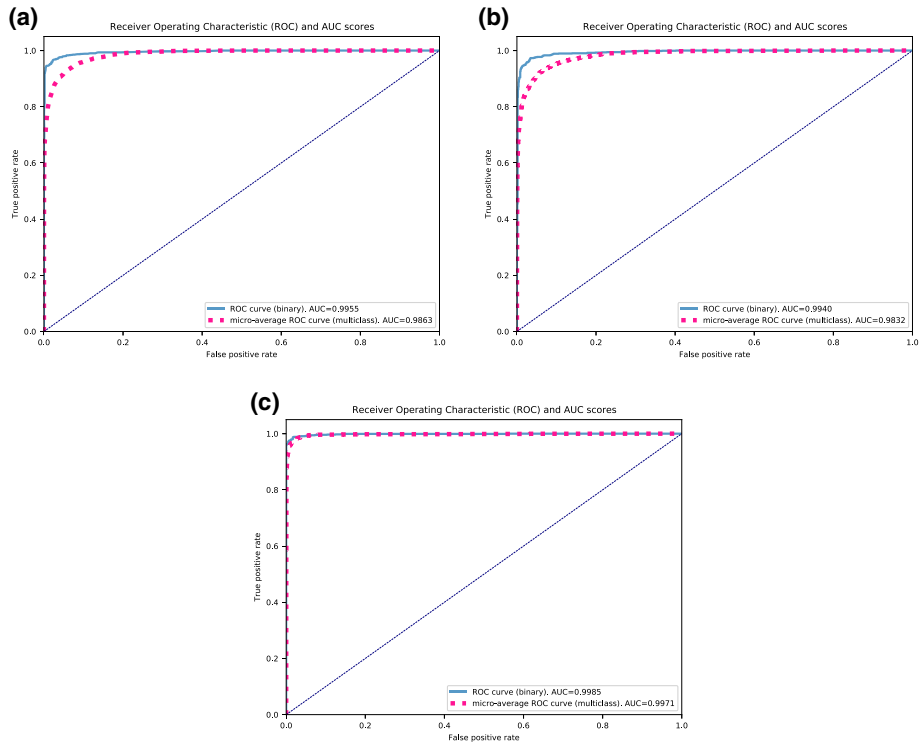


Fig. 6 ROC (Receiver Operating Characteristic) curves for binary and multiclass versions of (**a**) Random Forest, (**b**) Gradient Boosting Machine and (**c**) Convolutional Neural Network and their AUC (Area Under the ROC Curve) score. Multiclass ROC was calculated using micro-average strategy, that sums up the individual true positives, false positives, and false negatives for all classes

Conclusion

In this work, we investigated milk composition and performed adulterant detection on FTIR spectral data by classifying samples using deep and ensemble tree learners. We collected 4846 milk samples and manually adulterated 2376 samples, using different classifiers to train models that are capable of recognizing composition characteristics that adulterants cause in milk. The classification was performed using two types of data: the whole infrared spectra analyzed by CNN and the 10 component features extracted from the spectra analyzed by RF and GBM classifiers.

Both methods, whole infrared spectra analyzed by CNN and the ten component features extracted from the spectra analyzed by RF and GBM classifiers achieved high accuracy, however, the CNN obtained better results, which is intuitive since it uses a more dense dataset (spectral coordinates). In other words, the extraction of the components performed by the FTIR equipment is not as representative as the features recognized by the proposed CNN architecture. Classification accuracies range from 86.09% to 98.76%.

Nevertheless, some challenges remain as future work, like a more profound study on the models' interpretability, such as feature importance analysis and variable interactions. New analyses with multiple adulterations per sample and their effects on milk composition are also considered. Finally, we consider as an extension of this work a metaclassifier application, where the predictions of the deep and ensemble models could be combined, potentially achieving better performances.

Additional files

Additional file 1: Figure S1. On the utilization of deep and ensemble learning to detect milk adulteration. (PDF 56 kb)

Additional file 2: A dataset containing nearly 1000 readings from milk samples in the CSV format. FTIR component features and spectral data points are provided for each sample. (CSV 9075 kb)

Abbreviations

AUC: Area under the ROC curve; CNN: Convolutional neural network; ECG: Electrocardiogram; FTIR: Fourier transformed infrared; GBM: Gradient boosting machine; MLR: Multiple linear regression; MUN: Milk urea nitrogen; PCA: Principal component analysis; PLS: Partial least squares; RF: Random forest; ROC: Receiver operating characteristic; SNF: Solids non-fat

Acknowledgements

Not applicable.

Authors' contributions

HAN implemented the algorithms, performed the analysis with the classifiers and wrote the paper. WLFT, DCSZR and LMF collected milk samples and conducted the adulterations experiments. SVAC, LMF and RCOA supervised the experiments and assisted with results analysis and manuscript preparation. All authors read and approved the final version of this manuscript.

Funding

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Availability of data and materials

A sample of the dataset used for this study is included in the Additional file 1 accompanying this article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Federal Institute of Mato Grosso do Sul, Rua Ângelo Melão, 790, 79641-162 Três Lagoas, MS, Brazil. [2]Department of Computer Science, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, 31270-901 Belo Horizonte, MG, Brazil. [3]Veterinary School, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, 31270-901 Belo Horizonte, MG, Brazil. [4]Instituto Tecnológico Vale, R. Boaventura da Silva, 955, 66055-090 Belém, PA, Brazil. [5]Federal University of Pará R. Augusto Corrêa, 1, 66075-110 Belém, PA, Brazil.

Received: 14 March 2019 Accepted: 5 June 2019
Published online: 08 July 2019

References

1. Abadi M, et al. TensorFlow: A System for Large-scale Machine Learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. OSDI'16. Savannah: USENIX Association; 2016. p. 265–83.
2. Alves da Rocha R, Paiva IM, Anjos V, Furtado MAM, Valenzuela MJ. Quantification of whey in fluid milk using confocal raman microscopy and artificial neural network. J Dairy Sci. 2015;98(6):3559–67. https://doi.org/10.3168/jds.2014-8548.
3. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv. 2010;4:40–79. https://doi.org/10.1214/09-SS054.
4. Botelho BG, Reis N, Oliveira LS, Sena MM. Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. Food Chem. 2015;181:31–7. https://doi.org/10.1016/j.foodchem.2015.02.077.
5. Chollet F, et al. Keras. Microtome Publishing; 2015. Available at https://keras.io. Accessed 16 Aug 2018.
6. Cruz AG, Cadena RS, Faria JAF, Oliveira CAF, Cavalcanti RN, Bona E, Bolini HMA, Da Silva MAAP. Consumer acceptability and purchase intent of probiotic yoghurt with added glucose oxidase using sensometrics, artificial neural networks and logistic regression. Int J Dairy Technol. 2011;64(4):549–56. https://doi.org/10.1111/j.1471-0307.2011.00722.x.
7. de Carvalho BMA, de Carvalho LM, dos Reis Coimbra JS, Minim LA, de Souza Barcellos E, da Silva Júnior WF, Detmann E, de Carvalho GGP. Rapid detection of whey in milk powder samples by spectrophotometric and multivariate calibration. Food Chem. 2015;174:1–7. https://doi.org/10.1016/j.foodchem.2014.11.003.

8. Eisavi V, Homayouni S, Yazdi AM, Alimohammadi A. Land cover mapping based on random forest classification of multitemporal spectral and thermal images. Environ Monit Assess. 2015;187(5):1–14. https://doi.org/10.1007/s10661-015-4489-3.
9. Gondim CdS, Junqueira RG, de Souza SVC, Ruisánchez I, Callao MP. Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies. Food Chem. 2017;230:68–75. https://doi.org/10.1016/j.foodchem.2017.03.022.
10. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd. New York: Springer; 2017, p. 745.
11. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR; 2015. p. 448–56.
12. Kartheek M, Anton Smith A, Kottai Muthu A, Manavalan R. Determination of Adulterants in Food: A Review. J Chem Pharm Res. 2011;3(2):629–36.
13. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014. CoRR http://arxiv.org/abs/1412.6980. Accessed 2 July 2018.
14. Kuhn M, Johnson K. Applied Predictive Modeling, 1st. New York: Springer; 2013, p. 600. https://doi.org/10.1007/978-1-4614-6849-3.
15. Kumar A, Khadkevich M, Fugen C. Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes. 2018 IEEE Int Conf Acoust, Speech and Sig Process (ICASSP). 2018326–30. https://doi.org/10.1109/icassp.2017.7952132.
16. Li D, Zhang J, Zhang Q, Wei X. Classification of ECG signals based on 1D convolution neural network. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). Dalian: IEEE; 2017. p. 1–6. https://doi.org/10.1109/healthcom.2017.8210784.
17. Liu J, Osadchy M, Ashton L, Foster M, Solomon CJ, Gibson SJ. Deep convolutional neural networks for Raman spectrum recognition: A unified solution. Analyst. 2017;142(21):4067–74. https://doi.org/10.1039/c7an01371j.
18. Maas AL, Hannun AY, Ng AY. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In: Proceedings of the 30th International Conference on Machine Learning. Atlanta: Microtome Publishing; 2013.
19. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69. https://doi.org/10.1093/bib/bbw068.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
21. Polikar R. Ensemble based systems in decision making. IEEE Circ Syst Mag. 2006;6(3):21–45. https://doi.org/10.1109/MCAS.2006.1688199.
22. Santos PM, Pereira-Filho ER, Rodriguez-Saona LE. Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis. Food Chem. 2013;138(1):19–24. https://doi.org/10.1016/j.foodchem.2012.10.024.
23. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
24. Souza SS, Cruz AG, Walter EHM, Faria JAF, Celeghini RMS, Ferreira MMC, Granato D, de S. Sant'Ana A. Monitoring the authenticity of Brazilian UHT milk: A chemometric approach. Food Chem. 2011;124(2):692–5. https://doi.org/10.1016/j.foodchem.2010.06.074.
25. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929–58.
26. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014. Lecture Notes in Computer Science. Cham: Springer; 2014. p. 818–33.

Publisher's Note