UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA FUNDAMENTAL

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA EM PROBLEMAS FORENSE: Identificação de manchas de sangue humano e plantações de *Cannabis sativa* L.

Recife
2019

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA EM PROBLEMAS FORENSE: Identificação de manchas de sangue humano e plantações de *Cannabis sativa* L.

Tese de Doutorado apresentada ao Programa de Pós-graduação em Química da Universidade Federal de Pernambuco como requisito parcial à obtenção do Título de Doutor em Química

Área de concentração: Química Analítica

Orientadora: Maria Fernanda Pimentel
Co-orientador: Ricardo Saldanha Honorato
Supervisor no período sanduíche na Universidade de Copenhagen: Rasmus Bro

Recife
2019

Catalogação na fonte
Bibliotecária Arabelly Ascoli CRB4-2068

P436e Pereira, José Francielson Queiroz
Espectroscopia no infravermelho próximo e quimiometria em problemas forense: identificação de manchas de sangue humano e plantações de *Cannabis sativa L.* / José Francielson Queiroz Pereira. – 2019.
118 f.: fig., tab.

Orientadora: Maria Fernanda Pimentel
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN. Química. Recife, 2019.
Inclui referências e apêndices.

1. Química analítica. 2. *Cannabis sativa L.* 3. Sangue humano. 4. Modelo hierárquico. I. Pimentel, Maria Fernanda (orientadora). II. Título.

543 CDD (22. ed.) UFPE-CCEN 2020-21

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E QUIMIOMETRIA EM PROBLEMAS FORENSE: Identificação de manchas de sangue humano e plantações de *Cannabis sativa* L.

Tese de Doutorado apresentada ao Programa de Pós-graduação em Química da Universidade Federal de Pernambuco como requisito parcial à obtenção do Título de Doutor em Química

Aprovada em: 04 /11/2019

BANCA EXAMINADORA

Profª. Drª. Maria Fernanda Pimentel (Orientadora)
Universidade Federal de Pernambuco

Dr. Ricardo Saldanha Honorato (Co-orientador)
Polícia Federal

Profº. Dr. José Licarion Pinto Segundo Neto (Examinador Interno)
Universidade Federal de Pernambuco

Profº. Dr. Vagner Bezerra dos Santos (Examinador Interno)
Universidade Federal de Pernambuco

Profº. Dr. Ivo Milton Raimundo Júnior (Examinador Externo)
Universidade Estadual de Campinas

Profª. Drª. Andréa Monteiro Santana Silva Brito (Examinador Externo)
Universidade Federal Rural de Pernambuco

Recife
2019

A minha mãe Maria Aparecida pelo carinho e amor em cada gesto de dedicação infinita, pela compreensão e preocupação em todos os momentos difíceis.

Ao meu inesquecível pai João Inocêncio que me deu apoio e incentivo nessa jornada até seus últimos dias (*in memorian*).

Dedico.

AGRADECIMENTOS

Agradeço primeiramente a minha orientadora, Profª Drª Maria Fernanda Pimentel, que me transmite confiança, dedicação e comprometimento desde o início do meu mestrado. Em mais de seis anos de orientações e convivência a professora Fernanda ampliou meu horizonte acadêmico e pessoal. Com certeza impactou profundamente na minha vida do ponto de vista científico e de realizações pessoais. Sou muito grato por seus valorosos ensinamentos e sua dedicação inabalável.

Ao meu co-orientador o perito Dr. Ricardo Saldanha Honorato, meus sinceros agradecimentos pelos momentos de discussão de ideias e conselhos durante todas as etapas da pesquisa. Sempre levantou importantes questionamentos sobre a pesquisa e apontou diferentes soluções para cada problemática discutida.

Ao professor Rasmus Bro pela importante contribuição científica na pesquisa e pela acolhida junto ao grupo de pesquisas em quimiometria CAT da Universidade de Copenhague. Seus ensinamentos e contribuições foram essenciais para a conclusão da pesquisa.

Ao Dr. Everaldo Paulo de Medeiros pela disponibilidade da câmera de infravermelho essencial para a pesquisa.

A CNPq, pela bolsa concedida, ao INCTAA e ao NUQAAPE, pelo apoio ao projeto e a UFPE/DQF pelo suporte institucional.

Ao Laboratório de Combustíveis (LAC) e toda sua equipe que possibilitaram a realização da pesquisa dando suporte em todas as fases do trabalho.

Aos amigos do DQF que me proporcionaram muitos e bons momentos de discussão, estudo e descontração.

Aos amigos e colegas de trabalho Alianda, Eduardo, Jéssica, Lívia, Marcela, Rafaella, Maria Júlia, Thiago, Humberto, Vitor, Paulinha, Vanessa e muitos outros por todos os conselhos, ensinamentos e discussões que me ajudaram a vencer vários desafios. Minha pesquisa é, sem dúvidas, um reflexo de todos que contribuíram de alguma forma para minha pesquisa ou para minha

formação humana. Sou eternamente grato por todo o apoio que recebi nos momentos nos quais mais precisei.

Um agradecimento especial a Carolina por ter se prontificado a me ajudar em inúmeras situações, pelo treinamento técnico dado, por sua disponibilidade em tirar minhas dúvidas e por todo o suporte emocional que foi tão essencial.

Ao meu grande amigo Neirivaldo por ter sido um apoio em todos os momentos em que precisei.

Aos colegas do LAC, João, Vanessa, Gisele, Douglas, André e Cleciane que me ajudaram a resolver incontáveis questões de trabalho e sempre estiveram dispostos a ajudar.

Aos bons amigos Rodrigo, Alessandro, Bingru, Luciana, Viola e Pranse que estiveram comigo durante um ano de intercâmbio.

A todos da minha família pelo incentivo em continuar trilhando o caminho acadêmico e buscando as minhas realizações pessoais e profissionais.

RESUMO

Este trabalho aborda duas problemáticas de interesse forense, ambas estudadas empregando a espectroscopia na região do infravermelho próximo e técnicas quimiométricas de classificação. A primeira trata da identificação de *Cannabis sativa* Linneaus na presença de outras plantas típicas da região onde estão localizadas plantações ilegais no Estado de Pernambuco. Imagens hiperespectrais na região do infravermelho próximo (HSI-NIR) foram obtidas de amostras simulando plantações de *Cannabis sativa* L. contendo plantas similares, juntamente com solo. Para seleção de variáveis espectrais foi inicialmente empregada análise de componentes principais com *sparce* (sPCA*)*. A técnica de classificação modelagem independente e flexível de analogia de classe (SIMCA) foi então utilizada, empregando apenas quatro variáveis selecionadas na etapa anterior. *Sparse*-SIMCA demonstrou alta especificidade e sensibilidade (0,922 e 0,902, respectivamente) para identificação inequívoca de *Cannabis*, além de ser capaz de identificar as folhas de *Cannabis* nas imagens de forma clara e objetiva. O segundo problema de interesse forense abordado foi a identificação de manchas de sangue depositadas em substratos comuns em possíveis cenas de crime. Em um primeiro estudo foi desenvolvida uma metodologia não destrutiva e confirmatória para identificação de manchas de sangue humano (SH) usando um espectrômetro NIR portátil. Manchas de SH, sangue animal (SA) e de falso-positivos comuns (FPC) foram depositadas em pisos cerâmicos, porcelanatos, vidro e metal e os espectros obtidos com o espectrômetro portátil MicroNIR (Viavi), após três dias de secagem. Foram criados modelos individuais para cada substrato utilizando três técnicas de classificação/seleção de variáveis: análise discriminante linear - algoritmo genético (GA-LDA), LDA - Algoritmo de Projeções Sucessivas (SPA-LDA) e análise discriminante por mínimos quadrados parciais (PLS-DA). GA-LDA e PLS-DA apresentaram os melhores resultados para classificação de um conjunto de validação externa composto por SH, SA e FPC, obtendo especificidade e sensibilidade iguais a 1 em todos os casos. Análises combinando pisos dois a dois foram conduzidas seguindo a metodologia anterior, porém nesse caso foi implementado o pré-processamento dos mínimos quadrados ponderados generalizados para minimizar a influência dos diferentes substratos. Os valores de especificidade e sensibilidade para a previsão em todos os modelos ficaram acima de 0,980 para o sangue humano e sangue animal. O segundo estudo descreve uma metodologia para identificação de manchas de SH em tecidos utilizando HSI-NIR e modelos de classificação hierárquica. Manchas de SH, SA e FPC foram preparadas sobre tecidos coloridos e estampados de base sintética e de algodão. Uma parte das amostras foi utilizada para construir modelos hierárquicos formulado pela fusão de modelos de PCA e PLS-DA, para os tecidos de composições distintas. As imagens de todas as amostras foram submetidas aos modelos e foi realizada a identificação visual de todas amostras de SH e SA. Obteve-se uma classificação correta de 95% das amostras de SH, sendo que mesmo as amostras de SH não classificadas corretamente foram identificadas como sangue. Assim, o modelo minimiza a possibilidade de negligenciar evidências reais.

Palavras-chaves: *Cannabis sativa* L. Sangue humano. SIMCA. Classificação. Modelo hierárquico.

ABSTRACT

This work addresses two issues of forensic interest, both evaluated employing near infrared spectroscopy and chemometric classification techniques. The first work deals with the identification of *Cannabis sativa* Linneaus in the presence of other typical plants of the region where illegal plantations are located in the state of Pernambuco. Near Infrared Hyperspectral Images (HSI-NIR) were obtained from samples prepared by reproducing *Cannabis sativa* L. plantations containing similar plants and soil. Sparse principal component analysis (sPCA) was initially used to select spectral variables. The Soft Independent Modeling Class Analogy (SIMCA) classification technique was then used, employing only four variables selected in the previous step. Sparse-SIMCA demonstrated high specificity and sensitivity (0.922 and 0.902, respectively) for unambiguous identification of *Cannabis*, as well as being able to identify *Cannabis* leaves in images clearly and objectively. The second problem of forensic interest addressed was the identification of bloodstains deposited on common substrates in potential crime scenes. In a first study, a non-destructive and confirmatory methodology for identifying human bloodstains (HB) using a portable NIR spectrometer was developed. HB, animal blood (AB) and common false positive (CFP) stains were deposited on ceramic, porcelain, glass and metal slabs and the spectra was obtained with the MicroNIR (Viavi) portable spectrometer after three days of drying. Individual models were created for each substrate using three classification / variables selection techniques: Linear Discriminant Analysis - Genetic Algorithm (GA-LDA), Successive Projections Algorithm (SPA-LDA) and Partial Least Squares Discriminant Analysis (PLS-DA). GA-LDA and PLS-DA presented the best results for the classification of an external validation set composed by HB, AB and CFP, obtaining specificity and sensitivity equal to 1 in all cases. Analyzes combining two-by-two slab types were conducted following similar methodology, but in this case the Generalized Weighted Least Squares (GLSW) preprocessing was implemented to minimize the influence of the different substrates. Prediction specificity and sensitivity values in all models were above 0.980 for human and animal blood. The second study describes a methodology for identifying HB stains in fabrics using HSI-NIR and hierarchical classification models. HB, AB and CFP stains were prepared on colored and printed fabrics with synthetic and cotton base. Part of the samples were used to construct hierarchical models formulated by the fusion of PCA and PLS-DA models for samples prepared on different fabric types. The images of all samples were submitted to the models and a visual identification of all samples of HB and AB succeeded correctly. 95% of the HB samples were correctly classified, and even HB samples not correctly classified were identified as blood. Thus, the model minimizes the possibility of neglecting actual evidence.

Keywords: *Cannabis sativa* L Human blood. SIMCA. Classification. Hierarchical model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BA	Batom
BEC	Tecido bege de algodão
BLC	Tecido preto de algodão
BrS	Tecido branco sintético
DQF	Departamento de Química Fundamental
FIR	Infravermelho Distante (*Far Infrared*)
FN	Falso Negativo
FPC	Falso-Positivo Comum
FT-IR	Infravermelho por Transformada de Fourier (*Fourier Transform Infrared*)
GA	Algoritmo Genético (*Genetic Algorithm*)
GC-MS	Cromatografia Gasosa hifenada a Espectroscopia de Massas (*Gas Chromatography coupled to Mass spectroscopy*)
GE	Geleia
GLSW	Mínimos Quadrados Parciais Ponderados Generalizados (*Generalized Least Squared Weighted*)
HPCL	Cromatografia Líquida de Alta Eficiência (*High Performance Liquid Chromatogra*phy)
HSI	Imagens Hiperespectrais (*Hyperspectral Images*)
HSI-NIR	Imagens Hieprespectrais no Infravermelho Próximo (*Hyperspectral Images on Near Infrared*)
IMS	Espectroscopia de Massas por Mobilidade Iônica (*Ion Mobility Mass Spectrometry*)
INCTAA	Instituto Nacional de Ciência e Tecnologia de Tecnologias Analíticas Avançadas
IR	Infravermelho (*Infrared*)
KE	Ketchup
LDA	Análise Discriminante Linear (*Linear Discriminant Analysis*)
LV	Variáveis Latentes (*Latent Variables*)
NIR	Infravermelho Próximo (*Near Infrared*)
NMR	Ressonância Magnética Nuclear (*Nuclear Magnetic Resonance Spectroscopy*)
NUQAAPE	Núcleo de Química Avançada de Pernambuco
MIR	Infravermelho Médio (*Middle Infrared*)
MSC	Correção de Sinal Multiplicativo (*Multiplicative Signal Correction*)
PF	Polícia Federal
PC	Componentes Principais (*Principal Component*)
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
PI	Pimenta
PLS	Mínimos Quadrados Parciais (*Partial Least Squares*)
PLS-DA	Análise Discriminante por Mínimos Quadrados Parciais (*Partial Least Squares – Discriminant Analysis*)

PRC	Tecido estampado de algodão
PRS	Tecido estampado sintético
RGB	*Red, Blue, Green*
RMSEP	Erro Quadrático Médio para Previsão (*Root Mean Squared of Prediction*)
RoI	Região de Interesse (Region of Interest)
SA	Sangue Animal
SG	Savitzky-Golay
SH	Sangue Humano
SIMCA	Modelagem Independente e Flexível de Analogia de Classes (*Soft Independent Modeling Class Analysis*)
Sn	Sensibilidade
SNV	Variação Normal Padrão (*Standart Normal Variate*)
Sp	Especificidade
SPA	Algoritmo de Projeções Sucessivas (*Successive Projection Algorithm*)
sPCA	Análise de Componentes Principais com *Sparse* (*Sparse Principal Component Analysis*)
SVM-DA	Análise Discriminante por Máquinas de Vetores de Suporte (*Support Vector Machine*s - *Discriminant Analysis*)
SY	Shoyu
THC	Tetra-hidro-cannabinol (Tetrahydrocannabinol)
UFPE	Universidade Federal de Pernambuco
UFRPE	Universidade Federal Rural de Pernambuco
VIP	Importância das Variáveis para a Projeção (*Variable Importance in Projection*)
VG	Vinagre balsâmico
VI	Vinho tinto
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 APRESENTAÇÃO

As análises de evidências em cenas de crime e em materiais suspeitos são essenciais para o andamento de investigações e resolução dos crimes. Os métodos forenses empregados para análise de sangue são bastante variados, sendo os testes químicos e análise do DNA duas abordagens conhecidas e muito comuns em casos (JAMES; KISH; SUTTON, 2005). Os trabalhos que visam desenvolver novas metodologias para análise de evidências em geral focam na praticidade dos métodos, pois a dinâmica das investigações torna difícil a especialização dos peritos nas diferentes técnicas e metodologias. Entretanto, muitos dos métodos utilizados são pouco específicos e/ou tem alto custo, especialmente se tratando de amostras de fluidos corporais. Adicionalmente, muitas das metodologias utilizadas em perícias criminais são destrutivas e podem comprometer a integridade das amostras (VIRKLER; LEDNEV, 2009a).

A Cannabis sativa L., popularmente conhecida como maconha, é um narcótico bastante comum no Brasil. Uma das atribuições da Polícia Federal (PF) é encontrar e erradicar plantações ilegais de marijuana em todo o território nacional. A metodologia atualmente utilizada pela PF é baseada na inspeção visual de grandes áreas verdes feita por peritos experientes a bordo de helicópteros, com posterior utilização de drones equipados com câmeras digitais para obter imagens mais precisas. A identificação das plantações é relativamente demorada e subjetiva, dependendo muito da experiência dos peritos

As técnicas espectroscópicas podem ser uma alternativa no desenvolvimento de metodologias forenses, pois são técnicas rápidas, confiáveis e relativamente baratas. Os métodos baseados na espectroscopia de infravermelho são bastante utilizados em análises forenses no laboratório, porém ainda não são muito explorados para análises em locais de crime (SKOOG; HOLLER; NIEMAN, 2009). Os recentes avanços nas tecnologias de espectroscopia no infravermelho possibilitaram a construção de equipamentos de ultra-portáteis e de mapeamento químico para construção de imagens hiperespectrais, ideais para análise de evidências em cenas de crime e de materiais suspeitos (MARQUES *et al.*, 2016; MOBARAKI; AMIGO, 2018). Outra vantagem desses métodos é sua objetividade quando são combinados com métodos matemáticos e estatísticos de análise de dados químicos, provenientes da área de estudos conhecida como Quimiometria (WOLD, 1995).

Os métodos quimiométricos são indispensáveis para auxiliar a análise dos dados complexos, como os produzidos por espectrômetros e câmaras hiperespectrais, visando extrair informação útil. Essa combinação de técnicas de infravermelho e quimiométricas viabilizam análises rápidas, não-destrutivas, confiáveis e objetivas diretamente em materiais encontrados em locais de crime ou até mesmo em campo.

Diante do exposto, este trabalho buscou desenvolver metodologias para identificação de sangue humano em possíveis cenas de crime e uma metodologia para classificação de *Cannabis sativa* L na presença de outras plantas típicas da região onde estão localizadas plantações ilegais no Estado de Pernambuco, empregando a espectroscopia na região do infravermelho próximo. O texto foi dividido em três partes para facilitar a compreensão e separar as problemáticas estudadas devido às suas dissimilaridades.

Na primeira parte é apresentada a fundamentação teórica com relação à espectroscopia na região do infravermelho próximo e técnicas quimiométricas.

Na segunda parte estão descritas as metodologias desenvolvidas para identificação de manchas de sangue humano depositadas em (i) materiais comumente encontrados em cenas de crime (vidro e metal) e em pisos, utilizando um equipamento ultra portátil; e (ii) uma metodologia para identificação de manchas de sangue humano em tecidos coloridos utilizando uma câmera hiperespectral. Ambas as metodologias utilizam técnicas quimiométricas de reconhecimento de padrão não supervisionadas e supervisionadas para diferenciar as amostras de sangue humano de diferentes falso-positivos comuns e sangue de animal.

A terceira parte trata de uma metodologia para identificação e discriminação de plantações de *Cannabis sativa* L. de outras plantas similares utilizando imagens hiperespectrais e técnicas quimiométricas. Para seleção de variáveis espectrais foi inicialmente empregada análise de componentes principais com *sparse* (sPCA do inglês *Sparse Principal Component Analysis*) e para classificação foi empregada a modelagem independente e flexível de analogia de classe (SIMCA do inglês *Soft Independent Modeling Class Analysis*).

2 INTRODUÇÃO

2.1 INFRAVERMELHO

Os espectros de absorção, emissão e reflexão na região do infravermelho estão diretamente relacionados às transições de uma molécula de um estado vibracional ou rotacional para outro. A radiação no infravermelho está compreendida na região espectral que se inicia no comprimento de onda 780 nm e se estende até o comprimento de onda 10^6 nm do espectro eletromagnético da luz. A região correspondente ao infravermelho podem ser dividida em Infravermelho Próximo (NIR do inglês *near infrared*), Infravermelho Médio (MIR do inglês *middle infrared*) e Infravermelho Distante (FIR do inglês *Far Infrared*)(SKOOG; HOLLER; NIEMAN, 2009), uma vez que os fenômenos envolvidos, os instrumentos e as aplicações variam consideravelmente de acordo com a faixa do espectro de infravermelho utilizada. A separação da região do infravermelho nas três faixas específicas pode ser observada na Tabela1

Tabela 1 – Regiões do infravermelho.

Região	Número de onda (cm^{-1})	Comprimento de onda (nm)	Frequência (Hz)
NIR	12.800-4.000	780-2.500	$3{,}8\ 10^{14}$- $1{,}2\ 10^{14}$
MIR	4.000-200	2.500-50.000	$1{,}2\ 10^{14}$- $6{,}0\ 10^{12}$
FIR	200-10	50.000-1.000.000	$6{,}0\ 10^{12}$- $3{,}0\ 10^{11}$

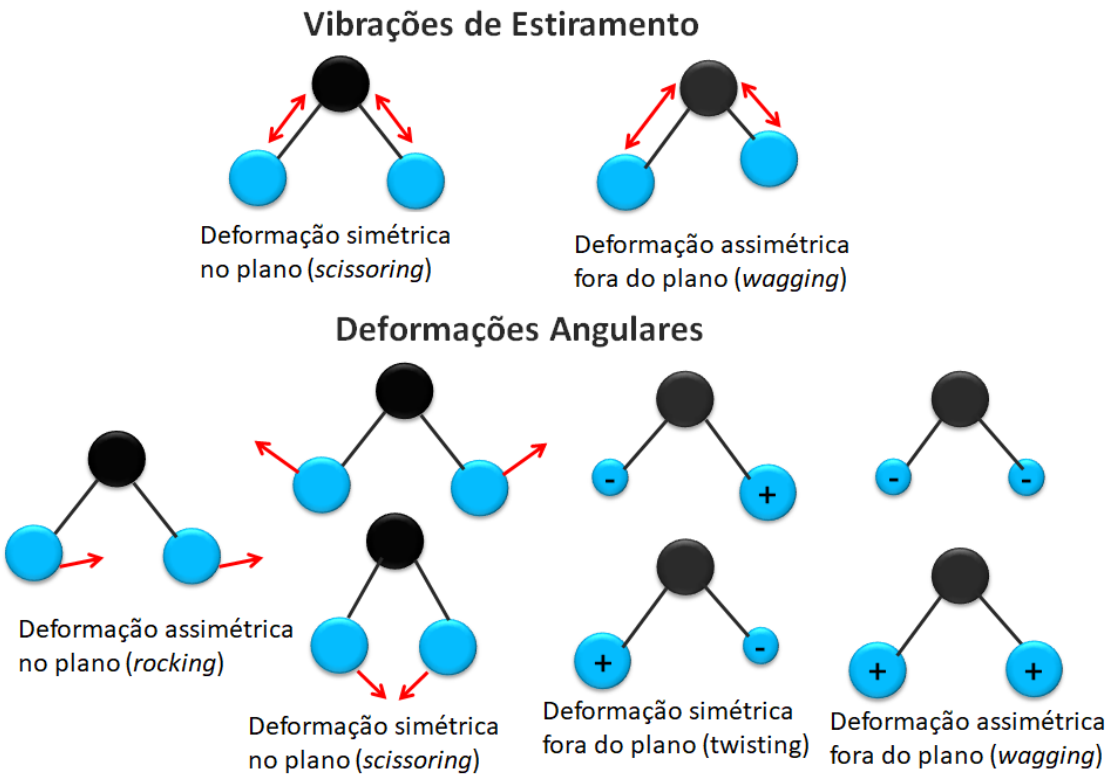
Fonte: Adaptado de Skoog *et al.* (2009)

Os sistemas moleculares absorvem radiação infravermelha em frequências (energias) discretas e depois a convertem em energias vibracionais e rotacionais. No processo de absorção, a radiação infravermelha que se iguala a frequência de vibração natural da molécula é absorvida, alterando a amplitude da vibração e, portanto, a energia do sistema molecular. Quando o sistema retorna ao estado fundamental, a energia absorvida é liberada em pacotes com energias equivalentes às transições entre níveis vibracionais subsequentes. No entanto, para que ocorra absorção da radiação infravermelha por uma molécula, deve haver uma mudança no seu momento dipolo decorrente das vibrações, gerando então um campo elétrico capaz de interagir com campo da radiação incidente (PAVIA *et al.*, 2009; SILVERSTEIN; WEBSTER; KIEMLE, 2005). Dessa

forma, moléculas diatômicas homonucleares, como O_2, N_2, Cl_2, H_2, etc., não absorvem na região do infravermelho (SKOOG; HOLLER; NIEMAN, 2009).

As vibrações moleculares são divididas em estiramento de ligação e deformações angulares. Uma vibração de estiramento é caracterizada pela variação da distância entre átomos ou entre grupos de átomos ligados, tomando o eixo de uma ligação como referência para deslocamento, podendo ser um estiramento simétrico ou assimétrico. A vibração de deformação angular por sua vez, se refere a variação do ângulo formado entre duas ligações adjacentes em um mesmo átomo, podendo ser uma deformação no plano de ligação ou fora dele (SKOOG; HOLLER; NIEMAN, 2009). Como um exemplo para melhor compreender os diferentes modos vibracionais, a Figura 1 representa um esquema para uma molécula triatômica angular.

Figura 1 – Estiramentos e deformações moleculares ativas no IR.



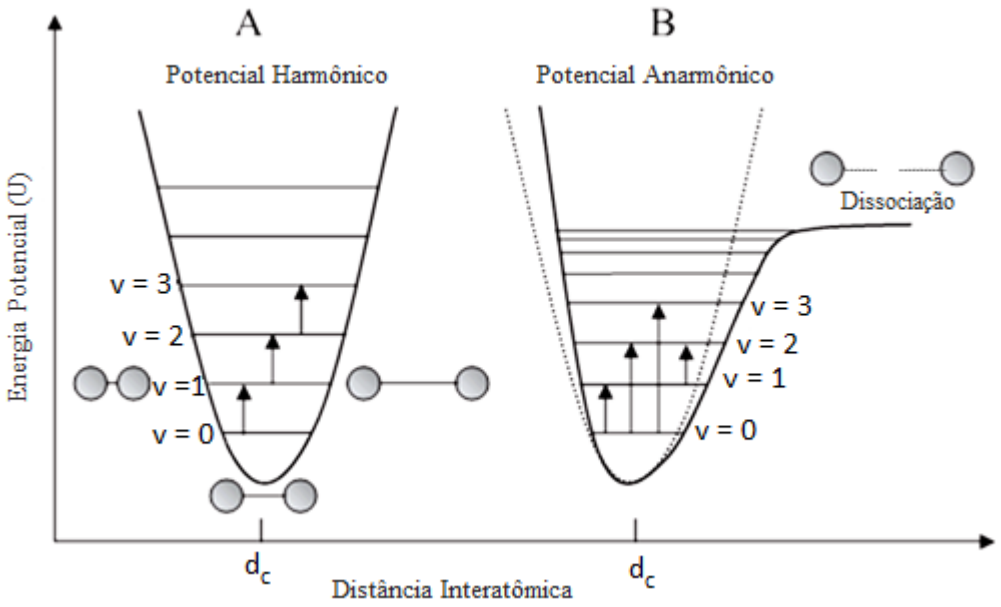
Fonte: Adaptado de Skoog *et al.* (2009)

Uma maneira de interpretar as vibrações das moléculas é considerar que esse sistema se assemelha a um oscilador harmônico, em que duas massas (átomos) estão ligadas por uma mola

(ligação covalente) que parte de um repouso e vibra quando uma fonte externa de energia provoca o deslocamento dos grupos/átomos ligados (PASQUINI, 2003). Contudo, esse modelo não é ideal para simular sistemas moleculares uma vez que esses sistemas não permitem que haja transições entre os níveis energéticos das moléculas. O modelo do oscilador harmônico considera apenas que as vibrações ocorrem entre dois níveis energéticos adjacentes ($\Delta\nu = \pm 1$) e que a energia entre os níveis é sempre igual independentemente do nível energético inicial, como ilustrado pela Figura 2.a. Adicionalmente, o modelo do oscilador harmônico não permite haver combinações de vibrações fundamentais e os sobretons de uma vibração ($\Delta\nu = 2$ ou maior)(PASQUINI, 2003).

O modelo do oscilador anarmônico é mais adequado para descrever as vibrações em uma molécula, uma vez que esse modelo considera os efeitos da repulsão coulombiana ao aproximar os átomos e a possibilidade de dissociação da ligação em condições extremas de afastamentos dos átomos. A repulsão coulombiana ocorre como consequência da aproximação das nuvens eletrônicas dos átomos ou grupos de átomos, que força os átomos a retornarem ao estado inicial. Já a dissociação da ligação pode ocorrer em uma situação em que a força restauradora da ligação está reduzida devido ao aumento da distância entre os átomos ligados. O modelo do oscilador anarmônico permite que haja transições fundamentais ($\Delta\nu = \pm 1$) e transições secundárias com $\Delta\nu = \pm 2$ ou 3, com valores discretos de energia. Essas transições secundárias explicam os sobretons de vibrações moleculares, e também a formação de bandas de combinação de vibrações diferentes (PASQUINI, 2003; SKOOG; HOLLER; NIEMAN, 2009). A Figura 2.b mostra um diagrama simples que permite entender as transições em um modelo de oscilador anarmônico em função da energia potencial.

Figura 2 – Diagrama de energia potencial para osciladores (A) harmônicos e (B) anarmônicos.



Fonte: Adaptado de Pasquini (2003).

As transições fundamentais geralmente apresentam sinais intensos e bem resolvidos na faixa do MIR. Na região NIR se encontram os sobretons e bandas de combinação das vibrações fundamentais das ligações O – H, N – H, C – H e S – H (SKOOG; HOLLER; NIEMAN, 2009). Essas ligações possibilitam a formação de dipolos intensos e, consequentemente, apresentam absorção da radiação (PASQUINI, 2003). Os instrumentos que operam na faixa NIR apresentam menor complexidade, são mais robustos e podem ser mais facilmente miniaturizados que os instrumentos MIR. Os componentes óticos utilizados em espectrômetros NIR são mais simples que os componentes necessários para os equipamentos MIR, além de serem mais resistentes às variações das condições ambientais como a umidade (PASQUINI, 2003). A facilidade para desenvolver componentes para equipamentos NIR impulsionou o desenvolvimento de instrumentos com aplicações mais específicas (SKOOG; HOLLER; NIEMAN, 2009).

2.2 TÉCNICAS QUIMIOMÉTRICAS

Os constantes avanços tecnológicos na instrumentação analítica implicaram também na necessidade de técnicas mais eficazes para a interpretação dos dados obtidos. Muitas vezes os

dados são complexos e difíceis de interpretar sem o auxílio de técnicas matemáticas adequadas que possam extrair as informações mais relevantes. Esse é o caso dos espectros NIR que apresentam grande sobreposição de bandas e correlações entre as variáveis espectrais, o que dificulta a interpretação visual ou univariada dos dados (BRERETON, 2003b; PASQUINI, 2018). O desenvolvimento dos métodos quimiométricos tem acompanhado a necessidade de interpretação dos dados multivariados de diferentes formas, seja através da exploração dos padrões ou a construção de modelos capazes de descrever o comportamento estatístico dos mesmos (BRO; SMILDE, 2014; WOLD, 1995).

2.2.1 A análise de Componentes Principais (PCA)

A análise de componentes principais é uma técnica exploratória de dados muito utilizada para identificar padrões de agrupamentos de amostras em função da máxima variância dos conjuntos de dados. A PCA é capaz de identificar os padrões de distribuição de amostras, que podem ter uma relação com a composição química das mesmas, e definir as varáveis mais importantes para descrever os padrões identificados. Um modelo PCA pode ser entendido como sendo resultado da combinação linear das variáveis de uma matriz de dados para produzir uma nova matriz com valores singulares capaz de representar as variáveis originais em um novo espaço de projeção (BRO; SMILDE, 2014). Considerando uma matriz de dados **X** com M linhas (amostras/objetos; $m = 1, 2, 3, \dots, M$) e N colunas (variáveis; $n = 1, 2, 3, \dots, \mathrm{N}$), cada elemento da matriz é descrito como x_{mn}. A combinação linear das variáveis pode, então, ser descrita pela expressão $\mathbf{t} = \boldsymbol{p_1}\, \boldsymbol{x_1} + \cdots + \boldsymbol{p_n}\, \boldsymbol{x_n}$, onde cada elemento de $\mathbf{x_n}$ é ponderado pelo coeficiente de pesos $\mathbf{p_n}$. Assim, uma matriz **T** formada pelos vetores **t** carrega parte da variância de **X** e essa quantidade de variância contida em **T** depende do número de elementos **t** utilizados para compor a matriz **T.** Como mostrado na equação 1, em que *l*= 1, ..., L representa o número de componentes ($\mathrm{PC}_{1\dots\mathrm{L}}$), a matriz **X** pode ser representada pela soma da variância capturada pelos L elementos **t** ponderados pelo vetor de pesos transposto **p**. A equação 2 traz a notação matricial resultante para L componentes:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \cdots + \mathbf{t}_\mathrm{L}\mathbf{p}_\mathrm{L}^T + \mathbf{E} \quad \text{Eq. 1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{Eq. 2}$$

Onde $\mathbf{P}^T = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_l]$(N *x* L) é a matriz de loadings transposta, $\mathbf{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l]$(M *x* L) é a matriz de escores e **E** é a matriz dos resíduos deixados pelo modelo. As implicações desta solução são que: a matriz de loadings contêm vetores ortogonais entre si e possuem norma igual a 1; as componentes dos escores são ortogonais entre si (BRO; SMILDE, 2014). O processo matemático completo pode ser encontrado no trabalho de Bro e Smilde (BRO; SMILDE, 2014) e no livro *Compreensive Chemometrics: chemical and biochemical Analysis* (BROWN; WALCZAK; TAULER, 2009).

As técnicas de análise exploratória são bastante úteis para compreensão dos conjuntos de dados e identificação de padrões. Contudo, quando o objetivo é construir um modelo para classificação de amostras futuras é comum a utilização de técnicas de reconhecimento de padrão supervisionadas. Na área forense, é bastante comum encontrar trabalhos que envolvem técnicas quimiométricas de classificação ou reconhecimento de padrão supervisionadas. As técnicas mais utilizadas para construir modelos de classificação se baseiam na modelagem de classes e na formulação de funções discriminantes.

2.2.2 Modelagem Independente e Flexível de Analogia de Classes (SIMCA)

A técnica de modelagem de classes mais conhecida é Modelagem Independente e Flexível de Analogia de Classes (SIMCA do inglês *Soft Independent Modeling Class Analysis*) na qual uma fronteira é construída para um determinado conjunto de amostras similares entre si. A fronteira é construída com as informações das amostras que pertencem ao grupo ou classe de interesse (BEEBE; PELL; SEASHOLS, 1998). Essa técnica é particularmente interessante quando as amostras da classe são homogêneas e as amostras não pertencentes à classe não são homogêneas. No entanto, uma possível desvantagem é que a modelagem SIMCA não busca a construção de um modelo que maximiza a diferença entre as diferentes classes, como no caso das análises discriminantes.

2.2.3 Análise Discriminante

As abordagens de classificação direta por meio da análise discriminante linear (LDA do inglês *Linear Discriminante Análises*) e análise discriminante por mínimos quadrados parciais (PLS-DA *Partial* do inglês *Least Squared Discriminant Analysis*), por sua vez, buscam construir limites bem definidos entre as duas ou mais classes pertencentes ao modelo.

PLS-DA é uma ferramenta quimiométrica bastante utilizada para diversas aplicações na química analítica e nas ciências forenses. As aplicações forense do PLS-DA em dados de infravermelho têm ocorrido com maior frequência devido a sua disponibilidade em diferentes softwares comerciais, como pode ser observado nos trabalhos de Mistek & Lednev (2015), Zhang *et al.* (2016), Zapata *et al.* (2015) e Grobério *et al.* (2015). Esta ferramenta partilha da mesma base conceitual da PCA, no entanto, para PLS-DA o comportamento de um conjunto de espectros **X** (variância) é determinado em função de uma ou mais propriedades **y.** O objetivo da técnica de mínimos quadrados parciais é construir um modelo capaz de descrever a variância interna do conjunto de amostras **X** e a do vetor **y**, empregando a covariância entre as duas matrizes (BRERETON; LLOYD, 2014; KJELDAHL; BRO, 2010; WOLD; SJOSTROM, 2001). Na PLS-DA o vetor **y** é um conjunto de números discretos, normalmente cada classe é relacionada a um índice numérico inteiro (Exemplo, -1 e 1, 0 e 1, 1 e 2; para duas classes) correspondentes às classes ou grupos aos quais as amostras presentes na matriz **X** pertencem. Isso implica que a modelagem das classes depende da representatividade individual de cada grupo (BRERETON; LLOYD, 2014). A matriz **X** e o vetor **y** podem ser decompostos em *scores* e *loadings* como feito pela PCA, e assim serem representados pelas equações 3 e 4:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\boldsymbol{T}} + \mathbf{E} \qquad \text{Eq. 3}$$

$$\mathbf{y} = \mathbf{U}\mathbf{q}^{\boldsymbol{T}} + \mathbf{f} \qquad \text{Eq. 4}$$

Em que **X** é a matriz dos espectros, **y** é um vetor de classes que se quer prever, **T e U** são as matrizes dos *scores* para os dois conjuntos de dados (**X** e **y**), **P** e **q** são matriz/vetor transpostos dos *loadings* e **E** e **f** matriz/vetor residuais. As sucessivas colunas da matriz de *scores* **T** (componentes do PLS) são ortogonais entre si, mas as linhas dos *loadings* **P** não são. Dessa forma,

TP nem sempre representam o máximo de variância em **X**, sendo maximizada a covariância entre **X** e **Y** (BRERETON; LLOYD, 2014). Considerando os erros **E** e **f** equivalentes, a relação de proporcionalidade entre os *scores* de **X** e **Y** pode ser expressa utilizando um vetor de pesos ***w*** para cada componente. A correlação entre **X** e **Y** pode ser expressa como a correlação entre os seus *scores* **U = w.T.** Considerando essa correlação entre os *scores*, os loadings podem ser calculados como:

$$\widehat{\boldsymbol{q}}^{\boldsymbol{T}} = \mathbf{U}^{T}.\mathbf{y} \qquad \text{Eq. 5}$$

$$\mathbf{P}^{T} = \mathbf{T}^{\boldsymbol{T}}.\mathbf{X} \qquad \text{Eq. 6}$$

O valor de **y** predito para uma nova amostra $\mathbf{X_{new}}$ pode ser calculado em função dos *scores* **U** obtidos da correlação com **T**, através da equação 7.

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{q}}^{\boldsymbol{T}}.\mathbf{U} \qquad \text{Eq. 7}$$

Substituindo a matriz **U** na equação 7 pelas relações anteriores, se obtém a equação 8. Os coeficientes de regressão do modelo são obtidos pelo produto $\widehat{\boldsymbol{q}}^{\boldsymbol{T}}.\mathbf{w}.\mathbf{P}^{\mathbf{T}}$ como mostra a equação 9:

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{q}}^{\boldsymbol{T}}.\mathbf{w}.\mathbf{X_{new}}.\mathbf{P}^{\boldsymbol{T}} \qquad \text{Eq. 8}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{xb} \qquad \text{Eq. 9}$$

Onde **b** é o vetor dos coeficientes de regressão com mesmo número de colunas que **X**. Dada uma amostra desconhecida **x** ($\mathbf{x} \in \mathbf{X_{new}}$) pode-se prever o valor correspondente em **y** através da equação 9 (BRERETON; LLOYD, 2014)

O valor $\widehat{\boldsymbol{y}}$ previsto para a amostra **x** pode assumir valores entre 0 e 1, considerando duas classes. Considerando classes com mesmo número de elementos e com comportamento normal, o limite entre as duas classes pode ser estabelecido em 0,5. Assim, valores de $\widehat{\boldsymbol{y}} \geq 0{,}5$ indicam que a amostra pertence à classe correspondente a 1. O procedimento matemático descrito obedece a situação em que **y** tem média zero, ou seja, as classes discriminadas são de mesmo tamanho. Caso

as classes tenham números de elementos diferentes, o limite entre as classes irá variar para mais ou para menos de 0,5, dependendo de qual classe é maior. O pacote de ferramentas PLSToolbox implementa um algoritmo flexível para definição de limite entre as classes, podendo ser acessados os resultados de classificação com limite fixo de 0,5 ou flexível, baseado na estatística Bayesiana (WOLD; SJOSTROM, 2001).

Como já mencionado, a análise discriminante linear é outro método de classificação que tem se mostrado bastante útil para aplicações no contexto forense (SILVA, 2013). Uma característica interessante desta técnica é a simplicidade da fronteira de discriminação das amostras, uma vez que a classificação é feita baseando-se num limiar entre as classes, ou seja, a fronteira não é modelada para cada classe individualmente como ocorre no SIMCA. Uma limitação relacionada com restrição das fronteiras no LDA e PLS-DA é que obrigatoriamente todas as amostras devem ser classificadas em uma das classes consideradas (BROWN; WALCZAK; TAULER, 2009).

A LDA busca novos vetores de projeção que sejam capazes de mostrar a melhor dissimilaridade entre as amostras das diferentes classes e que também apresentem a maior similaridade entre as amostras pertencentes a uma mesma classe. A similaridade ou dissimilaridade das amostras é calculada utilizando como critério as distâncias mínimas de Mahalanobis normalizadas e a probabilidade total de erro na classificação das amostras. Nesse sentido, a razão da variância entre as amostras de classes diferentes e a variância das amostras dentro de cada classe individualmente deve ser otimizada na etapa de treinamento. Considerando que as amostras se distribuem normalmente nas classes com médias μ_1 e μ_2, covariâncias iguais $\sum$ e probabilidade iguais, a regra de discriminação para classificar uma amostra $\boldsymbol{x}$ em um grupo $\mathbf{G_1}$ assume a seguinte forma:

$$(x - \mu_1)^t \sum^{-1} (x - \mu_1) \leq (x - \mu_2)^t \sum^{-1} (x - \mu_2) \qquad \text{Eq. 10}$$

A média populacional μ_1, e a covariância $\sum$ são estimativas calculadas a partir das amostras e permitem estimar a probabilidade de uma amostra pertencer a classe 1 ou 2 (BROWN; WALCZAK; TAULER, 2009).

Uma restrição imposta para a LDA é que o número de variáveis seja sempre menor que o número de amostras de treinamento. Dessa forma, quando o número de variáveis excede o de amostras é necessário utilizar alguma técnica para redução do número de variáveis antes de construir os modelos. A redução da dimensionalidade dos dados pode ser realizada utilizando a LDA acoplado a PCA ou ainda algoritmos específicos da seleção de variáveis, como por exemplo, o Algoritmo de Projeções Sucessivas (SPA do inglês *Successive Projection Algorithm*) e o Algoritmo Genético (GA do inglês *Genetic Algorithm*) que foram utilizados nesta tese (PONTES *et al.*, 2005).

O SPA foi inicialmente desenvolvido para a seleção de variáveis para calibração e se trata de um método determinístico que tem o objetivo de selecionar as variáveis menos correlacionadas e que carreguem o máximo de informação dos conjuntos de dados (PONTES *et al.*, 2005). O objetivo do algoritmo é reduzir a colinearidade das variáveis e assim minimizar a função de custo utilizando operações de projeção de vetores. A função de custo para o SPA aplicado à classificação se baseia no risco médio do erro de classificação das amostras de acordo com o conjunto de variáveis ortogonais selecionadas (ARAÚJO *et al.*, 2001). Por outro lado, o algoritmo GA é um método estocástico de seleção de variáveis que simula um processo evolutivo para encontrar o conjunto mais adequado de variáveis que permita a classificação corretada do maior número possível de amostras de um dado conjunto de treinamento (LEARDI; GONZÁLES, 1998). Por este método, variáveis são inicialmente selecionadas randomicamente e em seguida são combinadas com novas gerações de variáveis aleatórias com o objetivo de melhorar os resultados de classificação. Como em cada geração alguns indivíduos (variáveis) são predominantes para o resultado de classificação, os novos conjuntos contendo apenas essas variáveis são combinadas aleatoriamente para se obter o ponto ótimo para uma função de custo. Ao se alcançar o critério de parada, as variáveis selecionadas são analisadas para determinar quais combinações apresentam melhor acurácia na classificação e menor custo (LEARDI; GONZÁLES, 1998).

O método de modelagem de classes SIMCA pode ser considerado mais simples que os supracitados, uma vez que a utilização dessa técnica para classificação se limita em comparar a variância das amostras com a variância das classes de modo independente e sem considerar a proximidade entre as classes. O SIMCA incorpora a PCA para reduzir a dimensionalidade e

quantificar a variância dos dados. Dessa forma, o conjunto de amostras correspondente a uma determinada classe pode ser representado por certo número de PCs que carrega as informações que descrevem o comportamento de dispersão espacial das amostras no novo sistema de coordenadas criado pela PCA (BRERETON, 2003a). Como cada classe é modelada separadamente, o número de PCs utilizadas para cada classe varia dependendo da variância total das amostras contidas em cada conjunto de amostras. A atribuição das classes para as amostras desconhecidas depende de diferentes fatores como a comparação dos resíduos e o poder de modelagem usado para avaliar o melhor ajuste das amostras ao modelo (BRERETON, 2003a; BROWN; WALCZAK; TAULER, 2009).

Como mencionado, SIMCA usa modelos de PCA para determinar espaços multidimensionais com bordas e limites definidos pela variância das amostras de cada classe. As amostras desconhecidas são projetadas nessas estruturas previamente definidas e a predição da classe é determinada de acordo com o ajuste das amostras dentro da estrutura de uma classe, mais de uma classe ou nenhuma das classes modeladas (BEEBE; PELL; SEASHOLS, 1998; BRERETON, 2003a). Para uma determinada classe $\boldsymbol{g}$ representada por um modelo de PCA, a distância d_{ig} de uma amostra x_i para o centro deste modelo pode ser determinada pela equação 11:

$$d_{ig} = \sqrt{\left(\frac{Q_{ig}}{Q_{0,95,g}}\right)^2 + \left(\frac{T_{ig}^2}{T_{0,95,g}^2}\right)^2} \qquad \text{Eq. 11}$$

Onde d_{ig} é a distancia que define a borda do modelo (o limite do modelo $\boldsymbol{g}$), $\boldsymbol{Q_{ig}}$ é o resíduo da amostra $\boldsymbol{x_i}$, $Q_{0,95,g}$ é o resíduo para as amostras que compõe a classe $\boldsymbol{g}$, $\boldsymbol{T_{ig}^2}$ é a distância da projeção da amostra x_i para a classe $\boldsymbol{g}$ e $\boldsymbol{T_{0.95,g}^2}$ é a distância das amostras que compõem a classe $\boldsymbol{g}$ para o centro do modelo. Na equação 11, todos os limites do modelo foram determinados a 95% de confiança. Considerando esses limite, uma amostra desconhecida projetada nesse espaço de classes pode ser considerada em uma classe se $\left(Q_{ig}/Q_{0.95,g}\right) \leq 1$ and $\left(T_{ig}^2/T_{0.95,g}^2\right) \leq 1$ (BRERETON, 2003a).

2.2.4 Métricas de desempenho

De modo geral, o percentual de erro é a figura de mérito mais comum para avaliar o desempenho para as técnicas de classificação. Contudo, existem outras figuras de mérito importantes como a exatidão, acurácia, sensibilidade e a especificidade. A eficiência de um modelo de classificação pode ser medida em função da capacidade desse modelo em classificar corretamente as amostras em suas respectivas classes. Os parâmetros sensibilidade e especificidade, quando avaliados conjuntamente, indicam a eficiência do modelo.

A sensibilidade (Sn) pode ser definida como a razão entre as amostras classificadas em uma classe a que pertencem e o total de amostras que foram classificadas nesta classe específica, ex. habilidade de evitar falso negativo, sendo matematicamente definida pela equação 12. A especificidade (Sp) por sua vez, é definida como a razão entre as amostras não incluídas em uma classe a que não pertencem e o total de amostras não incluídas nesta classe, ex. habilidade de evitar falso-positivos, sendo representada matematicamente pela equação 13

$$\mathrm{Sn} = \mathrm{VP}/(\mathrm{VP} + \mathrm{FN}) \qquad \text{Eq. 12}$$

$$\mathrm{Sp} = \mathrm{VN}/(\mathrm{VN} + \mathrm{FP}) \qquad \text{Eq. 13}$$

Onde VP é o total de amostras que foram classificadas na classe considerada e que pertencem a classe, FN é o total de amostras classificadas como não pertencentes na classe considerada e que pertencem a classe e VN é o total de amostras classificadas como não pertencentes a classe considerada e que realmente não pertencem a classe. A sensibilidade e a especificidade são os parâmetros representam a melhor maneira de avaliar os erros mais comuns em classificação e, por isso, foram as métricas utilizadas para os modelos de classificação nesta tese (BOTELHO *et al.*, 2015; BROWN; WALCZAK; TAULER, 2009).

2.3 MODELOS HIERÁRQUICOS DE CLASSIFICAÇÃO POR FUSÃO DE TÉCNICAS QUIMIOMÉTRICAS

Os métodos convencionais para análise discriminantes, como por exemplo PLS-DA e LDA, são bastante adequados para discriminação de dois ou mais grupos de amostras distintos. Esses

métodos apresentam uma função discriminante que estabelece uma fronteira entre as diferentes classes de objetos (BRERETON; LLOYD, 2014). Um fator problemático para essas técnicas é a sua incapacidade de identificar amostras estranhas aos modelos (resíduos e influência altos) e que não pertencem a nenhum dos grupos de amostras utilizado para criar os modelos discriminantes. Nesse sentido, as técnicas de modelagem flexível de classes, como por exemplo SIMCA, apresentam a vantagem de possibilitar a criação de modelos de classificação para cada classe de amostras (BEEBE; PELL; SEASHOLS, 1998). Essas técnicas de modelagem flexível de classes permitem que as amostras com características muito diferentes das classes modeladas sejam classificadas como pertencentes a uma classe de amostras indefinidas. A desvantagem desses métodos reside no fato de que os modelos criados dependem fortemente da homogeneidade das classes e dificilmente apresentam bons resultados para distinguir classes muito similares.

Os modelos hierárquicos de classificação por fusão de técnicas quimiométricas possibilitam a formulação de um esquema de classificação que combina as características de métodos de modelagem de classes e de análise discriminante. Essa nova ferramenta quimiométrica permite que técnicas de reconhecimento de padrão não-supervisionadas e supervisionadas sejam utilizadas como funções de decisão dentro de um modelo de classificação mais robusto (TAKAMURA *et al.*, 2018).

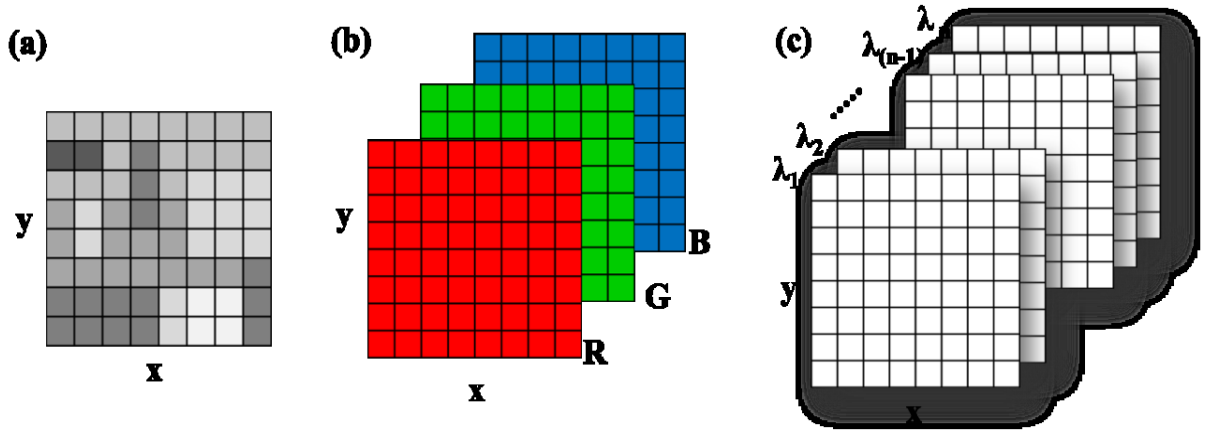
Dado um modelo PCA para um conjunto de amostras qualquer, para verificar se uma amostra é similar àquelas descritas pelo modelo PCA, pode-se utilizar um teste estatístico, como por exemplo o cálculo de Q residual, para comparar as amostras. Desse modo, o modelo de PCA serve como uma função de decisão para uma posterior classificação de amostras. O exemplo dado se assemelha ao método SIMCA, porém com menos restrições matemáticas. Uma técnica como LDA também pode ser usada como função de decisão dentro do método hierárquico de classificação, porém nesse caso a própria função discriminante do LDA é utilizada para realizar a tomada de decisão (escolher a classe das amostras). Diferentes técnicas podem ser combinadas para gerar um modelo de classificação hierárquico que considera todas as funções de decisão como critérios para definir a classe de uma determinada amostras e negligenciar amostras que são diferentes das amostras utilizadas para a construção dos modelos (TAKAMURA *et al.*, 2018).

Takamura e colaboradores (2018), descrevem diferentes situações nas quais o método hierárquico de fusão de técnicas é a escolha mais adequada para solucionar os problemas de classificação.

2.4 IMAGENS HIPERESPECTRAIS

As imagens digitais podem ser definidas como uma função de duas dimensões $f(x, y)$, em que x e y são cordenadas espaciais finitas e com valores discretos. Assim, um conjunto de coordenadas *xy* definem um pixel que pode assumir valores discretos ou um conjunto de valores. Quando os pixels de uma imagem apresentam mais de um valor de intensidade, função que descreve a imagem ganha mais uma dimensão passando a ser representada como $f(x, y, z)$ (GONZALEZ; WOODS, 2002). Quando a dimensão z assume um valor discreto a imagem resultante é chamada de imagem em escala de cinza. Para z contendo os três canais de cores vermelho, azul e verde (*red, blue, green* – RGB) a imagem é conhecida com RGB. As imagens hiperespectrais (*Hyperspectral Images* – HSI) por sua vez, são definidas como uma função $f(x, y, z)$ em que a terceira dimensão pode ser um vetor composto por valores correspondentes a uma determinada medida instrumental, como por exemplo espectros de infravermelho (PRATS-MONTALBÁN; DE JUAN; FERRER, 2011). Nesse terceiro tipo de imagem cada pixel carrega a informação espectral e, portanto, química das amostras. A Figura 3 apresenta as formas de imagens citadas.

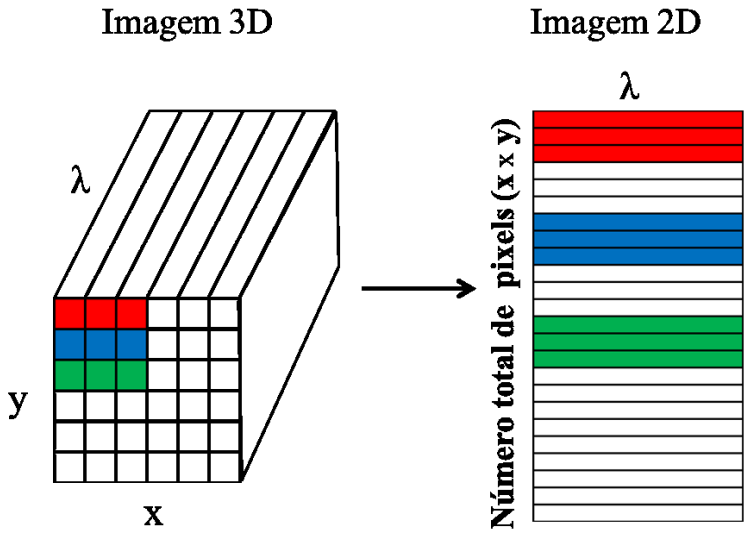
Figura 3 – (a) representação da imagem em escala de cinza, (b) da imagem RGB e seus canais de cores e (c) esquema de uma HSI com n comprimentos de onda.



Fonte: Do autor (2019).

O processamento das imagens hiperespectrais inclui algumas etapas básicas que podem ser implementadas sem a necessidade de utilizar ferramentas específicas para tratamento de imagens. Os cubos de imagens HSI podem ser desdobrados em duas dimensões, convertendo as coordenadas espaciais em linhas que representam as amostras individuais de um conjunto de dados e a coordenada z permanecem como as variáveis do conjunto de dados (DE JUAN *et al.*, 2004). Dessa forma, todas as ferramentas quimiométricas clássicas podem ser aplicadas no conjunto de dados bidimensional. A decomposição do cubo ocorre de acordo com o esquema demonstrado na Figura 4:

Figura 4 – Desdobramento da imagem de 3 dimensões para 2 dimensões.



Fonte: Do autor (2019).

Após aplicar o desdobramento, as imagens podem ser remontadas. Esse processo de interconversão entre uma matriz 2D e 3D pode ocorrer em qualquer etapa do tratamento de dados. Uma desvantagem dessa abordagem de tratamento de imagens é a perda da informação de contraste espacial entre os pixels. Para evitar essas perdas de informações resultantes do desdobramento das imagens é importante avaliar as imagens brutas através de projeções das imagens para cada comprimento de onda do espectro.

3 IDENTIFICAÇÃO DE MANCHAS DE SANGUE HUMANO EM POSSÍVEIS CENAS DE CRIME UTILIZANDO EQUIPAMENTO PORTÁTIL E CÂMERA HIPERESPECTRAL

3.1 IDENTIFICAÇÃO DE MANCHAS DE SANGUE HUMANO

Cenas de crime são ambientes complexos que apresentam um grande volume de informações sobre o crime, as vítimas, os suspeitos e a dinâmica da ação. O sangue é talvez um dos mais importantes traços forenses que podem ser encontrados em cenas de crime (JAMES; KISH; SUTTON, 2005) e pode confirmar a presença de um indivíduo, fornecendo informações relacionadas com a atividade ou ação, além de por meio de análise de DNA levar a eventual identificação de um suspeito (BREMMER; NADORT; *et al.*, 2011). A análise de DNA é um dos recursos mais recorrentes em investigações criminais, contudo o método demanda tempo e recursos elevados e sua eficácia depende de uma seleção cuidados das evidências afim de evitar erros (JAMES; KISH; SUTTON, 2005). A presença de outras substâncias com aspecto visual similar ao sangue (tais como molhos de pimenta, vinho tinto, ketchup, molho shoyu, etc.) também dificulta o trabalho dos peritos durante a coleta das evidências. Portanto, para obter melhor custo-benefício e eficiência nas análises forense, é de grande importância a correta identificação da natureza de machas suspeitas encontradas em cenas de crime.

Os métodos utilizados para análise de cenas de crime têm evoluído e se multiplicado de acordo com o desenvolvimento científico, surgimento de novas técnicas analíticas e a necessidade de cada caso. Algumas técnicas caíram em desuso e outras se popularizaram ao ponto de serem largamente utilizadas sem prévia avaliação da sua efetividade e compatibilidade com o caso específico (JAMES; KISH; SUTTON, 2005). De acordo com uma recente revisão da literatura (ZAPATA; OSSA; GARCÍA-RUIZ, 2015), existem diversos métodos que podem ser utilizados para a detecção e confirmação da presença de traços de sangue diretamente em locais de crime. Esses métodos incluem a inspeção visual das manchas usando o microscópio; métodos químicos como os largamente difundidos testes do Luminol ou Kastle-Mayer e o teste de microcristais de Takayama ou Teichmann; métodos espectroscópicos como as fontes de luz alternadas e absorção UV-Vis; e métodos imunológicos como os kits de teste para anticorpos RSIDTM ou ABA Card□

Hematrace (JAMES; KISH; SUTTON, 2005; STOILOVIC, 1991; ZAPATA; OSSA; GARCÍA-RUIZ, 2015). De modo geral, esses métodos são presuntivos e pouco específicos para identificar sangue humano. Dentre os problemas que eles apresentam podem-se destacar alguns: são afetados por diversos interferentes comuns em diversos ambientes; alguns métodos requerem ambiente escuro ou uso de cotonetes para coleta das evidências; usam reagentes que podem invalidar uma subsequente análise de DNA ou até mesmo destruir a amostra; alguns métodos apresentam reatividade com mais de um tipo de fluido corporal; e muitos outros podem apresentar falso-positivo com substâncias comumente encontradas (VIRKLER; LEDNEV, 2009a).

Portanto, existe uma necessidade de novos métodos que não sejam apenas presuntivos, mas também confirmatórios, não destrutivos e específicos para identificar sangue humano. Além disso, os estudos mais atuais exprimem a necessidade de equipamentos portáteis, que apresentem fácil manuseio e que forneçam informações confiáveis e úteis para a rápida identificação de sangue humano diretamente em cenas de crime (BREMMER; EDELMAN; *et al.*, 2011).

Neste sentido, as técnicas de espectroscopia vibracional como Raman e Infravermelho têm mostrado grande potencial com inúmeras aplicações para a ciência forense (CHALMERS; EDWARDS; HARGREAVES, 2012; SILVA, SANTOS; BRAZ; PIMENTEL, 2019) e particularmente para identificação de traços de sangue humano (ZAPATA; OSSA; GARCÍA-RUIZ, 2015), devido ao fato de que o sangue humano apresenta assinatura química bastante característica. Adicionalmente, estas técnicas permitem análises rápidas, não-invasivas e não-destrutivas das amostras. A espectroscopia Raman tem recebido mais destaque nesse tipo de aplicação quando comparada à espectroscopia no infravermelho. Em grande parte, essa tendência decorre do fato de que a água presente no sangue tem pouca influência nos espectros obtidos com espectroscopia Raman (VIRKLER; LEDNEV, 2008, 2009a, 2009b, 2010). Contudo, a espectroscopia IR tem vantagens práticas em relação à Raman, uma vez que a fluorescência e a luz do ambiente externo não influenciam na qualidade dos espectros adquiridos, fato este que é bastante importante para análise direta em cenas de crime (VIRKLER; LEDNEV, 2010).

O constante desenvolvimento na tecnologia resultou em grandes avanços na instrumentação analítica nos últimos anos, como por exemplo a miniaturização dos equipamentos Raman e NIR

utilizados em diversos métodos de análise (LUTZ *et al.*, 2014). Dispositivos portáteis (que podem ser deslocados) e de mão (com dimensões comparáveis a uma mão) são agora importantes instrumentos para diferentes setores incluindo ciências forenses, onde as análises têm se tornado rápidas, mais efetivas e adaptáveis para diferentes condições de análise, além de contribuir para a redução na quantidade de tempo de resposta e de recursos consumidos em laboratório (CHALMERS; EDWARDS; HARGREAVES, 2012; MARQUES *et al.*, 2016).

Até o momento, os trabalhos publicados empregando a espectroscopia de infravermelho médio para discriminação de manchas de sangue consistem principalmente da comparação do perfil espectral de amostras analisadas em equipamentos de bancada (ELKINS, 2011; ORPHANOU, 2015). Os trabalhos que utilizam a espectroscopia de infravermelho próximo para aquisição dos espectros geralmente apresentam uma interpretação estatística e matemática mais completa dos dados. O trabalho de Botonjic-Sehic *et al.* (2009) é um exemplo de utilização da espectroscopia NIR combinada com técnicas quimiométricas para avaliação do tempo de degradação de manchas de sangue. Nesse trabalho, os autores avaliam a mudança do perfil dos espectros NIR de amostras de sangue em decorrência da conversão da desoxihemoglobina (HbO_2) em metoxihemoglobina (MetHb). As amostras foram preparadas sobre gaze e placas de vidro, sendo analisadas durante as primeiras horas e diariamente por 30 dias. A banda de água (~1400 nm e ~1900 nm) dominou os espectros nas primeiras horas de análise e um modelo de regressão por mínimos quadrados foi construído. O modelo construído apenas funcionou devido à informação relacionada a variação da banda de água na primeira hora. Um outro modelo de regressão foi construído com os espectros coletados com mais de 1h de secagem e o desvio padrão para previsão de 3 amostras independentes foi de cerca de 2,3 h em um período total de 590 h (BOTONJIC- SEHIC *et al.*, 2009). Vale ressaltar que o trabalho de Botonjic-Sehic *et al.* (2009) não deixa claro como os modelos foram criados e como foram determinadas os desvios no valor predito.

Um outro exemplo interessante é o trabalho de Edelman *et al.* (2012) que utilizou a espectroscopia NIR combinada com refletância na região do visível e o método de regressão por mínimos quadrados parciais para analisar manchas de sangue em algodão. Nesse mesmo estudo foi avaliado o efeito do tempo no perfil espectral de amostras de sangue depositado em algodão. Os autores utilizaram dois equipamentos de bancada com faixas de trabalho de 400-2500 nm e 800-

2778 nm para aquisição dos espectros. Os resultados obtidos demonstraram que manchas de sangue podem ser distinguidas das demais substâncias de cor vermelha e de algodão colorido com 100% de sensitividade e especificidade. Além disso, o erro quadrático médio relativo da previsão (RMSEP: *Root Mean Squared of Prediction*) para a datação das manchas de sangue em algodão escuro foi de 8,9% (EDELMAN *et al.*, 2012). A espectroscopia na região NIR mostrou-se adequada para a datação de manchas de sangue em curto prazo, mas apresentou algumas limitações na avaliação da contribuição nos espectros dos diferentes componentes presentes no sangue e interpretação das bandas de absorção. Uma outra aplicação da espectroscopia NIR combinada com PLS-DA foi demonstrada por Zhang *et al.* (2016) para discriminação de sangue humano e sangue de animais. As amostras foram analisadas por transmitância em cubetas de 5 mm. Nesse trabalho, os autores construíram um modelo capaz de discriminar amostras de sangue humano, sangue de rato e sangue de macaco, obtendo 100% de correta classificação. Um conjunto de 7 amostras de sangue de porco foi testado no modelo construído e todas foram classificadas como sangue animal. Embora o método desenvolvido por Zhang *et al*. (2016) tenha obtido alta eficiência, ele demanda volume expressivo de amostras de sangue, adicionado um agente químico anti-coagulante e o número de variáveis latentes utilizado foi alto (8LV).

Edelman *et al.* (2012) também avaliaram a possibilidade de empregar HSI-NIR para a detecção para discriminar amostras de sangue de outras substâncias de cor similar em diversos substrato com texturas e cores diferentes. Em um trabalho posterior, Edelman *et al.* (2013) demonstraram o potencial das imagens hiperespectrais visível-NIR (400-1000 nm) para identificação de sangue humano depositado em algodão branco. Foram utilizados 30 falsos-positivos visuais para discriminar das amostras de sangue. A identificação das manchas de sangue foi realizada pela comparação do perfil espectral das amostras considerando o limite/limiar de correlação superior a 0,98 (EDELMAN; VAN LEEUWEN; AALDERS, 2013).

A utilização de técnicas quimiométricas para auxiliar na interpretação das informações químicas contidas nos espectros NIR é uma prática bem estabelecida na química analítica e apresenta diversas aplicações em diversos campos de pesquisa como alimentos, combustíveis e forense (MARQUES *et al.*, 2016; ORPHANOU, 2015; SILVA *et al.*, 2014; SILVA; BRAZ; PIMENTEL, 2019) . Os estudos relacionados à identificação de sangue humano em cenas de crime

ainda carecem de resultados expressivos com relação a metodologias rápidas e precisas para identificar e classificar sangue humano em diferentes materiais encontrados nesses locais. Portanto, o presente trabalho teve como objetivo investigar o potencial de diferentes estratégias para análise de manchas suspeitas em diferentes materiais utilizando a espectroscopia NIR e métodos de classificação multivariada.

3.1.1 Objetivo geral para identificação de manchas de sangue

- Desenvolver uma metodologia usando um espectrômetro ultra portátil e câmera de imagem na região NIR para identificação e análise confirmatória, não invasiva e não destrutiva de manchas de sangue humano depositadas em diferentes pisos e tecidos.

3.1.1.1 Objetivos específicos para identificação de manchas de sangue

- Avaliar a eficácia de diferentes técnicas de reconhecimento de padrão supervisionadas tais como PLS-DA, SPA-LDA, GA-LDA e SIMCA, para identificação e classificação de manchas de sangue humano entre outras substâncias de aspecto similar depositadas em um mesmo tipo de substrato (piso), empregando um espectrômetro portátil.

- Aplicar a metodologia desenvolvida na primeira etapa do trabalho para desenvolver modelos capazes de identificar sangue humano depositadas em diferentes tipos de pisos.

-Avaliar o potencial de modelos hierárquicos com fusão de técnicas quimiométricas para identificação de manchas de sangue humano em tecidos sintéticos e de algodão, empregando uma câmera de imagens.

3.2 REVISÃO DA LITERATURA SOBRE IDENTIFICAÇÃO DE MANCHAS DE SANGUE

As manchas de sangue são evidências comuns em diferentes crimes violentos, sendo também importantes evidências para investigação e elucidação dos crimes por permitir inferir, por exemplo, sobre a dinâmica do crime, presença de suspeitos, tempo decorrido do crime, tempo de *post-mortem* (JAMES; KISH; SUTTON, 2005). Após identificar o sangue, a etapa de caracterização do material genético contido na amostra consiste principalmente da análise do DNA (BREMMER *et al.*, 2012). Embora o DNA seja o método mais específico para análise de sangue,

a coleta adequada da amostra é uma etapa extremamente importante para o sucesso da análise. Além disso, uma coleta adequada evita que falsos-positivos sejam levados ao perito e reduz o tempo e custo das investigações. A presença de substâncias com aspecto visual similar ao sangue dificulta a coleta das evidências e por isso foram desenvolvidos diversos métodos para identificação de sangue diretamente em materiais encontrados em cenas de crime (CHEMELLO, 2007).

Os métodos utilizados para identificação prévia de evidências são conhecidos como testes presuntivos, que podem ser entendidos como testes que, quando positivos, permitem a coleta para posterior teste confirmatório como DNA (JAMES; KISH; SUTTON, 2005). Os testes presuntivos mais comuns são os métodos baseados em reações químicas colorimétricas ou luminescentes, mas a inspeção visual das cenas de crime utilizando filtros de luz coloridas ainda é uma técnica bastante comum (ZAPATA; OSSA; GARCÍA-RUIZ, 2015). A maior parte dos métodos presuntivos para detecção de sangue utilizam reagentes que são bastante sensíveis como a benzidina usada para o teste de Adler, a fenolftaleína para o teste de Kastle-Mayer e 3-aminophthalhydrazida conhecida como luminol que sofrem oxidação na presença de hemoglobina do sangue (BARNI *et al.*, 2007; JAMES; KISH; SUTTON, 2005). Embora bastante sensíveis, esses métodos apresentam inúmeros falsos-positivos e podem anular a evidência devido degradação química dos componentes do sangue que impedem uma posterior análise confirmatória (LARKIN; GANNICLIFFE, 2008). Também existem os testes de formação de cristais que são baseados em reações químicas com os radicais heme não proteicos do sangue, porém eles exigem uma inspeção visual no microscópio para confirmar a presença de sangue. Isso torna as análises por micro cristais demoradas e caras (JAMES; KISH; SUTTON, 2005).

Diante destas limitações, várias alternativas estão sendo avaliadas no campo da espectroscopia. A grande vantagem dos métodos espectroscópicos é o fato de que as análises são rápidas, quase não exigem preparo de amostras e não alteram a composição química das mesmas (ZAPATA; OSSA; GARCÍA-RUIZ, 2015).

Seidl *et al*. (2008) demonstram o potencial de identificação de diversos fluidos corporais e, em especial sangue, utilizando uma fonte de luz ultravioleta para distinguir as manchas de sangue depositadas em substratos com cores variadas. Embora tenha obtido sucesso na identificação das

manchas de sangue, a técnica demonstrou ser pouco específica quando confrontada com substâncias de cor similar ao sangue. A difração de raios X também foi avaliada para identificação de sangue por Trombka e colaboradores (2002), demonstrando ser uma técnica adequada para identificação de manchas através do metal ferro presente no sangue. A ressonância magnética nuclear e a espectroscopia de massas também foram estudadas, porém as espectroscopias Raman e infravermelho são as mais estudas para identificação de fluidos corporais em cenas de crime (ZAPATA; OSSA; GARCÍA-RUIZ, 2015).

A espectroscopia Raman tem sido largamente explorada para identificação de sangue e discriminação de sangue humano de sangue de várias espécies de animais. Vários trabalhos reportam o uso da espectroscopia Raman para identificação de sangue através da análise dos perfis espectrais (SIKIRZHYTSKI; VIRKLER; LEDNEV, 2010a; VIRKLER; LEDNEV, 2008) e também utilizando técnicas quimiométricas para tratamento dos dados e discriminação de amostras (DOTY; LEDNEV, 2018; MCLAUGHLIN; DOTY; LEDNEV, 2014; SIKIRZHYTSKAYA; SIKIRZHYTSKI; LEDNEV, 2012; SIKIRZHYTSKI; SIKIRZHYTSKAYA; LEDNEV, 2012; VIRKLER; LEDNEV, 2009b). Nas pesquisas que envolvem métodos de análise multivariada dos espectros, os autores buscaram caracterizar as amostras e diferenciar as espécies doadoras do sangue. Virkler & Lednev (2009b) utilizaram PCA para determinar se amostras de sangue humano e de animais apresentavam diferenças químicas significativas. Em um trabalho posterior Sikirzhytski e colaboradores (2012) utilizaram ferramentas quimiométricas mais complexas e adequadas para discriminação de amostras para diferenciar amostras de misturas sangue e sêmen em diferentes proporções. Seguindo a mesma linha de pesquisa, Mclaughlin *et al.* (2014) utilizaram PLS-DA em espectros Raman para discriminar sangue humano de sangue animal e obtiveram resultados com acurácia de 100% para a discriminação, incluindo em casos de testes-cego. Esses trabalhos têm em comum a metodologia para a obtenção dos espectros Raman, que envolve a deposição das amostras em placas de alumínio para posterior análise com o equipamento. Essa método de análise reflete a limitação experimental da técnica Raman referente ao efeito da fluorescência nos espectros além da dificuldade otimização experimental para reduzir a influência dos substratos nos espectros (CHALMERS; EDWARDS; HARGREAVES, 2012).

Diferente do que se observa na espectroscopia Raman, as técnicas de análise por infravermelho são mais simples de otimizar e não apresentam o efeito indesejável da luz ambiente, podendo ser aplicadas em condições mais adversas (CHALMERS; EDWARDS; HARGREAVES, 2012). A região do infravermelho próximo (NIR) é mais explorada que o infravermelho médio devido à sua simplicidade instrumental e a gama de equipamentos NIR disponíveis. Além do mais, a espectroscopia NIR avançou bastante com o desenvolvimento tecnológico na instrumentação que possibilitou a construção de câmeras de infravermelho próximo (GOWEN *et al.*, 2015) e a miniaturização dos espectrômetros resultando em equipamentos portáteis de mão (PEREIRA *et al.*, 2017; SILVA, C. *et al.*, 2017).

Zhang e colaboradores (2016), em um trabalho similar ao de Virkler e Lednev (2009b) que utilizaram espectroscopia Raman, demonstraram o potencial para a discriminação de sangue de animais e humanos através da espectroscopia NIR e PLS-DA. Os espectros foram obtidos a partir de sangue humano, de macaco e de rato utilizando a técnica de transmitância, sendo as amostras analisadas dentro de cubetas (ZHANG *et al.*, 2016). Esse trabalho é uma exceção, uma vez que foge da lógica que havia se estabelecido por outros autores que buscaram mais proximidade com casos reais forense e analisaram manchas de sangue diretamente em superfícies comuns em locais de crime. A questão dos substratos coloridos foi estudada por Edelman *et al.* (2012), que utilizaram algodões com diferentes cores para servir de substrato para depositar sangue humano e outras 30 substâncias similares à sangue, sendo todas as amostras analisadas através da espectroscopia de reflectância vis-NIR.

Outra abordagem foi avaliada por nosso grupo de pesquisa que utilizou um equipamento NIR portátil e técnicas quimiométricas para analisar e discriminar manchas de sangue humano, sangue animal e outras 7 substâncias com aspecto visual similar a sangue depositadas em pisos e materiais habitualmente presentes em locais de crime (PEREIRA *et al.*, 2017). Esse trabalho será discutido com mais detalhes nesta tese. Morillas *et al.* (2018) também utilizaram um espectrômetro NIR portátil para discriminar sangue humano de outras substâncias visualmente similares depositadas em placas de vidro, ladrilhos, madeira, couro, acrílico e algodão. Os resultados dessa pesquisa apresentaram taxas de verdadeiro positivos de 94% e taxa de verdadeiros negativos de 86% para o melhor modelo. Entretanto, os autores não descrevem em detalhes qual a técnica

quimiométrica empregada. Os autores mencionaram apenas que as ferramentas quimiométricas utilizadas eram baseadas em regressão PLS.

3.3 METODOLOGIA PARA IDENTIFICAÇÃO DE MANCHAS DE SANGUE HUMANO

O estudo sobre a identificação de manchas de sangue humano seguiu duas linhas de pesquisa que abordam dois tópicos bastante diferentes. A primeira aborda a problemática de identificação de sangue e distinção entre espécies a partir de manchas encontradas em materiais comuns em possíveis cenas de crime (vidro, metal, cerâmica e porcelanato), além da construção de modelos que se aplicam em dois tipos de piso. Nessa primeira linha de pesquisa o foco foi mantido na utilização de um espectrômetro NIR portátil, uma vez que é difícil o transporte da amostra para o laboratório, particularmente no caso dos pisos.

A segunda linha aborda o problema da identificação e discriminação de manchas de sangue humano em tecidos com cores diferentes, tecidos estampados e tecidos com formulações de fibra diferentes. Para esse estudo, os dados espectrais foram obtidos através de uma câmera HSI-NIR e modelados utilizando o método hierárquico de classificação por meio de fusão de técnicas quimiométricas (PCA e PLS-DA).

Como as pesquisas envolvem a coleta e manipulação de sangue humano, o projeto foi submetido ao Comitê de Ética em Pesquisa envolvendo Seres Humanos da Universidade Federal de Pernambuco (CEP-UFPE), sendo aprovado e recebendo o parecer nº1.059.225, Os colaboradores que forneceram as amostras de sangue assinaram o Termo de Consentimento de Livre Esclarecimento (TCLE), que se encontra no Apêndice 1.

3.3.1 Preparo de amostras para pisos, placas de vidro e metal

O desenvolvimento de modelos com amostras de apenas um dos substratos e de modelos combinando dois tipos de pisos ocorreram de forma independente e em momentos diferentes da pesquisa. Dessa forma, o desenvolvimento de cada tipo de modelo será tratados como dois estudos independentes. Inicialmente, para avaliar o potencial dos equipamentos portáteis, um estudo foi realizado para construir modelos individuais para cada um dos substratos. Numa segunda etapa,

manchas de sangue depositadas em dois substratos similares (tipos de pisos) foram utilizadas para construção dos modelos combinando essas informações.

3.3.1.1 Amostras para construção dos modelos para substratos individuais

Amostras de sangue humano de 31 doadores (14 homens e 17 mulheres), foram inicialmente adquiridas. O sangue foi coletado diretamente dos capilares dos dedos dos doadores usando agulhas individuais e esterilizadas por um colaborador qualificado. Amostras de sangue animal provenientes de um cachorro e um gato foram adquiridas junto ao hospital veterinário da Universidade Federal Rural de Pernambuco (UFRPE). Adicionalmente, 5 produtos de cor avermelhada e visualmente similares ao sangue (vinagre balsâmico, batom vermelho, vinho tinto, pimenta e molho shoyo) foram utilizados para preparar amostras de falso-positivos comuns. Dependendo do aspecto físico das amostras, elas foram depositadas diretamente sobre os substratos utilizando uma pipeta de Pasteur (2 gotas) em 4 diferentes tipos de substrato: placas de piso porcelanato na cor bege, placas brancas de piso cerâmico, placas de vidro transparente e a parte metálica de cutelos idênticos. Todos os substratos foram previamente lavados com água e sabão, e em seguida foram deixados em condições ambiente até a completa secagem.

O volume de sangue humano das manchas preparadas não foi controlado. A Tabela 2 resume o preparo das amostras e o número de manchas depositadas em cada substrato. O número de manchas depositadas em cada substrato foi diferente devido à variação do volume de sangue extraído de cada doador. As 220 manchas foram mantidas em condição ambiente, com temperatura variando entre 20 e 26ºC, para secar durante três dias antes da análise que foi realizada em uma única batelada com o espectrômetro NIR ultra portátil.

Tabela 2 – Resumo do preparo de amostras e distribuição das manchas em cada substrato.

Substratos	Número de Manchas		
	Sangue Humano	Sangue Animal	Falso-Positivos Comuns
Porcelanato	27 (7 doadores)	10	15
Cerâmica	36 (9 doadores)	10	15
Vidro	33 (9 doadores)	10	15
Metal	25 (6 doadores)	9	15

Fonte: Do autor (2019)

3.3.1.2 Amostras para construção de modelos para dois substratos (modelos cruzados)

Uma nova base de dados foi gerada a partir da coletada de novas amostras de sangue humano (14 homens e 7 mulheres) sendo 21 voluntários no total, bem como sangue de 4 animais (3 carneiros e 1 cavalo) utilizando o mesmo procedimento descrito no item 4.1.1. Além dessas, também foram preparadas novas amostras de falso-positivos comuns com as mesmas substâncias utilizadas na 1ª etapa da pesquisa e mais 2 novos produtos de coloração avermelhada (ketchup e geleia). Os substratos utilizados nessa etapa se restringiram apenas a placas de piso, porém com mais variabilidade, sendo 3 tipos de cerâmicas e 2 tipos de porcelanatos. As manchas foram preparadas sobre estes 5 substratos seguindo os mesmos protocolos utilizados para o preparo de amostras da etapa anterior deste trabalho. Após cerca de 2 horas de secagem das amostras em condições ambiente (temperatura entre 20 e 26 ºC), todas as manchas foram analisadas utilizando o espectrômetro NIR ultra portátil. A Tabela 3 resume o preparo das amostras e o número de espectros produzidos em cada modelo.

Tabela 3 – Resumo do preparo e aquisição de espectros para os modelos construídos empregando dois substratos (modelos cruzados).

Substratos	Amostra (nº de espectros)		
	Sangue humano	Sangue Animal	Falso-Positivos Comuns
Porcelanato 1 x cerâmica 1	116 (11 doadores)	65	213
Cerâmica 1 x porcelanato 2	96(10 doadores)	65	153
Cerâmica 1 x cerâmica 2	110 (13 doadores)	63	161
Cerâmica 3 x cerâmica 1	116 (13 doadores)	64	153
Cerâmica 2 x cerâmica 3	122 (14 doadores)	63	104
Porcelanato 1 x porcelanato 2	108 (11 doadores)	66	156

Fonte: Do autor (2019)

3.3.2 Amostras de sangue em tecido

Tecidos naturais a base de algodão e tecidos sintéticos a base de náilon, poliéster e acrílico, foram adquiridos no comércio da cidade de Recife, PE. Foram obtidos tecidos nas cores bege, branco, preto, vermelho e estampados. Os tecidos estampados de algodão apresentam estampas impressas sobre o tecido como uma pintura na superfície dos fios, enquanto os tecidos sintéticos apresentam estampas por meio de tintas absorvidas pelos fios do tecido. Um total de 10 tipos de tecidos foram utilizados no preparo das amostras. O sangue humano foi coletado de 16 voluntários (9 homens e 5 mulheres) através dos vasos capilares do dedo de cada doador utilizando o mesmo procedimento anteriormente descrito. Em cada tecido foram preparadas entre 4 e 6 manchas de sangue humano de doadores diferentes como nos exemplos da Figura 5.

Figura 5 – Manchas de sangue humano em tecidos intercalados na ordem algodão e sintético.

Fonte: Do autor (2019)

Sangue de cabra e cavalo foram utilizados para preparar 2 manchas em cada tecido. O sangue fresco foi obtido junto ao hospital veterinário da UFRPE e utilizado para preparar as manchas no local de coleta. Adicionalmente, manchas com falsos-positivos comuns (ketchup, vinho tinto, vinagre balsâmico, molho shoyu, molho de pimenta, batom e geleia) com aspecto visual similar à sangue também foram preparas em todos os substratos. Todas as amostras foram armazenadas no laboratório, com temperatura variando de 24 à 28ºC, durante 4 dias até a análise.

3.3.3 Instrumentação para análise das manchas de sangue

3.3.3.1 Aquisição dos espectros nas manchas preparadas em placas de metal, vidro e pisos

Espectros de infravermelho próximo foram coletados em cada mancha utilizando o espectrômetro MicroNir 1700 da Viavi. Este instrumento tem um filtro linear variável acoplado a um arranjo de detectores de Arseneto de Índio e Gálio (InGaAs), duas lâmpadas de tungstênio como fonte de radiação infravermelha e uma interface USB para aquisição e transferência de dados. O MicroNir 1700 tem dimensões de 45 mm de diâmetro e 42 mm de altura, pesando 60 g. O dispositivo trabalha na faixa espectral de 908 nm a 1676 nm e apresenta resolução espectral de 12,5 nm. As medidas foram realizadas com 64 varreduras e tempo de integração de 5 milissegundos, sendo adquiridos três espectros em diferentes posições em cada gota para contemplar a máxima variabilidade de cada amostra. Devido à alta transparência do vidro, o que

permite que a radiação atravesse a placa, um suporte de Teflon foi posicionado abaixo das placas no momento da aquisição para evitar absorção da bancada.

3.3.3.2 Aquisição das imagens das manchas de preparadas em tecido

A câmera SisuChema da Specim foi utilizada para aquisição das HSI-NIR. A lente usada produzia imagens com dimensões aproximadas de 15×5 cm e pixels com 156×156 µm. A faixa espectral do equipamento foi de 928 nm a 2524 nm, com velocidade de amostragem de 30 mm/s, resolução espacial de 6,3 nm e resolução espectral de 10 nm. Uma placa de teflon foi colocada na base da bandeja do equipamento para servir de base para as amostras.

3.3.4 Pré-processamentos e construção dos modelos para manchas preparadas em placas de vidro, metal e pisos.

No sentido de reduzir os efeitos resultantes de variações instrumentais e efeitos indesejáveis, diferentes pré-processamentos foram aplicados nos espectros, tais como: 1ª e 2ª derivadas com filtro de suavização Savitzky-Golay (SG), variando o número de pontos da janela (de 3 a 15 pontos) e com polinômio de segunda ordem; suavização com filtro SG (janela com 3 a 15 pontos e polinômio de segunda ordem); Correção de Sinal Multiplicativo (MSC do inglês *Multiplicative Signal Correction*); Variação Normal Padrão (SNV do inglês *Standart Normal Variate*); normalização pela área, faixa e máximo; e centralização na média. O melhor resultado foi escolhido baseado em inspeção visual dos espectros corrigidos e através dos resultados obtidos pelo método SIMCA, construído para as amostras depositadas em placas de porcelanato.

Além dos pré-processamentos supracitados, a correção por Mínimos Quadrados Parciais Ponderados Generalizados (GLSW do inglês *Generalized Least Squared Weighted*) também foi testada para os espectros dos modelos construídos para manchas depositadas em dois tipos de pisos (que foram nomeados de modelos cruzados). O pré-processamento mais adequado para os espectros foi definido através da inspeção visual e dos resultados de classificação através da PLS-DA.

Adicionalmente, as amostras preparadas para construção dos modelos individuais foram utilizadas para construir modelos LDA com seleção de variáveis utilizando o algoritmo genético e

o algoritmo de projeções sucessivas. Os diferentes tratamentos foram aplicados para cada substrato para a construção dos modelos de classificação individuais com os dados da 1ª etapa. Três diferentes classes foram consideradas: Sangue Humano (SH), Sangue Animal (SA) e Falso Positivos Comuns (FPC).

As amostras de sangue animal e falsos-positivos comuns depositadas em cada substrato foram divididas em conjunto treinamento (70% das amostras) e validação externa (30% das amostras). O número de amostras de sangue humano no conjunto de treinamento variou de acordo com o substrato, uma vez que elas foram selecionadas de acordo com a análise de componentes principais. Os gráficos dos scores foram analisados para verificar a presença de *outliers* e selecionar as amostras com grande variabilidade para compor o conjunto de treinamento, deixando as amostras de sangue de dois doadores para compor o conjunto de validação externa para cada tipo de substrato. Na PCA, foi utilizado o método da validação cruzada total. Os mesmos conjuntos de treinamento e de validação externa foram utilizados para realização do LDA, SIMCA e PLS-DA. Na aplicação do método LDA, 66% das amostras do conjunto treinamento foram utilizados para a calibração e 34% para a validação interna.

Os dois métodos de seleção de variáveis foram aplicados (GA e SPA) com o objetivo de reduzir a dimensão dos dados e minimizar colinearidades. O algoritmo SPA foi programado para selecionar no mínimo 3 variáveis. No algoritmo GA foram realizadas 100 gerações com 200 cromossomos em cada uma e os resultados da mutação e do cruzamento foram mantidos em 10% e 60%, respectivamente. O processo de seleção de variáveis foi repetido 10 vezes para reduzir flutuações e ajustar a função de custo, sendo que o melhor resultado para função de custo definiu as variáveis a serem utilizadas.

O PLS-DA foi a técnica de classificação escolhida para a segunda etapa desta pesquisa (modelos cruzados), sendo que PCA também foi utilizada para remoção de *outliers*. As amostras provenientes de dois tipos de substratos foram utilizadas para constituir um modelo único. Assim, foram construídos 6 modelos no total, fazendo a combinação em pares das amostras preparadas nos diferentes substratos.

Neste caso, as amostras de FPC e SA foram novamente separadas em conjunto de treinamento e conjunto de validação externa na proporção de 70% e 30%, respectivamente. No entanto, as amostras de SH utilizadas para compor o conjunto de predição/validação externa foram extraídas dos dois pisos de cada modelo, sendo as amostras correspondentes a um doador diferente em cada piso.

O software utilizado para realizar o tratamento de dados foi o Matlab (MATLAB® R2010a 7.10.0.499, MathWorks) e os modelos PLS-DA foram executados para todos os dados utilizando o PLS_Toolbox (Eigenvector Research, Inc).

3.3.5 Pré-processamento das imagens e construção dos modelos de tecidos

As imagens hiperespectrais NIR sofrem influência da superfície irregular das amostras, sendo o espalhamento da radiação e o ruído espectral consequências comuns nesse tipo de dados. Os pixels e comprimentos de onda anômalos são outros dois efeitos matemáticos bastante recorrentes em HSI-NIR. Suavização com filtro Savitzky-Golay (janela variando de 7 a 21 pontos, polinômio de segunda ordem), derivada Savitzky-Golay de 1ª e 2ª ordem, SNV, MSC e GLSW foram as técnicas de pré-processamento testadas para corrigir ou suavizar os efeitos físicos na matriz de espectros.

A seleção de regiões de interesse (RoI, do inglês *Region of Interest*) foi utilizada para obtenção de pixels/espectros das imagens de cada tipo de amostra. Os pixels foram escolhidos aleatoriamente dentro das RoI selecionadas no centro das manchas de cada amostra. Os limites das áreas de cada mancha foram determinados pela inspeção visual das imagens de cada amostra, considerando a reconstrução das imagens em função do valor médio dos pixels. Os espectros/pixels selecionados foram utilizados para construir modelos hierárquicos de classificação baseados na fusão de técnicas quimiométricas bastante conhecidas como PCA e PLS-DA.

Os modelos hierárquicos de fusão de técnicas quimiométricas foram construídos de modo que permitisse que as amostras preparadas em um tecido (ex.: tecido vermelho) fossem mantidas fora de todas as etapas de treinamento, que inclui avaliação dos espectros através de PCA e PLS-DA, para servir de conjunto de validação externa. Portanto, foram construídos modelos

hierárquicos combinando as amostras de sangue dos demais tecidos (ex. preto, branco, bege e estampado) com um mesmo tipo de fibra (algodão ou sintético).

3.4 RESULTADOS E DISCUSSÃO PARA IDENTIFICAÇÃO DE MANCHAS DE SANGUE

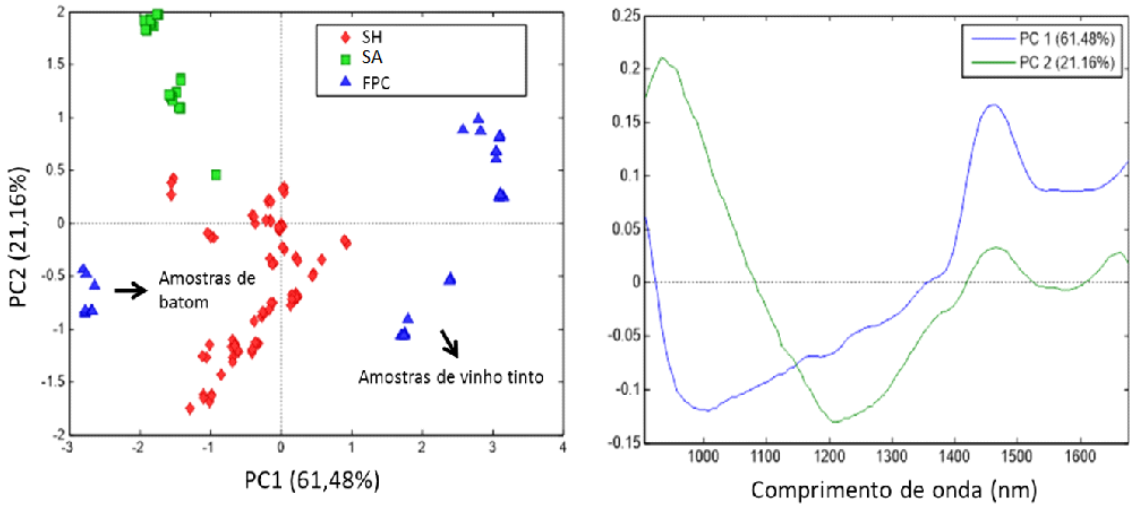
3.4.1 Modelos construídos para as manchas depositadas em substratos individuais

Inicialmente, os espectros das amostras de sangue humano, sangue animal e de falso-positivos comuns preparadas na placa de porcelana foram usadas para avaliar a melhor combinação de técnicas de pré-processamento espectral. Para isso, os espectros brutos e pré-processados foram constantemente comparados para observar os efeitos de cada pré-processamento nos espectros. Além disso, PCA foi aplicada aos dados depois de cada pré-processamento, com o objetivo de observar a distribuição das amostras. Os gráficos de *scores* da PCA foram então utilizados para verificar a existência de padrões e/ou agrupamentos, além da detecção e remoção de amostras anômalas (*outliers*). Em seguida, os modelos SIMCA foram construídos para determinar o pré-processamento que possibilitasse a discriminação mais efetiva das amostras das três classes pré-estabelecidas: Sangue Humano, Sangue Animal e Falso-Positivos Comuns.

Seguindo a metodologia descrita, observou-se que a combinação de SNV com a normalização pela faixa e centralização na média foi mais eficiente para minimizar o efeito de espalhamento da radiação e permitiu classificar corretamente (95% de confiança) 100% das amostras de SH e FPC, sendo que apenas 3 amostras de SA não foram classificadas em nenhuma das classes. Os gráficos de *scores* da PC1 versus PC2 e de *loadings* para as amostras de SH, SA e FPC depositadas na placa de porcelana após o pré-processamento podem ser vistos na Figura 6. Observando o gráfico dos *scores,* verifica-se que cada classe formou um agrupamento distinto em relação às demais, mas as amostras de FPC apresentaram uma distribuição muito mais dispersa que as amostras das outras classes. As amostras de batom têm valores de *scores* negativos na PC1, enquanto que as demais amostras de FPC apresentam valores fortemente positivos de *score*. Estas amostras foram bastante influenciadas pelas absorções na região de 900 nm a 1300 nm, como pode ser observado no gráfico de *loadings* da PC1. As demais amostras de FPC foram influenciadas pela banda em ~1460 nm, com valores positivos para os *loadings*. Esta banda pode ser atribuída ao primeiro sobretom de estiramento $O-H$ de álcool ou ácido carboxílico, provavelmente presentes

dentre os compostos da degradação das amostras de vinho. A dispersão das amostras nos escores da PC2 é resultante, principalmente, da banda de água por volta de ~970 nm e a banda em ~1210 nm proveniente dos sobretons de estiramento $C-H$ (KUENSTNER; NORRIS; MCCARTHY, 1994)

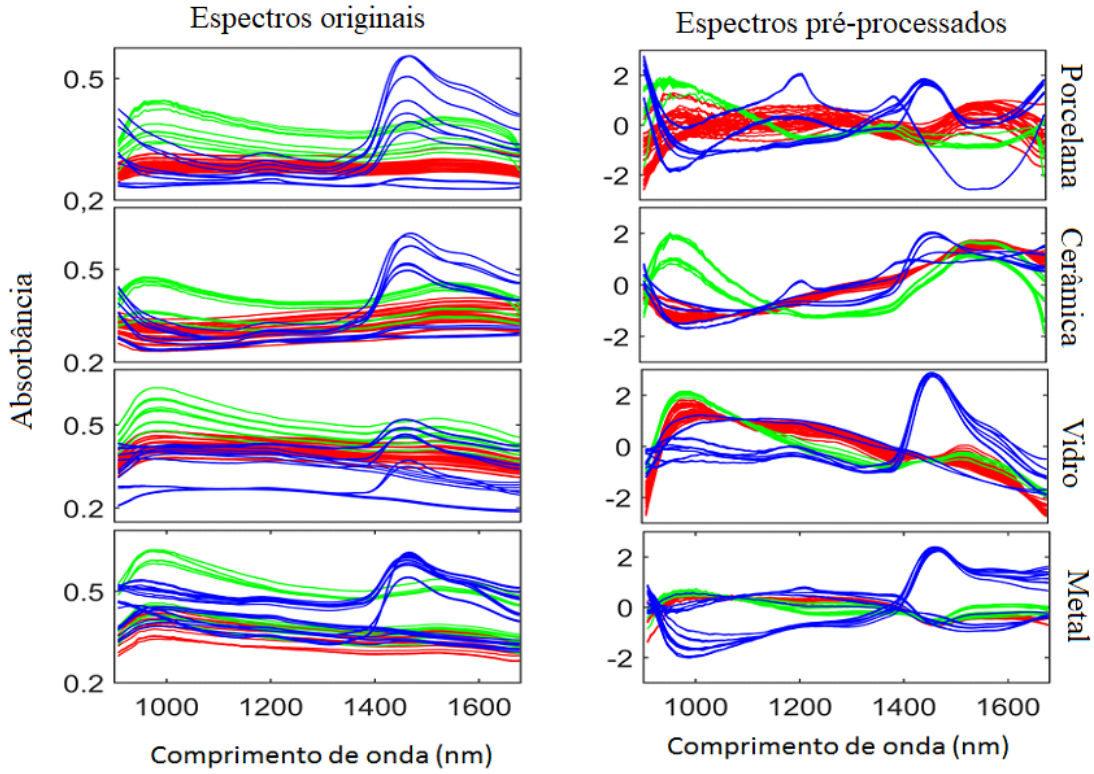
Figura 6 – gráfico dos *scores* e *loadings* da PCA das amostras de sangue humano (vermelho), sangue animal (verde) e falso-positivos comuns (azul) depositadas no substrato de porcelana após pré-processamento com SNV e normalização pela faixa.



Fonte: Do autor (2019).

Após selecionar o melhor pré-processamento, o mesmo método foi aplicado para as amostras nos diferentes substratos. A Figura 7 apresenta os espectros brutos e pré-processados para as amostras de SA, SH e FPC nos diferentes substratos.

Figura 7 – Espectros originais (coluna da esquerda) e espectros pré-processados (coluna da direita) de todas as amostras de treinamento em diferentes substratos. Espectros de SH (vermelho), espectros de SA (verde) e espectros de FPC (azul).



Fonte: Do autor (2019).

A significativa variação da absorção espectral no SH, SA e FPC mostra a grande dependência do espectro em função da absorção dos substratos, especialmente na faixa entre 900 nm a 1300 nm. Apesar da dificuldade de interpretação e identificação de bandas na região NIR, em torno de 930 nm se encontram as vibrações do terceiro sobretom de estiramento $-CH$ possivelmente relacionadoas à oxihemoglobina, e por volta de 1570 nm encontram-se as vibrações do primeiro sobretom do estiramento $N-H$ dos grupos amida presentes na proteína albumina característica no sangue (EDELMAN, GERDA; VAN LEEUWEN; AALDERS, 2012; KUENSTNER; NORRIS; MCCARTHY, 1994; WORKMAN JR; WEYER, 2012). Porém, fica claro que a região de 1300-1700 nm exibe a maioria das informações espectrais referentes ao sangue, uma vez que é a região em que o SH e o SA são mais parecidos entre si e diferentes dos FPC.

Utilizando o mesmo procedimento desenvolvido para o porcelanato, foram construídos modelos SIMCA para os demais substratos. O percentual de classificação correta para as amostras preparadas sobre o metal, a cerâmica e o vidro, foi de 80%, 90% e 100% em seus respectivos modelos (95% de confiança). Além disso, as amostras que não foram classificadas corretamente na classe a que pertence também não foram classificadas nas demais classes, o que aponta para uma especificidade igual a 1 em todos os modelos de cada substrato. Uma vez que a técnica de modelagem de classes SIMCA é um método de classificação flexível, amostras com características diferentes daquelas utilizadas para modelar as classes não são obrigatoriamente classificadas, podendo ser classificadas apenas como inconclusivas por não pertencer a nenhuma classe do modelo SIMCA. Do ponto de vista de uma investigação forense, essa característica do SIMCA é muito importante por permitir que possíveis evidências de sangue humano que não tenham sido claramente identificadas como pertencentes à classe sangue humano não sejam rejeitadas como prova. Nestes casos de amostras inconclusivas, as evidências seriam coletadas e analisadas através de outro teste confirmatório para verificar a identidade.

Com relação as amostras de sangue animal, a percentagem de classificação correta foi de 62,5% para amostras preparadas na cerâmica e de 70% para as amostras preparadas no porcelanato e no metal. O melhor resultado foi obtido para as amostras preparadas sobre o vidro que foram classificadas com 100% acerto. Este resultado pode ser justificado pela pouca quantidade de amostras de SA disponíveis para modelagem das classes em cada substrato. Assim, espera-se que aumentando o número de amostras utilizadas para criar o modelo, esse problema seja contornado. No entanto, vale ressaltar que o objetivo deste estudo foi classificar corretamente as amostras de SH e evitar que estas fossem confundidas com amostras de SA.

A Tabela 4 mostra os resultados de classificação com mais detalhes e as características dos modelos. Como pode ser observado, os modelos foram construídos usando até 3 PCs para cada classe em todos os substratos, sendo que a variância capturada foi acima de 90% para todos os casos. Este aspecto dos modelos demonstra a simplicidade e eficiência da modelagem de classes SIMCA.

Tabela 4 – Resultados e características dos modelos de classificação SIMCA mantendo 5% de significância.

	Conjuntos de dado				Classificação		
Substratos	Classes	Treinamento		Validação Externa	Modelos		
		nº de espectros	nº de PC's	nº de espectros	SH	SA	FPC
Porcelanato	SH	57 (5 doadores)	3	24 (2 doadores)	100,0%	---	---
	SA	20	3	10	---	70,0%	---
	FPC	30	2	15	---	---	100,0%
Cerâmica	SH	81 (7 doadores)	3	27 (2 doadores)	88,9%	---	---
	SA	20	2	8	---	62,5%	---
	FPC	30	3	15	---	---	93,0%
Vidro	SH	81 (7 doadores)	2	18 (2 doadores)	100,0%	---	---
	SA	20	3	7	---	100,0%	---
	FPC	30	3	15	---	---	80,0%
Metal	SH	45 (4 doadores)	3	15 (2 doadores)	80,0%	---	---
	SA	20	3	10	---	70,0%	---
	FPC	30	2	15	---	---	100,0%

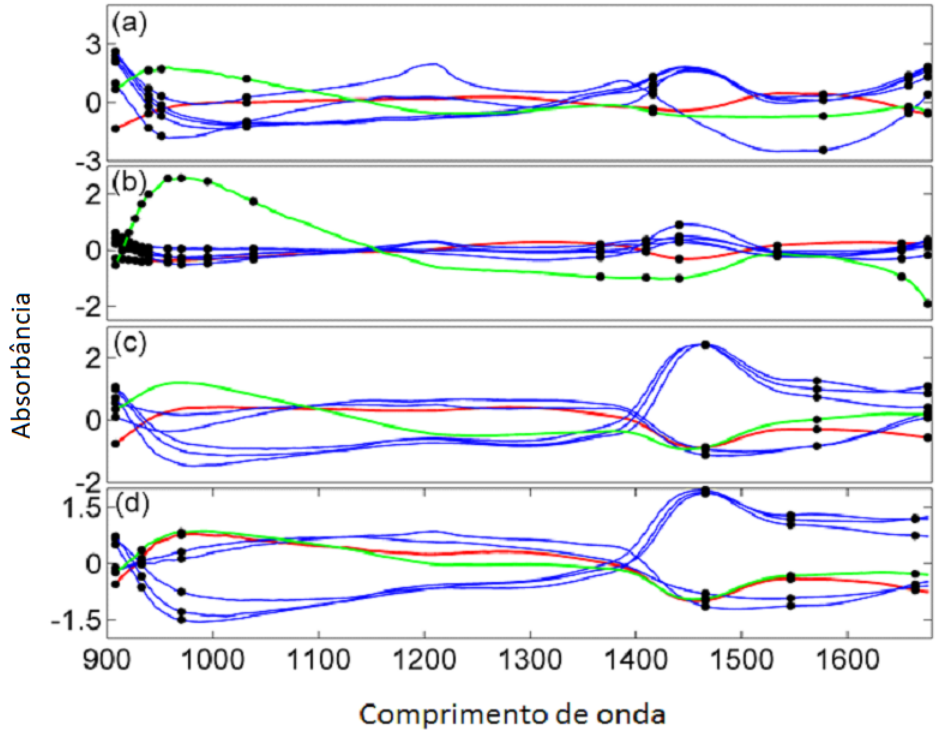
Fonte: Do autor (2019).

Os mesmos dados utilizados para a modelagem SIMCA foram também utilizados para desenvolver modelos de análise discriminante linear. Para a construção do modelo LDA, o conjunto de treinamento utilizado na modelagem SIMCA foi dividido em conjunto de calibração (66% das amostras) e de validação interna (34% das amostras), como requer o algoritmo de LDA utilizado (BOTELHO *et al.*, 2015; SILVA *et al.*, 2014). Como descrito na metodologia, o LDA necessita de uma redução do número de varáveis para ser executado apropriamente e, portanto, a seleção das variáveis tem fundamental importância no modelo final.

O algoritmo SPA selecionou diferentes números de variáveis espectrais para as amostras em cada tipo de substrato como pode ser observado na Figura 8. As oito variáveis selecionadas para os modelos construídos para as manchas em placas de porcelanato foram na região de 900 nm

à 1150 nm correspondente à vibração do terceiro sobretom $H-O-H$ da água e a vibração do terceiro sobretom do estiramento $-CH$ e $-CH2$ da oxihemoglobina (930 nm); a banda em 1577 nm corresponde à vibração do estiramento $N-H$ do grupo amida presente na proteína albumina. As dezesseis variáveis selecionadas para os modelos construídos para as amostras na placa de cerâmica contêm a banda em 1430 nm correspondente ao estiramento $H-O-H$ da água e a banda em 1530 nm provavelmente correspondente ao estiramento $N-H$ do grupo amido da albumina. Para as amostras das placas de vidro, foram selecionadas apenas quatro variáveis, sendo duas delas sem informação, uma em 1460 nm correspondente vibração do estiramento da água e a outra em 1570 nm que possivelmente corresponde a vibração do estiramento $N-H$ do grupo amida da albumina. As seis variáveis selecionadas para o modelo construído com as amostras preparadas no metal foram escolhidas na região de 900-1150 nm; em 1550 nm que pode estar relacionada com à vibração do estiramento $N-H$ do grupo amida da albumina ou o estiramento e vibração do primeiro sobretom do $O-H$ da água; e na região em torno de 1466 nm que corresponde as vibrações do $O-H$ do álcool (BOTELHO *et al.*, 2015).

Figura 8 – Espectros centrados na média de todas as amostras e variáveis selecionadas pelo algoritmo SPA em (a) porcelanato (8 variáveis), (b) cerâmica (16 variáveis), (c) vidro (4 variáveis) e (d) metal (6 variáveis). Espectros azuis para FPC, espectros vermelhos para SH, espectros verdes para SA e pontos pretos para as variáveis selecionadas.

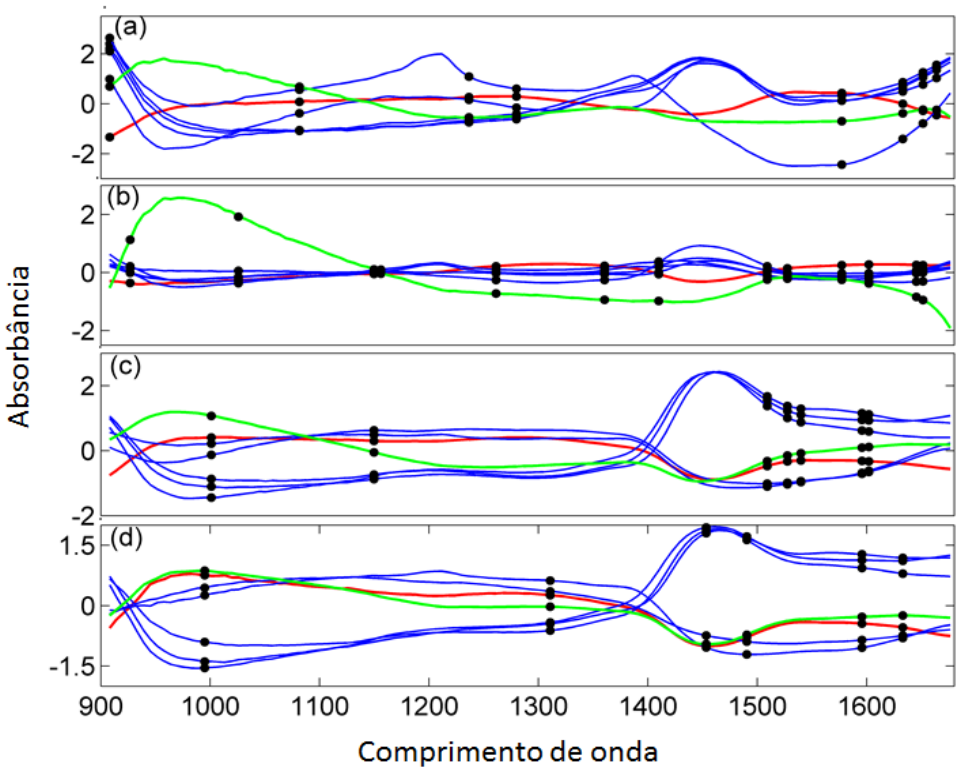


Fonte: Do autor (2019).

As vaiáveis selecionadas pelo algoritmo GA foram escolhidas quando o menor valor para função de custo foi obtido. Uma vez que o GA pode selecionar diferentes variáveis para um mesmo espectro a cada vez que é aplicado, os modelos foram construídos 10 vezes. Dessa forma, foram escolhidas as variáveis com a maior representatividade espectral das bandas e o menor valor da função custo para cada modelo GA-LDA. O algoritmo GA selecionou 8, 13, 7 e 6 variáveis para os modelos construídos para as manchas depositadas em porcelanato, cerâmica, vidro e metal, respectivamente. O modelo para as amostras preparadas no porcelanato foi construído com 5 variáveis entre 1080 nm e 1570 nm, sendo que a banda 1280 nm provavelmente corresponde às vibrações $O-H$ de álcool e as demais variáveis na banda entorno de 1650 nm correspondente ao primeiro sobretom $C-H$. No modelo para as amostras depositadas em cerâmica foram selecionadas 13 variáveis que já foram identificadas e atribuídas anteriormente no modelo SPA-LDA da cerâmica. Com relação ao modelo para as manchas preparadas no vidro, foram

selecionadas 7 variáveis, sendo 1 delas entre 1000 nm e 1150 nm correspondendo a vibração do segundo sobretom C – H e as outras 6 na região entre 1500 nm e 1700 nm provavelmente correspondentes à vibração do estiramento N – H. As 6 variáveis escolhidas para as amostras do substrato metálico foram selecionadas em 1000 nm correspondente a vibração do segundo sobretom C – H, em 1450 nm referente a vibração do estiramento da água e em 1600 nm relativo a vibração do estiramento N – H do grupo amida da proteína albumina (BOTELHO *et al.*, 2015; WORKMAN JR; WEYER, 2012). A Figura 9 apresenta os espectros médios para todas as amostras e as bandas identificadas a partir das variáveis selecionadas pelo algoritmo genético.

Figura 9 – Espectros centrados na média de todas as amostras e variáveis selecionadas pelo algoritmo GA em (a) porcelanato (8 variáveis), (b) cerâmica (13 variáveis), (c) vidro (7 variáveis) e (d) metal (6 variáveis). Espectros em azul para FPC, vermelho para SH, verde para SA e pontos pretos para identificar as variáveis selecionadas.



Fonte: Do autor (2019).

Os modelos PLS-DA e LDA, realizados com as variáveis selecionadas pelos algoritmos SPA e GA, foram utilizados para classificação das amostras do conjunto de validação externa de todos os substratos. É válido ressaltar que os mesmos conjuntos de calibração e validação interna

foram usados para os dois métodos de classificação. Os resultados dos modelos PLS-DA, SPA-LDA e GA-LDA estão resumidos na Tabela 4.

Observando a Tabela 4, é possível notar que no modelo SPA-LDA para as manchas depositadas no substrato cerâmica, uma amostra de FPC (vinho tinto) foi erroneamente classificada como SH, sendo esse resultado provavelmente uma consequência do uso de muitas variáveis selecionadas na região de 900 nm a 1300 nm onde todos os tipos de as amostras apresentam absorções similares. Esse tipo de erro implica em um falso-positivo, que é aceitável do ponto de vista forense, pois não resulta na exclusão de uma evidência real do processo de investigação. Adicionalmente, uma amostra de SH preparada no metal foi incorretamente classificada como pertencente à classe SA. Esse erro provavelmente se deve a maior parte das variáveis selecionadas não mostrar diferenças espectrais entre os dois compostos, com exceção de uma única variável localizada em torno de 1600 nm. Esse último erro de classificação (falso negativo) é preocupante uma vez que daria indício para que uma importante evidência fosse negligenciada ou até descartada.

Os modelos GA-LDA apresentaram melhores resultados que os modelos SPA-LDA, obtendo 100% de classificação correta para as amostras de validação externa. O algoritmo GA foi capaz de selecionar variáveis representativas de toda a faixa espectral (ver Figura 9) e, portanto, incluiu a maioria das bandas informativas, fato este que não ocorreu com os modelos SPA-LDA das amostras no vidro e metal.

Os modelos PLS-DA foram construídos utilizando a validação cruzada total. O número de Variáveis Latentes (LV) e o limiar (*threshold*) de classificação foram automaticamente calculados para cada modelo pelo PLS-Toolbox, que utiliza aproximação Bayesiana (PASTORE *et al*. 2011;ARMSTRONG & HIBBERT,2009). O número de variáveis latentes selecionadas para cada modelo foi 3, 2, 5 e 5 para as amostras em porcelanato, cerâmica, vidro e metal, respectivamente. Em quase todos os modelos, os valores de limite (*threshold*) para predição das amostras de sangue humano ficaram entorno de 0,5, sendo o valor de *threshold* mais alto igual a 0,8 apenas para as amostras preparadas na cerâmica. Este fato, provavelmente se deve à alta variabilidade das amostras de SA e FPC nesse caso, porém isto não prejudicou o potencial de classificação do modelo.

Os modelos PLS-DA obtiveram 100% de classificação correta para amostras de validação externa de todos os substratos, resultado similar ao obtido com GA-LDA. Como pode ser visto na Tabela 5, a sensibilidade e especificidade para os modelos PLS-DA e GA-LDA foram iguais a 1. Neste aspecto os modelos SPA-LDA para a cerâmica e para o metal apresentaram uma diminuição dos valores de sensibilidade e especificidade devido aos erros de classificação mencionados anteriormente. No caso específico dos modelos de classificação, o decréscimo da especificidade está relacionado ao decréscimo da sensibilidade.

Tabela 5 – Resultados de classificação para conjunto de validação externa usando SPA-LDA, GA-LDA e PLS-DA. Sn: sensibilidade; Sp: especificidade.

Classes por modelo		Nº de amostras	SPA-LDA			GA-LDA			PLS-DA		
			Classificação (%)	Sn	Sp	Classificação (%)	Sn	Sp	Classificação (%)	Sn	Sp
Porcelanato	SH	24	100	1	1	100	1	1	100	1	1
	SA	10	100	1	1	100	1	1	100	1	1
	FPC	15	100	1	1	100	1	1	100	1	1
Cerâmica	SH	27	100	1	0,958	100	1	1	100	1	1
	SA	8	100	1	1	100	1	1	100	1	1
	FPC	15	93,4	0,937	1	100	1	1	100	1	1
Vidro	SH	24	100	1	1	100	1	1	100	1	1
	SA	9	100	1	1	100	1	1	100	1	1
	FPC	15	100	1	1	100	1	1	100	1	1
Metal	SH	15	93,4	0,937	1	100	1	1	100	1	1
	SA	11	100	1	0,967	100	1	1	100	1	1
	FPC	15	100	1	1	100	1	1	100	1	1

Fonte: Do autor (2019).

3.4.2 Modelos construídos para dois substratos (modelos cruzados)

A escolha do PLS-DA para a construção dos modelos para as manchas depositadas em dois substratos (modelos cruzados) foi realizada após observar o excelente desempenho que a técnica

obteve na primeira etapa da pesquisa. As placas de 2 tipos de porcelanato e 3 tipos de cerâmica foram combinadas como descrito na metodologia, produzindo 6 modelos mistos.

Uma consequência dessa metodologia foi a adição de variabilidade dos pisos ao conjunto total de dados. Desse modo, além de realizada uma suavização e centrar os dados na média, o pré-processamento GLSW também foi aplicado com o objetivo de reduzir a influência das diferenças entre os dois substratos em que foram preparadas as amostras usadas para construir cada modelo e maximizar a informação correspondente às amostras. O fator α responsável por determinar o quanto de informação dos dados deve ser suprimida foi fixado em 0,02 que é o valor sugerido pelo algoritmo.

A escolha dos doadores de SH cujas amostras integram o conjunto de validação externa foi realizada de forma aleatória entre os voluntários, porém garantido que pelo menos espectros das manchas de sangue de 2 doadores estão presentes nas amostras de validação externa para cada piso. As amostras de SA e FPC foram divididas em treinamento e validação externa na proporção de 70% e 30 %, respectivamente. Os modelos foram construídos utilizando a validação interna subgrupos aleatórios (10 partições) e com número de VL variando entre 3 e 6.

De acordo com a Tabela 6, os modelos que combinam amostras provenientes de substratos de tipos diferentes (por exemplo, cerâmica x porcelanato) apresentaram sensibilidade igual a 1 e especificidade com um pequeno decréscimo apenas em um dos casos, apresentando o valor de 0,982. Por sua vez, os modelos compostos de amostras preparadas em substratos com mesma natureza (por exemplo, cerâmica x cerâmica) apresentam decréscimo nos valores de sensibilidade e especificidade na maioria dos casos. A composição do pigmento das placas pode ter influenciado no modelo, uma vez que os substratos do mesmo tipo apresentam colorações diferentes. Embora a redução de especificidade tenha sido observada, não foi detectado falso negativo. Portanto, em um caso real de aplicação forense todas as amostras de sangue humano seriam coletadas e, mesmo com a presença de poucos falso-positivos, todas as evidências importantes (manchas de SH) não seriam negligenciadas.

Tabela 6 – Resultados de classificação para o conjunto de treinamento e de validação externa. Sensibilidade (Sn), Especificidade (Sp).

Substratos	Amostras	Validação por blocos aleatórios				Predição		
		LV	Sn	Sp	Erro da classe	Sn	Sp	Erro da classe
Porcelanato1	FPC		1,0	1,0	0	1,0	0,98	0,009
x	SA	3	1,0	1,0	0	1,0	1,0	0
Cerâmica1	SH		1,0	1,0	0	1,0	1,0	0
Cerâmica 1	FPC		1,0	0,99	0,004	1,0	1,0	
x	SA	3	1,0	1,0	0	1,0	1,0	0
Porcelanato 2	SH		1,0	1,0	0	1,0	1,0	
Cerâmica 1	FPC		1,0	1,0		1,0	0,98	0,006
x	SA	3	1,0	1,0	0	1,0	1,0	0
Cerâmica 2	SH		1,0	1,0		1,0	0,98	0,007
Cerâmica 3	FPC		1,0	1,0		1,0	1,0	
x	SA	3	1,0	1,0	0	1,0	1,0	0
Cerâmica 1	SH		1,0	1,0		1,0	1.0	
Porcelanato1	FPC		1,0	1,0	0	0,962	1,0	0,019
x	SA	6	1,0	0,985	0,018	1,0	0,98	0,005
Porcelanato 2	SH		0,988	0,982	0,015	0,975	0,96	0,033
Cerâmica 2	FPC		1,0	1,0	0	1,0	1,00	0
x	SA	4	1,0	1,0	0	1,0	0,98	0,007
Cêramica 3	SH		1,0	0,981	0,010	1,0	0,952	0,024

Fonte: Do autor (2019).

3.4.3 Identificação de manchas de sangue em tecidos

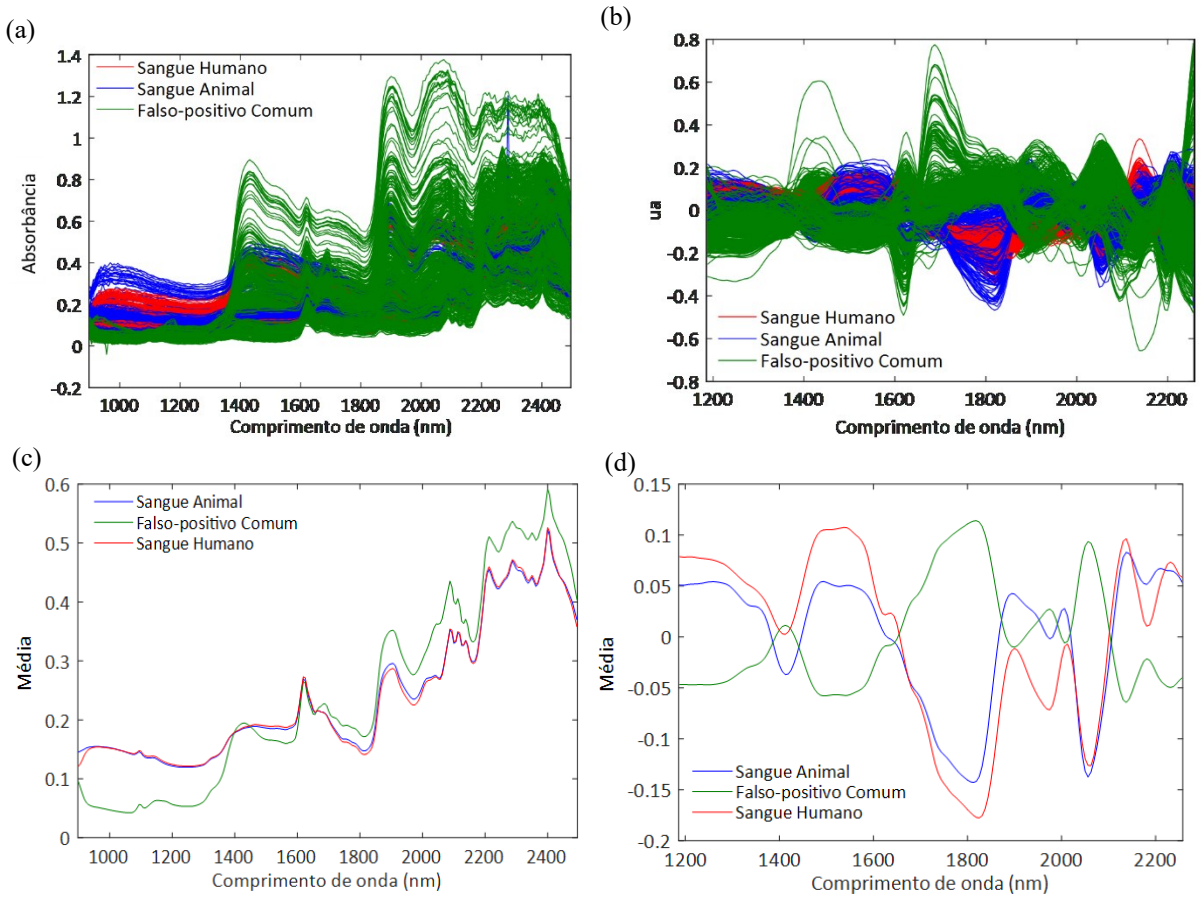
3.4.3.1 Análise dos espectros

A seleção de RoI foi aplicada em uma imagem de cada tipo de amostra (Sangue Humano, Sangue de Animal, Falso-Positivo Comum) para os dois tipos de tecido (sintético e de algodão) para definir uma área que corresponda às manchas preparadas. Em seguida, foram selecionados 50 pixels de cada um dos 7 FPC, 150 pixels de SA (75 de sangue de cavalo e 75 de sangue de cabra) e 150 pixels de SH (sangue de uma mancha escolhida aleatoriamente) em cada uma das manchas nos 10 tecido diferentes (5 tecidos sintéticos e 5 tecidos de algodão). A seleção dos pixels foi realizada aleatoriamente dentro da área das manchas para tornar a seleção o mais independente possível.

As amostras foram divididas em dois conjuntos de dados de acordo com o tipo de tecido (sintético e de algodão) para posterior construção dos modelos hierárquicos de fusão de técnicas quimiométricas. Essa divisão foi necessária devido à grande diferença entre as amostras preparadas sobre o tecido sintético e de algodão observada através de um modelo de PCA com todas as amostras. Desse modo, o pré-processamento espectral foi realizado individualmente em cada conjunto de dados mesmo eles apresentando efeitos indesejados semelhantes. Contudo, o pré-processamento mais adequado para os dois conjuntos de dados foi praticamente o mesmo. Todos os espectros apresentavam excesso de ruído espectral nos primeiros comprimentos de onda e na última parte dos espectros, além de acentuado espalhamento espectral com pouca correlação com as informações químicas das amostras. Esses efeitos físicos foram corrigidos através da redução da faixa de trabalho para 1187-2265 nm, seguida de suavização utilizando filtro Savitzky-Golay com janela de 11 pontos e polinômio de segunda ordem e SNV. Como os tecidos com mesma formulação química, algodão por exemplo, apresentam texturas diferentes resultantes das características dos fios utilizados na confecção, também foi aplicada uma técnica de pré-processamento que suprime o efeito das diferenças entre as matrizes em função do substrato. A técnica em questão foi o GLSW, que foi implementado utilizando as classes de tecidos com cores diferentes para fazer a correção e obteve um índice de supressão $\alpha = \mathbf{0,91}$ para as amostras preparadas sobre os tecidos sintéticos e $\alpha = \mathbf{1,4}$ para as amostras preparadas sobre os tecidos de algodão.

Observando a Figura 10.a, que apresenta o gráfico dos espectros para todas as amostras de treinamento preparadas nos tecidos sintéticos, fica evidente a grande variabilidade dos espectros. Além disso, os espectros apresentam forte efeito do espalhamento da radiação e ruído. Como pode ser visto na Figura 10.b, mesmo após o pré-processamento, os espectros ainda parecem bastante dispersos dentro de cada classe. Essa confusão é consequência da grande variabilidade das amostras, da técnica de aquisição dos espectros (imageamento por reflectância) e número elevado de amostras. Isso é demonstrado quando comparamos os espectros médios de cada classe antes e depois do pré-processamento (ver Figura 10.c.d), que evidencia a similaridade entre os espectros de sangue de conjuntos de imagens bastante distintas.

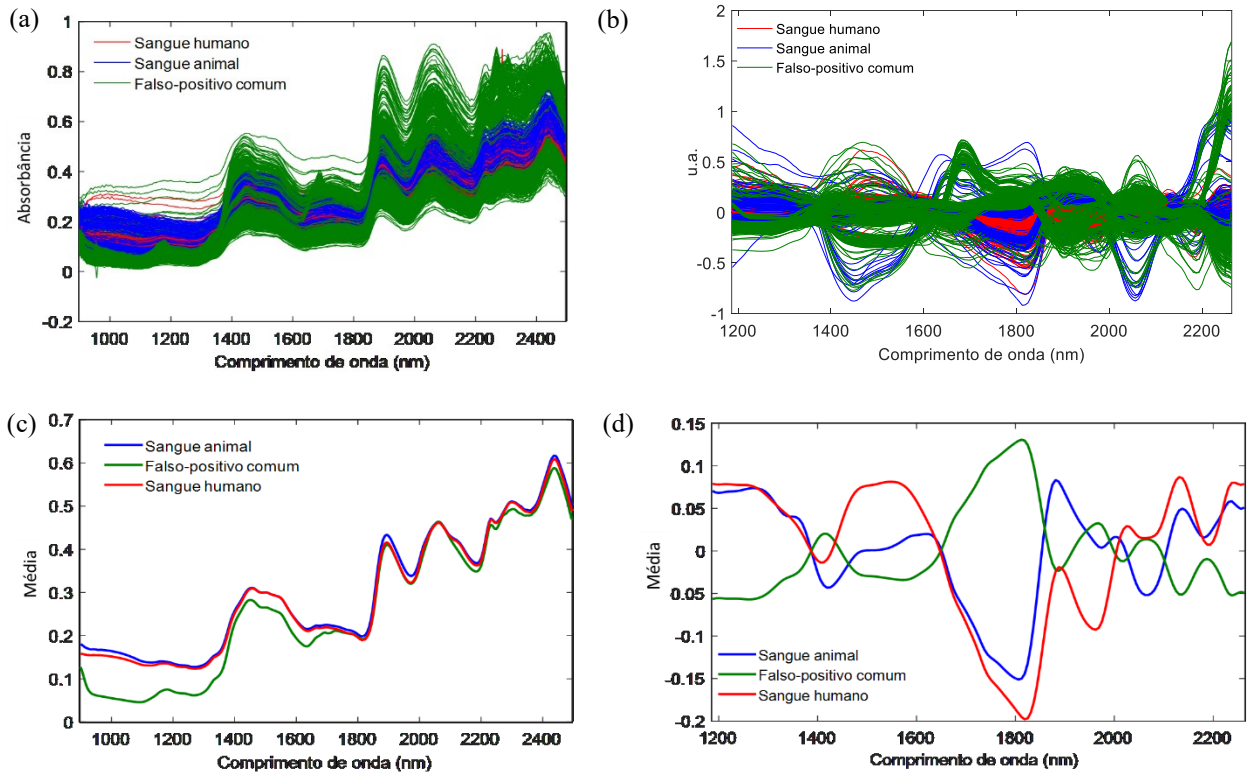
Figura 10 – Gráfico dos espectros originais (a) e espectros pré-processados (b) para as amostras preparadas sobre os tecidos sintéticos; e gráfico dos espectros originais médios (c) e espectros médios pré-processados (d).



Fonte: Do autor (2019).

O mesmo pode ser observado na Figura 11 que apresenta os gráficos dos espectros das manchas preparadas sobre os tecidos de algodão antes e depois do pré-processamento. A dispersão e ruído espectrais foram parcialmente corrigidos. Vale ressaltar que os gráficos dos espectros médios (Figura 11.c.d) não são uma boa representação para as amostras de FPC, uma vez que a variabilidade dessas amostras é muito alta e acaba sendo mascarada através do cálculo da média.

Figura 11 – Gráfico dos espectros originais (a) e espectros pré-processados (b) para as amostras preparadas sobre os tecidos de algodão; e gráfico dos espectros originais médios (c) e espectros médios pré-processados (d).



Fonte: Do autor (2019).

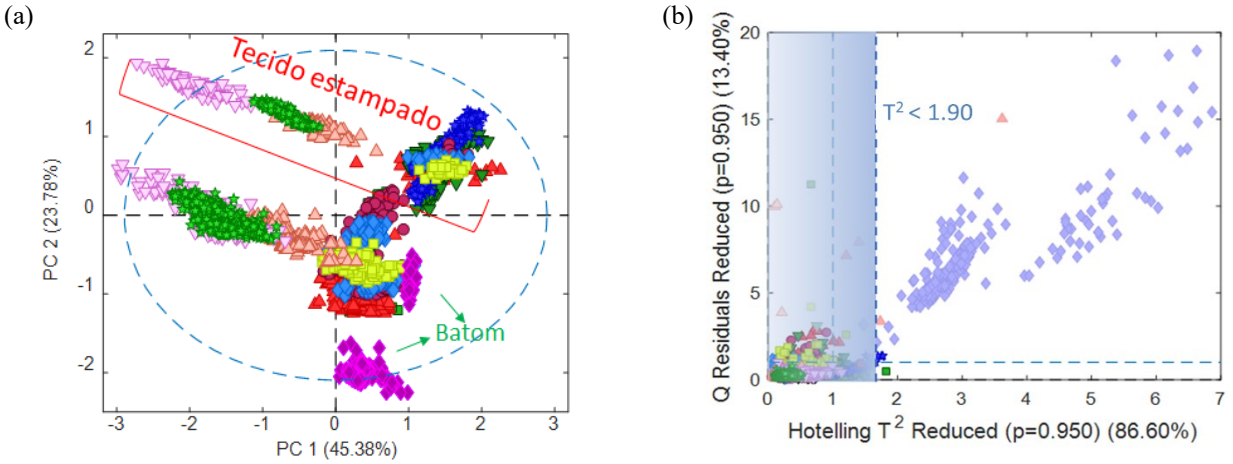
Após concluir o pré-processamento, os conjuntos de treinamento das amostras de SH, SA e FPC preparadas sobre os tecidos sintético e de algodão foram submetidos à PCA e PLS-DA. A aplicação de cada técnica quimiométrica e sua avaliação foram utilizadas como critérios para a posterior construção do modelo hierárquico de fusão de técnicas quimiométricas. A descrição detalhada para a construção de um modelo hierárquico de fusão de técnicas quimiométricas é apresentada a seguir, deixando as amostras preparadas sobre o tecido bege de fora do modelo para posterior utilização como conjunto de validação externa.

3.4.3.2 Modelos hierárquicos para manchas em tecido sintético

O primeiro passo do tratamento dos dados foi construir modelos de PCA individuais para as amostras preparadas sobre os tecidos sintéticos e para os tecidos a base de algodão com o objetivo de remover os pixels anômalos e as amostras mais diferentes dentro do conjunto total das amostras. Observando a Figura 12.a dos escores da PCA para os espectros das manchas preparadas no tecido sintético, foi possível identificar alguns *outliers* e também tendência de separação entre os espectros das manchas preparadas sobre o tecido sintético estampado e os espectros das manchas preparadas sobre os tecidos sintético lisos coloridos. Desse modo, decidiu-se remover primeiro os espectros referentes à classe "batom" e na etapa seguinte foram removidas as amostras preparadas no tecido estampado.

O modelo PCA foi recalculado sem as amostras de batom e, em seguida, essas amostras foram projetadas dentro do novo modelo de PCA com o objetivo de avaliar sua similaridade com o modelo. A Figura 12.b mostra o gráfico de resíduos para o modelo de PCA construído com as amostras restantes e a projeção das amostras excluídas (espectros do batom e *outliers*). Através deste gráfico foi possível definir um valor mínimo para T^2 de Hotelling reduzido ($T^2 < 1,9$), com o qual as amostras de batom foram mantidas fora dos limites do modelo. Esse critério para definir os limites para os modelos de PCA é similar ao utilizado na modelagem SIMCA.

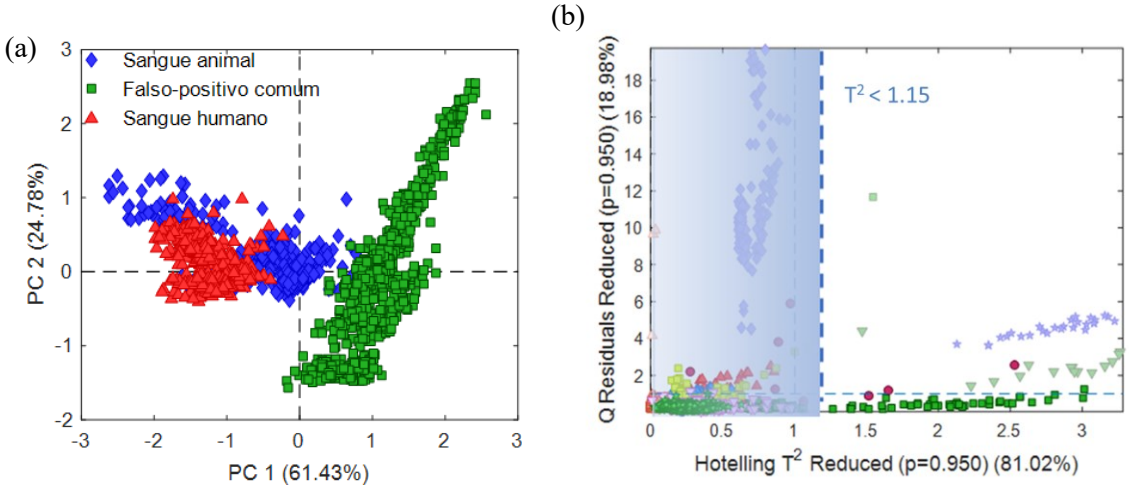
Figura 12 – Gráfico dos *scores* das duas primeiras PC's para o conjunto de espectros referentes às amostras preparadas sobre tecidos sintéticos (a) e gráficos de resíduos do modelo, incluindo a projeção das amostras de batom considerando o limite $T^2 < 1{,}90$.



Fonte: Do autor (2019).

As amostras preparadas sobre os tecidos sintéticos estampado foram removidas do conjunto de treinamento restante e outra PCA foi construída (Figura 13.a). As amostras removidas foram projetadas no modelo e foi definido um valor de $T^2 < 1{,}15$ para estabelecer os limites do modelo, deixando fora do modelo as amostras removidas antes do cálculo da PCA. Observando a Figura 13.b é possível identificar amostras excluídas dentro do limite estabelecido para essa PCA, contudo essas amostras correspondem a classe “batom” que já foram removidas na etapa anterior e, portanto, têm um filtro próprio. Vale ressaltar que esses modelos de PCA serão utilizadas como filtros ou funções de decisão para incluir/excluir amostras progressivamente no modelo hierárquico. Vale ressaltar que o limite estabelecido e as imagens recobertas pelo retângulo sombreado são edições gráficas aplicadas no gráfico de influência apenas para fins de compreensão da montagem dos modelos.

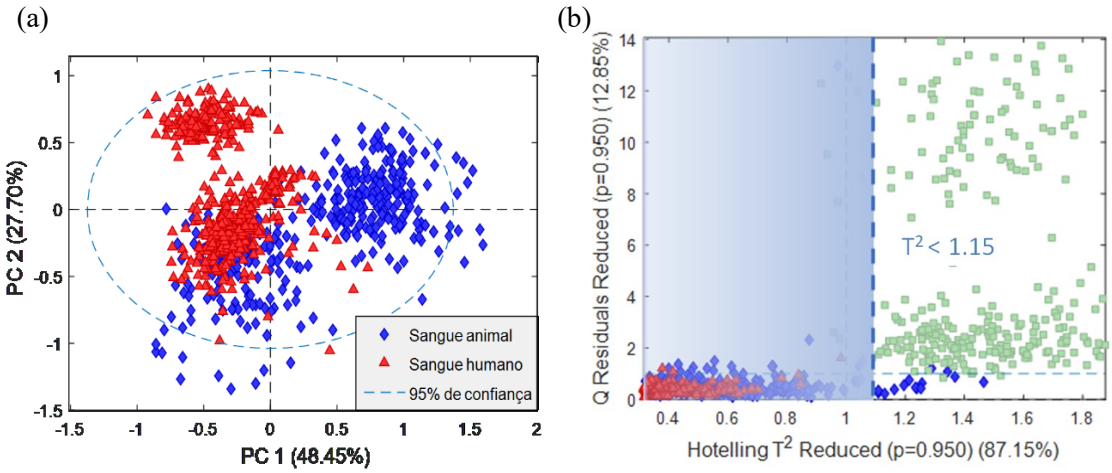
Figura 13 – Gráfico dos *scores* das duas primeiras PC's para o conjunto de amostras preparadas sobre tecidos sintéticos coloridos (a) e gráficos de resíduos do modelo, incluindo a projeção das amostras preparadas no tecido estampado e amostras de batom, considerando o limite $T^2 < 1{,}15$.



Fonte: Do autor (2019).

Após as duas PCAs, as amostras de FPC preparadas nos tecidos sintéticos coloridos foram removidas do conjunto principal e uma nova PCA foi construída apenas com as amostras de SH e SA (Figura 14.a). Os espectros de FPC removidos foram projetados neste novo modelo e o valor de $T^2 < 1{,}15$ foi estabelecido como limite do modelo para manter apenas as amostras de sangue dentro do modelo, como pode ser observado na Figura 14.b.

Figura 14 – Gráfico dos *scores* das duas primeiras PC's para as amostras de SH e SA preparadas sobre tecidos sintéticos coloridos (a) e gráficos de resíduos do modelo, incluindo a projeção das amostras de FPC excluídas e o limite estabelecido de $T^2 < 1{,}15$.

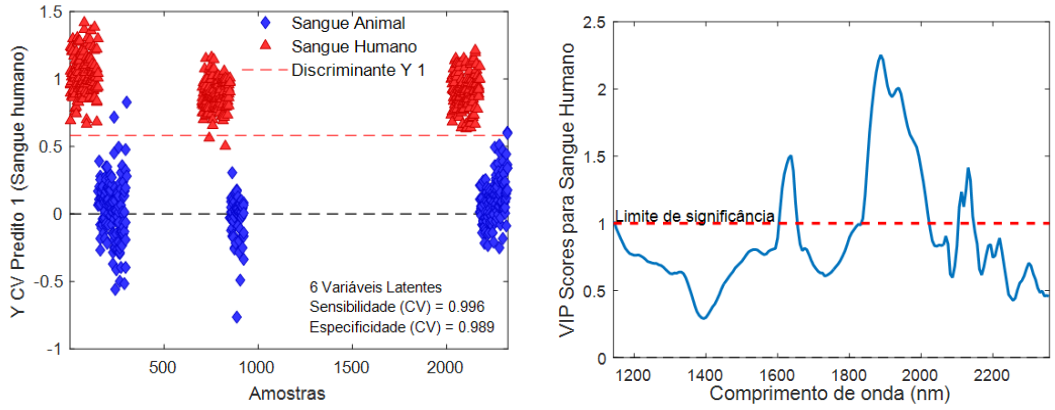


Fonte: Do autor (2019).

Como o conjunto de amostras de SA e SH são muito similares, a PCA não permitiu fazer a distinção clara entre estas amostras e por isso, os espectros restantes foram utilizados para construir um modelo PLS-DA com duas classes capaz de discriminar as amostras. O modelo PLS-DA foi construído considerando cerca de 420 espectros de SH e 400 espectros de SA, aplicando validação cruzada *venetian blinds* (com 10 subgrupos e 12% das amostras para validação). O modelo resultante apresentou alta sensibilidade (0,996) e alta especificidade (0,989) para a validação interna das amostras. A Figura 15.a apresenta os *scores* do modelo PLS-DA que denota a discriminação entre as amostras de SH e SA. A Figura 15.b, que apresenta o gráfico de influência das variáveis originais para o modelo discriminante. Vale ressaltar a região de bandas de combinação apresentou a maior influência na discriminação das amostras de sangue humano e de animal. As bandas mais influentes para discriminação entre as amostras de sangue nesse modelo PLS-DA foram: banda de combinação entre $C-N$ e $N-H$ provavelmente de amida II (~1630nm); a banda de combinação de estiramento $O-H$ e $C=O$ de ácido carboxílico (~1890nm) e banda de estiramento de $O-H$ da água (~1940 nm)(WORKMAN JR; WEYER, 2012).

Figura 15 – Gráfico dos *scores* do modelo PLS-DA com 6 variáveis latentes (sn=0,996, sp=0,989) para as amostras de SH e SA preparadas nos tecidos sintéticos coloridos (a) e gráfico dos VIP *scores* para a discriminação das duas classes (b).

(a) (b)

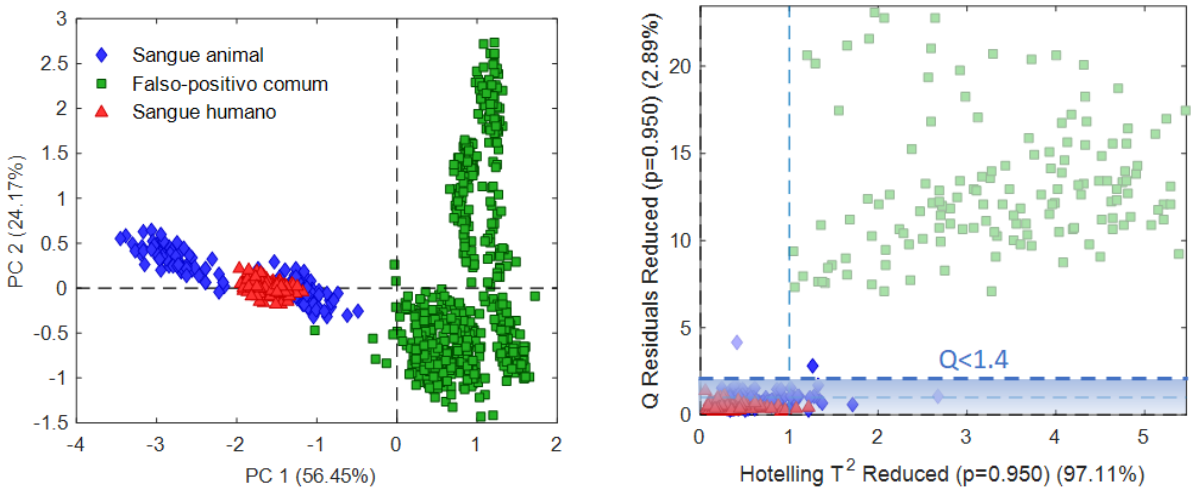


Fonte: Do autor (2019).

O próximo passo na construção do modelo foi recuperar as amostras preparadas sobre o tecido sintético estampado e prosseguir com o mesmo procedimento adotado para as amostras preparadas nos tecidos coloridos. As amostras de sangue e de FPC apresentam um comportamento bastante distinto, como pode ser visto na Figura 16.a. Assim, as amostras de FPC foram removidas e o modelo de PCA foi refeito apenas com as amostras de sangue humano e animal. Observando o gráfico de resíduos para o novo modelo de PCA (Figura 16.b), foi possível definir um valor de $Q_{residual} < 1{,}40$ para manter as amostras de FPC fora do limiar do modelo em relação a distância para o centro do modelo.

Figura 16 – Gráfico dos *scores* das duas primeiras PCs para as amostras preparadas no tecido sintético estampado (exceto batom)(a) e gráficos de resíduos do modelo PCA para as amostras de SH e SA, incluindo a projeção das amostras de FPC excluídas, considerando o limite estabelecido de Qr < 1,40.

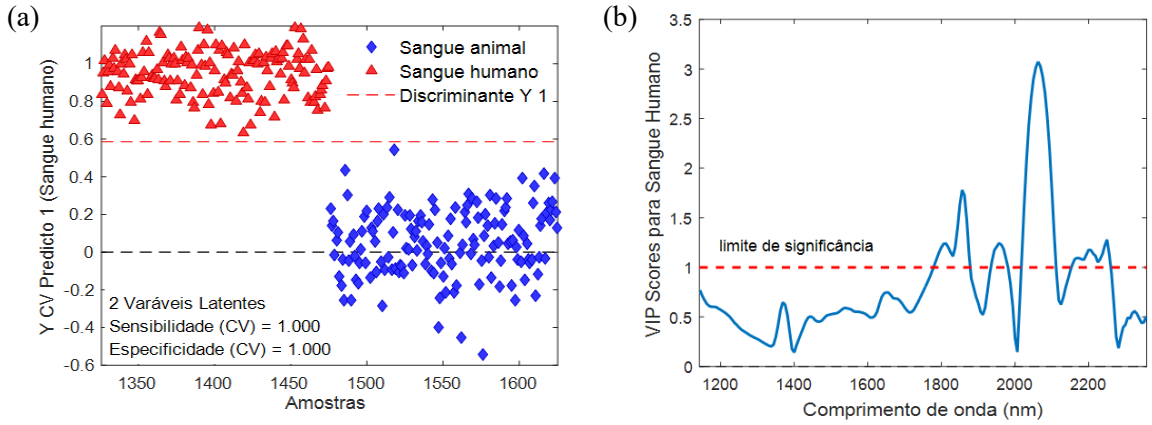
(a) (b)



Fonte: Do autor (2019).

As amostras de sangue humano e animal restantes foram utilizadas para construir um modelo PLS-DA específico para as amostras preparadas sobre tecido estampado, mantendo a mesma configuração do modelo construído anteriormente, aplicando validação cruzada *venetian blinds* (com 10 subgrupos e 12% das amostras para validação interna). Os valores de sensibilidade e de especificidade evidenciam o grande potencial da técnica PLS-DA para discriminar as amostras de SH e SA, resultando em 100% de classificação correta com apenas 2 LV (ver Figura 17.a). As variáveis mais importantes para a discriminação observada no gráfico de *scores* do modelo PLS-DA podem ser associadas à albumina, hemoglobina e globulina (ver Figura 17.b). Essas proteínas apresentam banda intensa entre 2050 nm e 2060 nm, correspondente à combinações das bandas de estiramento $N-H$ da albumina, e uma banda mais fraca entre 2160 nm e 2180 nm associada à combinação da deformação angular $N-H$, estiramento $C=O$ e estiramento $C-N$ (EDELMAN, GERDA *et al.*, 2012; WORKMAN JR; WEYER, 2012).

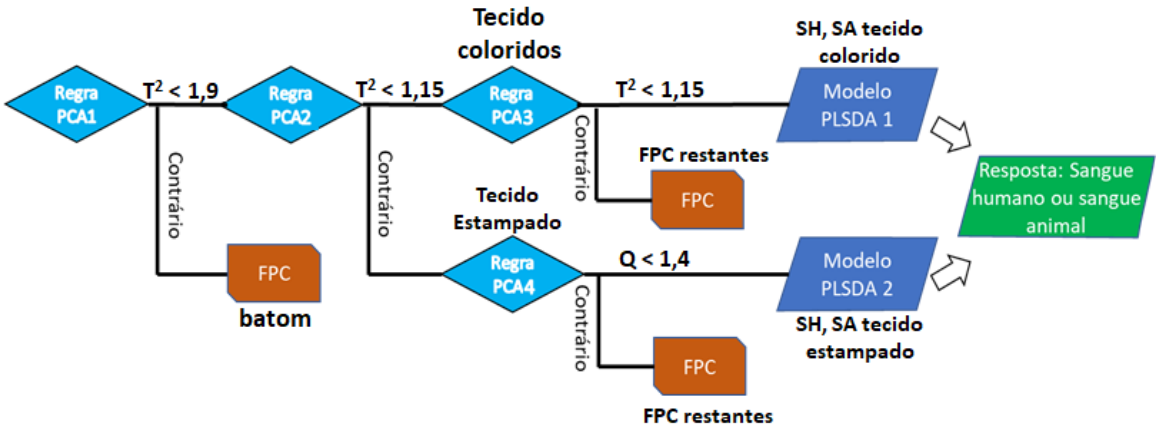
Figura 17 - Gráfico dos scores do modelo PLS-DA com 2 variáveis latentes (sn=1,000, sp=1,000) para as amostras de SH e SA preparadas no tecido sintético estampado (a) e gráfico dos VIP scores para a discriminação das duas classes (b).



Fonte: Do autor (2019).

Considerando todos os modelos PCA e PLS-DA, foi construído o modelo hierárquico de fusão de técnicas quimiométricas. O modelo hierárquico representado graficamente pela Figura 18 utiliza os modelos de PCA como funções de decisão para definir em que condições uma amostra foram considerada FPC ou não-sangue (valores limite de $Q_{residual}$ ou T^2 de Hotteling) e os modelos PLS-DA como funções de classificação que define quais amostras pertencem a classe sangue humano ou sangue animal.

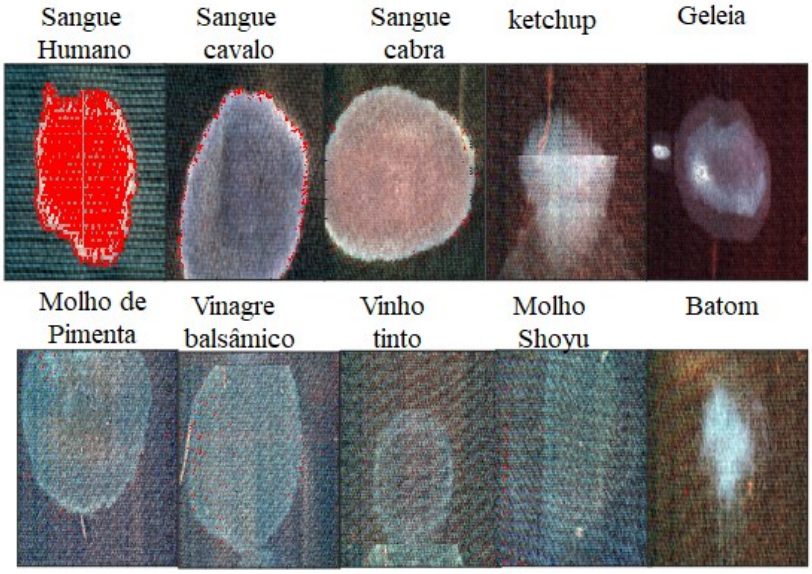
Figura 18 – Representação gráfica do modelo hierárquico de fusão de técnicas quimiométricas (PCA e PLS-DA) para as amostras preparadas nos tecidos sintéticos.



Fonte: Do autor (2019).

O modelo hierárquico construído sem as amostras preparadas no tecido bege foi utilizado para classificar os pixels das imagens de SH, SA e FPC em relação a classe SH. A Figura 19 ilustra os resultados da classificação das amostras depositadas no tecido bege (conjunto de validação externa). Como pode ser visto na Figura 19, na imagem de sangue humano praticamente todos os pixels referentes à mancha foram corretamente classificados como sangue humano. A imagem da amostra de sangue de cavalo mostra que apenas uma pequena quantidade de pixels da borda da mancha foi classificada como sendo sangue humano (Falso-positivo). Para as demais amostras de FPC e sangue de cabra, apenas poucos pixels ou nenhum pixel foram classificados como sangue humano.

Figura 19 – Modelo hierárquico construído para manchas em tecidos sintéticos - Projeção dos resultados de previsão da classe SH sobre as imagens falsas para o tecido bege.

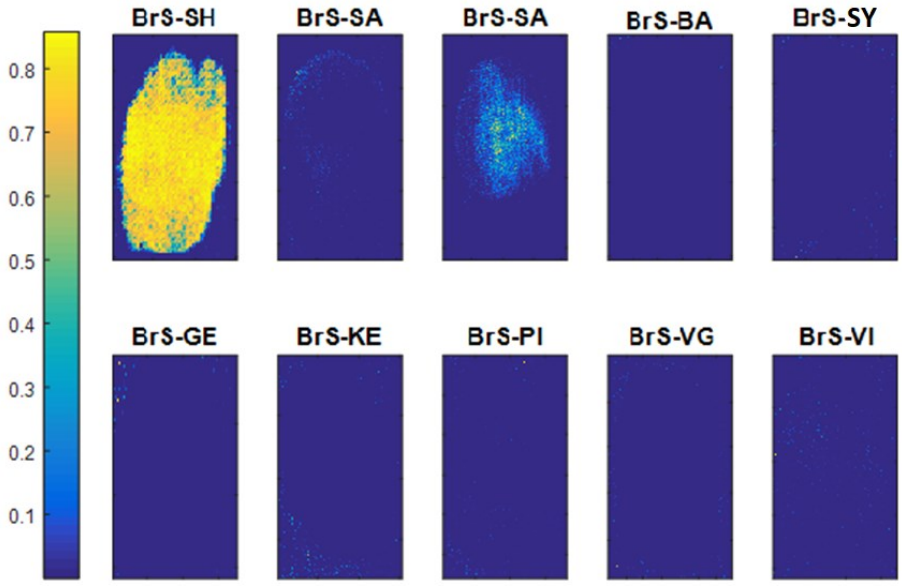


Fonte: Do autor (2019).

Esse modelo hierárquico de classificação também foi utilizado nas amostras preparadas nos tecidos que fizeram parte do processo de construção do modelo (tecido preto, branco, vermelho, estampado). Vale ressaltar que as amostras de sangue humano utilizadas para construir os modelos e as amostras de validação externa, para um mesmo tecido, não pertencem ao mesmo doador e, portanto, podem ser consideradas amostras de validação externas. A Figura 20 apresenta as imagens reconstruídas utilizando os resultados de *scores* da classificação para as amostras depositadas no tecido branco utilizando o modelo hierárquico sem o tecido bege. Os *scores* da

previsão estão expressos em termos da probabilidade de as amostras pertencerem a classe “sangue humano”, sendo que pixels com valor acima de 0,5 representam os espectros classificados como SH. Assim, as imagens referentes à sangue animal (sangue de cabra e sangue de cavalo) apresentam alguns pixels que se assemelham ao sangue humano, porem se usarmos um *threshold* igual a 0,5 não há problemas de classificação incorreta.

Figura 20 – Imagens de *scores* da previsão pelo modelo hierárquico para as amostras preparadas no tecido branco. Onde BrS representa os tecidos brancos sintéticos, SH o sangue humano, SA o sangue animal, BA o batom, SY o molho shoyu, GE a geleia, KE o ketchup, PI a pimenta, VG o vinagra balsâmico e VI o vinho tinto.



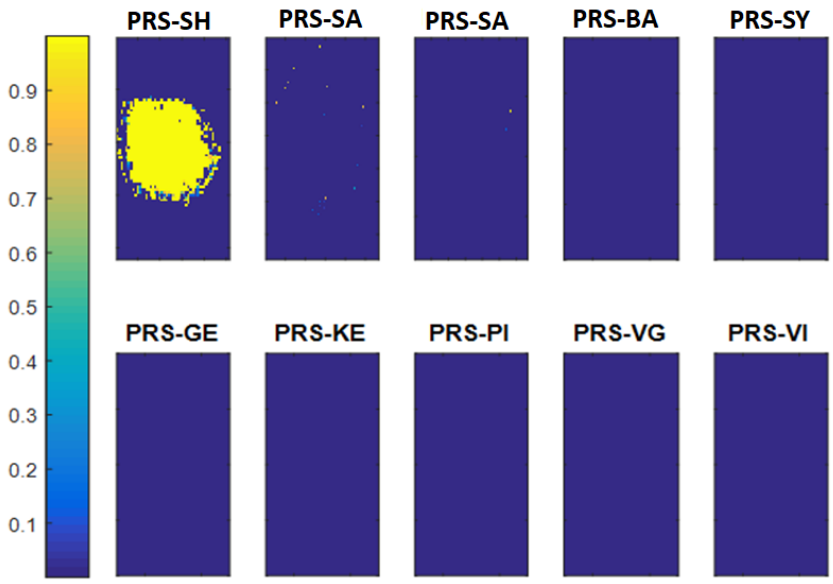
Fonte: Do autor (2019).

O mesmo procedimento de construção desse modelo hierárquico, que não incluiu no conjunto de treinamento as amostras preparadas sobre o tecido bege, foi adotado para construir modelos hierárquicos excluindo do conjunto de treinamento as manchas nos demais tecidos coloridos (vermelho, preto, branco) um a um, de forma a usá-las como conjunto de validação externa. Vale ressaltar que mesmo utilizando manchas de sangue de diferentes doadores, elas foram preparadas no mesmo tipo de tecido e por isso foram mantidas externas todas as amostras preparadas em um tecido por vez para construir os modelos e garantir uma validação totalmente externa. Apenas no caso do tecido estampado sintético não foi possível mantê-lo externo ao modelo

por que só tínhamos um único tipo de tecido estampado, porém manchas de sangue de doadores diferentes foram utilizadas na construção e na validação externa do modelo hierárquico.

A Figura 21 apresenta os resultados de classificação para as amostras preparadas em tecido estampado. Nesse caso o modelo foi ainda mais confiável, uma vez que apenas as amostras de SH e SA apresentam pixels com probabilidade de serem classificados como SH. No exemplo da Figura 20 alguns poucos pixels de FPC apresentam probabilidade de serem classificados como SH principalmente em decorrência da informação dos tecidos. Isso fica claro quando observado que esses pixels de FPC na verdade pertencem ao sinal de fundo da imagem, ou seja, representam os tecidos.

Figura 21 – Imagens de *scores* da previsão pelo modelo hierárquico para as amostras preparadas no tecido sintético estampado. Onde PRS representa os tecidos estampados sintéticos, SH o sangue humano, SA o sangue animal, BA o batom, SY o molho shoyu, GE a geleia, KE o ketchup, PI a pimenta, VG o vinagra balsâmico e VI o vinho tinto.



Fonte: Do autor (2019)

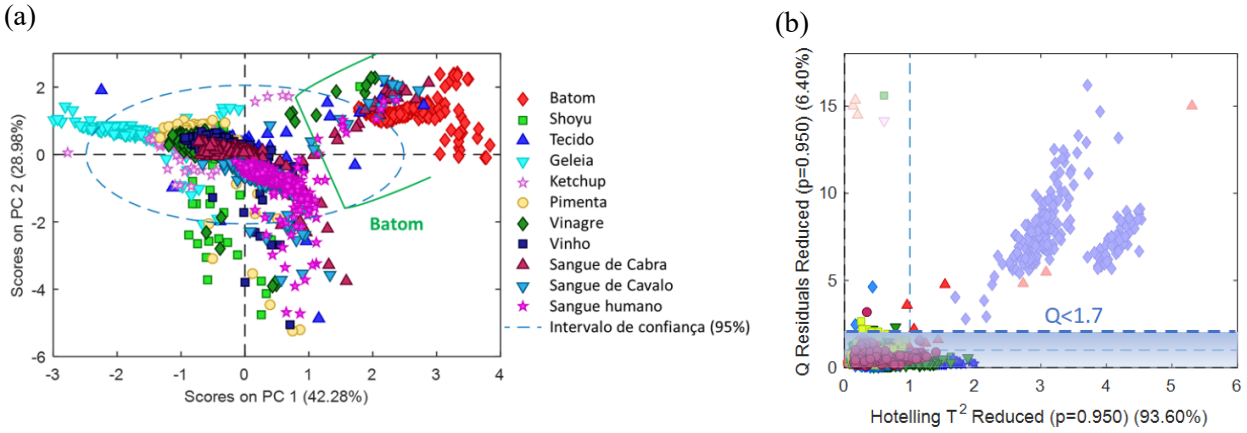
Os mesmos resultados se reproduziram para os demais tecidos sintéticos, nos quais as amostras de sangue humano foram corretamente classificadas e, embora pixels aleatórios tenham sido classificados como SH, estes não configuram falso-positivos por não apresentarem o padrão de uma mancha e pelo fato da maioria deles serem referentes ao tecido.

3.4.3.3 Modelos hierárquicos para manchas em tecido de algodão

Modelos hierárquico por fusão de técnicas quimiométricas foram construídos para as amostras preparadas sobre os tecidos de algodão utilizando o mesmo procedimento utilizado para a construção do modelo das amostras preparadas no tecido sintético. O primeiro passo foi construir um modelo PCA para todas as amostras para verificar a presença de grupos distintos e *outliers*. Como pode ser observado na Figura 22.a, as amostras preparadas sobre tecido estampado apresentam o mesmo padrão das amostras preparadas nos tecidos coloridos e essa é a principal diferença entre os modelos hierárquicos para amostras do tecido sintético e de algodão.

Com relação as amostras da classe "batom", observa-se que elas se destacam do conjunto principal das amostras de SH, SA e de alguns FPC. Assim, foram removidas as amostras de batom e de algumas amostras que também se destacaram do conjunto principal e uma nova PCA foi construída. As amostras removidas foram projetadas no gráfico de resíduos da PCA e foi definido um valor de Qresidual < 1,7 para manter as amostras excluídas fora dos limites do modelo (ver Figura 22.b).

Figura 22 – Gráfico dos *scores* da PCA para os espectros referentes à todas as amostras preparadas sobre tecidos de algodão (a) e gráfico de resíduos do modelo PCA reconstruído sem as amostras excluídas (batom e *outliers*), incluindo a projeção das amostras excluídas, considerando o limite estabelecido de Qr < 1,70.

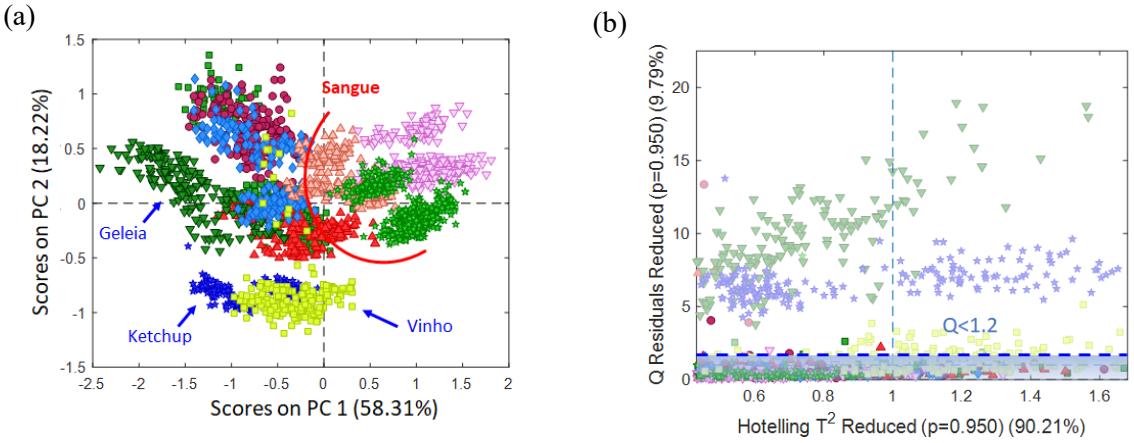


Fonte: Do autor (2019)

O mesmo procedimento foi adotado para remover as amostras de geleia, ketchup e vinho tinto que se destacaram do grupo das demais amostras onde se encontravam as amostras de sangue, como pode ser visto na Figura 23.a. Embora a separação seja observada nos *scores* de PC1 x PC2,

quando as amostras foram removidas e a PCA foi recalculada, a projeção das amostras de vinho no gráfico dos resíduos dessa nova PCA não apresentou uma separação clara em relação ao modelo (ver Figura 23.b). A utilização de um valor de Qresidual <1,2 para manter as amostras excluídas fora dos limites do modelo foi eficiente para praticamente todas as amostras, porém uma pequena parte das amostras de vinho permaneceram dentro deste limite.

Figura 23 – Gráfico dos *scores* da PCA para os espectros referentes às amostras restantes preparadas nos tecidos de algodão (a) e gráfico de resíduos do novo modelo PCA reconstruído após excluir as amostras de vinho, geleia e ketchup, incluindo a projeção das amostras excluídas e considerando o limite estabelecido de Qr < 1,20 para as amostras restantes.

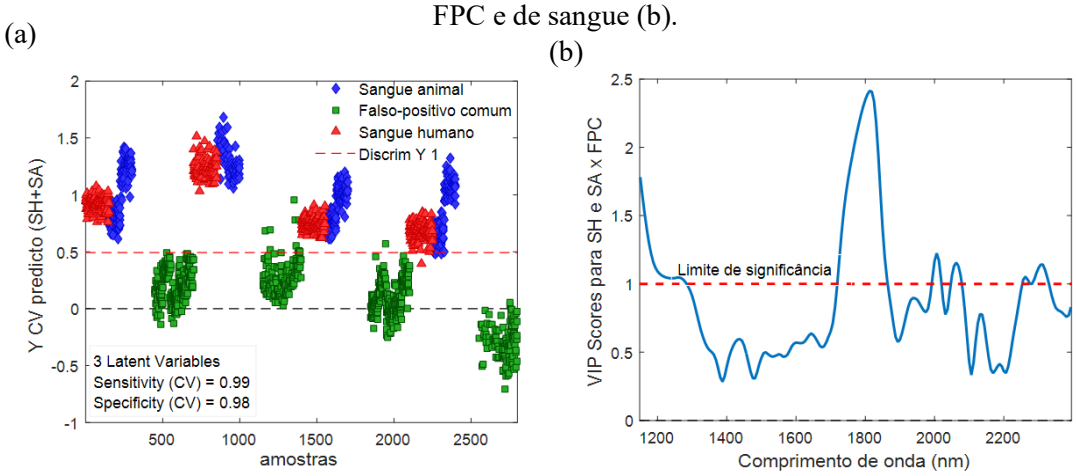


Fonte: Do autor (2019).

A dificuldade de usar PCA como critério para discriminar amostras ficou clara, mostrando não ser possível diferenciar as demais amostras de FPC das amostras de SH e SA. Assim, um modelo PLS-DA foi construído com as amostras de FPC compondo uma classe e as amostras de SH e SA juntas compondo a outra classe. Esse modelo foi construído considerando 3 LV e a validação interna foi realizada pelo método de validação cruzada customizada, em que cada tecido foi mantido externo ao modelo de calibração e em seguida predito pelo modelo construído com as demais. Assim cada bloco foi removido e predito no modelo construído com as demais amostras e o erro quadrático médio foi calculado para a avaliação do modelo PLS-DA construído (Figura 24.a). A especificidade e sensibilidade calculadas para o modelo foram iguais a 0,98 e 1,00, respectivamente, demonstrando a eficiência de discriminar os FPC das amostras de sangue utilizando a técnica de reconhecimento de padrão supervisionada.

Observando a Figura 24.b do gráfico de VIP *scores*, foi possível identificar as bandas mais importantes para a discriminação das amostras de sangue e FPC. Deste modo, o gráfico de VIP *scores* do modelo permite fazer a interpretação do resultado da classificação, em função da composição química das amostras. A banda larga entre 1720 nm e 1840 nm é a mais influente para a discriminação e provavelmente caracteriza os FPC pois essa região apresenta sinais de estiramento $C-H$ (1740 nm), banda de combinação de $O-H$ de água (1790 nm) e também possível combinação de estiramento $O-H/C-H$ (1820 nm). As bandas menos significativas em 2070 nm e 2210 nm podem estar relacionadas com estiramento $N-H$ de proteínas e deformação angular de $C-H$ de lipídios ou ácidos graxos (WORKMAN JR; WEYER, 2012).

Figura 24 – Gráfico dos *scores* da classificação PLS-DA para as amostras de SH e SA preparadas nos tecidos de algodão (a) e gráfico dos VIP *scores* para a discriminação das amostras de FPC e de sangue (b).



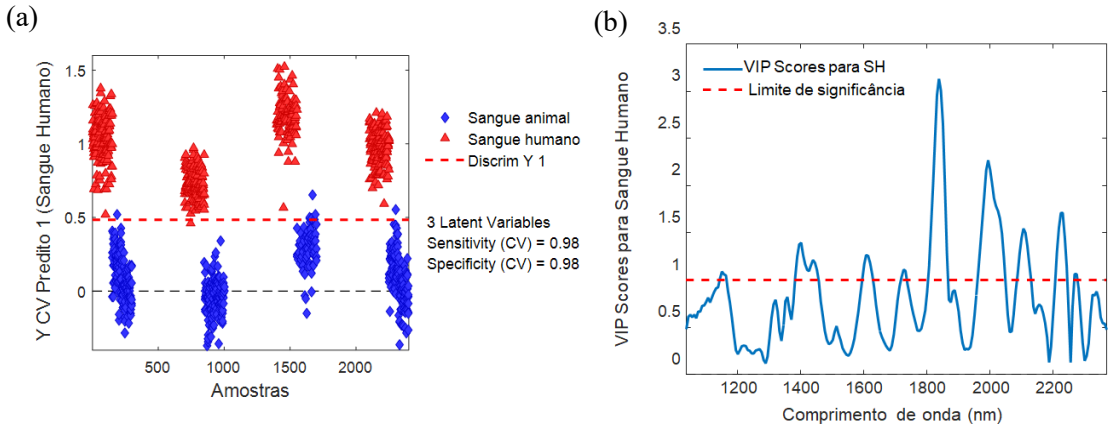
Fonte: Do autor (2019).

O próximo passo na construção do modelo hierárquico foi desenvolver um modelo para discriminar as amostras de SH e SA restantes. Novamente um PLS-DA foi construído como a mesma configuração do modelo anterior, mas considerando apenas as classes SH e SA. A Figura 25.a, que mostra o gráfico dos *scores* do PLS-DA, indica que as amostras referentes a cada tecido se agrupam, mas a diferença entre as amostras de classes diferentes é mais significativa e resulta numa discriminação com alta especificidade (98%).

As variáveis significativas para a separação das classes SH e SA podem ser observadas na Figura 25.b que apresenta os VIP *scores* para o modelo PLS-DA. Em 1400 nm observa-se a banda

de estiramento C – H e/ou banda O – H de água. Entre 1990 nm e 2020 nm é possível identificar bandas de estiramento N – H e combinação N – H, provavelmente relacionadas com as proteínas do sangue. Além dessa banda larga, foi possível identificar a banda em 2110 nm correlacionada com estiramento N – H de proteínas. Várias outras bandas contribuíram para a discriminação das amostras, porém são difíceis de atribuir devido a sobreposição das informações.

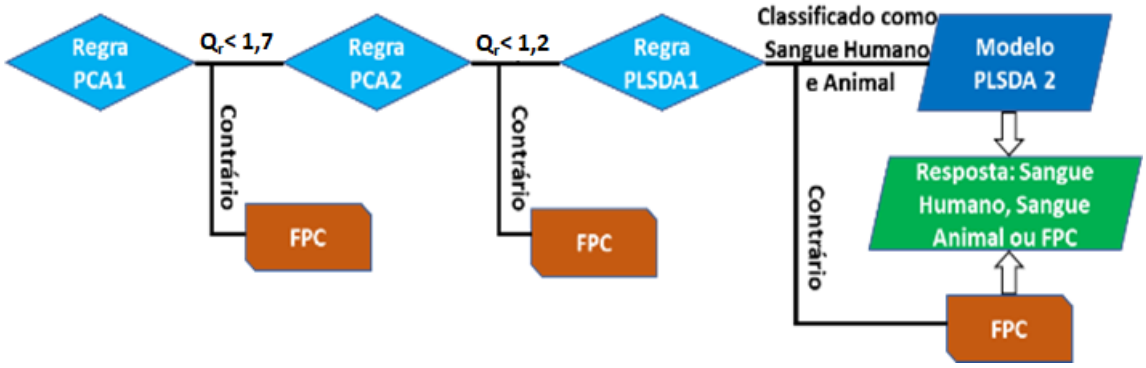
Figura 25 – Gráfico dos *scores* da classificação PLS-DA para as amostras de SH preparadas nos tecidos de algodão (a) e gráfico dos VIP scores para a discriminação das amostras de SH e SA (b).



Fonte: Do autor (2019).

Os modelos de PCA e PLS-DA construídos utilizando as amostras preparadas sobre os tecidos de algodão foram associados para criar o modelo hierárquico de classificação. O modelo hierárquico é representado graficamente pela Figura 26, onde faz uso dos modelos de PCA como funções de decisão para definir em que condições uma amostra foi considerada FPC ou "não-sangue", além da própria função discriminante de um modelo PLS-DA para discriminar os FPC mais similares ao sangue. O segundo modelo PLS-DA foi utilizado como funções de classificação que define quais amostras pertencem a classe sangue humano ou sangue animal.

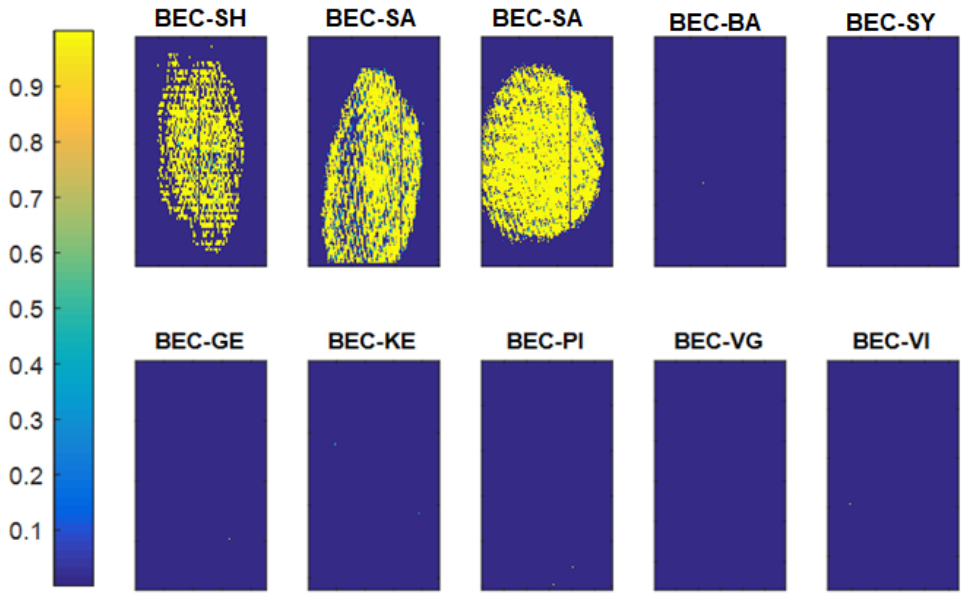
Figura 26 – Representação gráfica do modelo hierárquico de fusão de técnicas quimiométricas (PCA e PLS-DA) para as amostras preparadas nos tecidos de algodão.



Fonte: Do autor (2019).

O modelo hierárquico construído para as amostras preparadas nos tecidos de algodão, deixando as amostras preparadas no tecido bege para validação externa, foi utilizado para classificar os pixels das imagens de SH, SA e FPC em relação a classe SH. A Figura 27 ilustra os resultados da classificação das amostras depositadas no tecido bege (conjunto de validação externa). Como pode ser visto na Figura 27, na imagem de sangue humano praticamente todos os pixels referentes a mancha foram corretamente classificados como sangue humano. As amostras de sangue animal também foram classificadas como sangue humano, o que configura falso-positivo. As amostras de FPC não apresentaram problemas de classificação, tendo apenas poucos pixels aleatórios com baixa probabilidade de serem SH.

Figura 27 – Imagens de *scores* da previsão pelo modelo hierárquico para as amostras preparadas no tecido bege de algodão. Onde BEC representa os tecidos de algodão na cor bege, SH o sangue humano, SA o sangue animal, BA o batom, SY o molho shoyu, GE a geleia, KE o ketchup, PI a pimenta, VG o vinagra balsâmico e VI o vinho tinto.

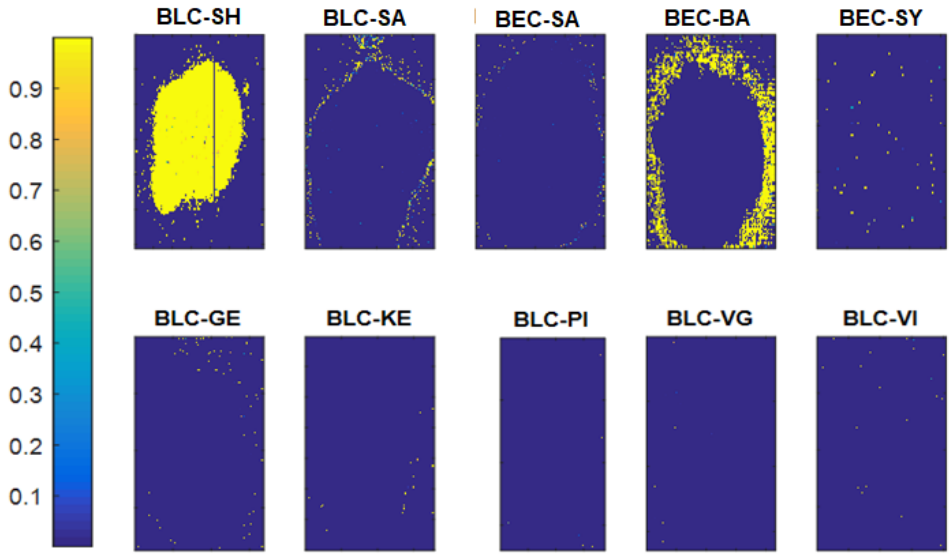


Fonte: Do autor (2019).

Embora tenha havido classificação incorreta de amostras de sangue animal, esse resultado ainda é útil do ponto de vista forense, uma vez que todas amostras de sangue foram identificadas e não seriam negligenciadas durante uma coleta de amostras aplicando esse método desenvolvido. A previsão das amostras preparadas nos demais tecidos (vermelho, preto, branco e estampado) foi realizada da mesma forma utilizando esse modelo hierárquico sem o tecido bege. Observando a Figura 27, é possível identificar padrões na imagem resultantes da textura das fibras do tecido de algodão. Esse efeito da textura não foi identificado nas amostras preparadas nos tecidos sintéticos.

Um caso interessante foi a previsão das amostras preparadas sobre o tecido preto, em que a maioria dos pixels de sangue humano foram corretamente classificados, mas para a amostra de batom a porção referente ao tecido apresentou pixels sendo classificados como sangue humano (ver Figura 28).

Figura 28 – Imagens de *scores* da previsão pelo modelo hierárquico para as amostras preparadas no tecido preto de algodão. Onde BLC representa os tecidos de algodão na cor preta, SH o sangue humano, SA o sangue animal, BA o batom, SY o molho shoyu, GE a geleia, KE o ketchup, PI a pimenta, VG o vinagra balsâmico e VI o vinho tinto.

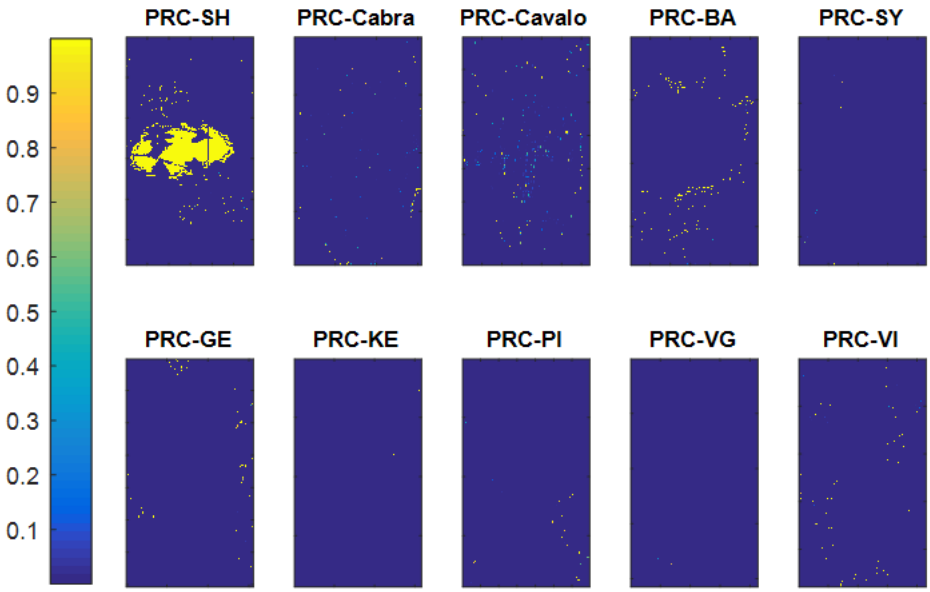


Fonte: Do autor (2019).

Esse efeito observado na Figura 28 para a amostras de batom se reproduz em menor intensidade nas amostras de sangue animal e em alguns outros FPC. Vale ressaltar que a área específica das manchas onde se encontravam as amostras não foram incorretamente classificadas.

Outro caso interessante de se observar foi o resultado para a classificação das amostras preparadas sobre o tecido estampado. Como pode ser visto na Figura 29, não houve problemas de classificação incorreta de falsos-positivos (SA e FPC). Contudo, a imagem de *scores* para a amostras de SH apresentam espaços sem pixels no centro da mancha. Esse efeito resulta da composição desse tecido estampado, no qual desenhos superficiais recobrem a fibra do tecido e modificam a composição da imagem das amostras depositadas nele. Quando essas imagens de previsão são comparadas com a foto do tecido na Figura 5, observa-se que os espaços vazios na imagem de previsão correspondem as pétalas das flores estampadas no tecido.

Figura 29 – Imagens de *scores* da previsão pelo modelo hierárquico para as amostras preparadas no tecido estampado de algodão. Onde PRC representa os tecidos de algodão estampados, SH o sangue humano, SA o sangue animal, BA o batom, SY o molho shoyu, GE a geleia, KE o ketchup, PI a pimenta, VG o vinagra balsâmico e VI o vinho tinto.



Fonte: Do autor (2019).

Para as demais amostras a classificação apresentou resultados similares, sem problemas significativos para a previsão utilizando os modelos hierárquicos de classificação por fusão de técnicas quimiométricas.

3.5 CONCLUSÃO

3.5.1 Identificação de manchas de sangue em substratos individuais e pisos combinados

A partir dos diferentes modelos avaliados na primeira etapa deste trabalho, construídos para distinguir o sangue humano do sangue animal e de possíveis falsos positivos comuns em diferentes substratos, o SIMCA mostrou valores de sensibilidade e especificidade de 1 para o sangue humano somente em substratos de porcelana e vidro. Do mesmo modo, o SPA-LDA mostrou uma diminuição na sensibilidade e também em valores de especificidade para amostras de sangue humano em substratos cerâmicos e metálicos. Por outro lado, PLS-DA e GA-LDA mostraram valores de sensibilidade e especificidade de 1 para todos os substratos, o que demonstra a eficiência de ambos os modelos para classificar corretamente as manchas de sangue humano.

Na segunda etapa do trabalho, os modelos PLS-DA apresentaram um grande potencial de obtenção de sensibilidade igual 1 em todos os modelos, exceto um cujo valor foi 0,988. Este resultado é muito importante para a investigação forense, uma vez que não houve falsos negativos e a especificidade foi próxima de 1 em todos os modelos, com poucos casos de falsos positivos.

Compreendendo que sempre é necessário realizar mais pesquisa, devemos enfatizar o grande potencial do espectrômetro portátil MicroNIR 1700 associado a modelos de classificação eficientes para a identificação confirmatória in situ de manchas de sangue em cenários de crime real.

3.5.2 Identificação de manchas de sangue em tecidos

Os modelos de PCA para as amostras preparadas nos tecidos sintéticos demonstraram um potencial para discriminação das amostras de sangue e de falsos-positivos comuns maior que os modelos construídos para as amostras preparadas nos tecidos de algodão. Isso ficou evidente na estrutura dos modelos hierárquicos de classificação para os tecidos sintéticos que demandaram modelos mais robustos apenas para distinguir as amostras de SH e SA. As amostras preparadas nos tecidos de algodão apresentaram maior complexidade para definição dos limites para os modelos de PCA, além de demandarem um modelo PLS-DA para discriminar as amostras de SH e SA das amostras de FPC restantes das etapas de separação realizadas usando PCA. Isso pode ter acontecido em decorrência da maior influência do tecido nos modelos, provavelmente devido a textura ou tipo de pigmentos das fibras.

Os modelos hierárquicos de classificação construídos para as amostras preparadas no tecido sintético apresentaram sensibilidade de 100% para identificar as manchas de sangue no tecido e especificidade de 98% para distinguir as amostras de SH e SA. Os modelos hierárquicos referentes às amostras preparadas nos tecidos de algodão também demostraram alta sensibilidade para identificar as manchas de sangue nos tecidos (95%) e especificidade de 98%. Desse modo, embora o modelo tenha obtido alta sensibilidade para a identificação das manchas, ainda é possível otimizar esse procedimento de identificação das manchas de sangue para as amostras preparadas nos tecidos de algodão.

Em suma, o método de fusão de técnicas quimiométricas utilizado no trabalho apresentou grande potencial para identificação e classificação de amostras de sangue em condições adversas, como manchas encontradas sobre substratos complexos que são os tecidos coloridos e estampados.

4 IDENTIFICAÇÃO DE *CANNABIS SATIVA* L. ATRAVÉS DE IMAGENS HIPERESPECTRAIS NA REGIÃO DO NIR E SELEÇÃO DE VARIÁVEIS SPARSE-PCA PARA CRIAÇÃO DE MODELOS SIMCA DE CLASSE ÚNICA

4.1 IDENTIFICAÇÃO DE *CANNABIS SATIVA* L.

A *Cannabis sativa* Linnaeus. popularmente conhecida como marijuana, é um narcótico comum no Brasil, sendo consumida como droga recreacional. As plantações ilegais de marijuana estão espalhadas por todo o país e especialmente em regiões mais remotas com acesso a fontes de água (POLICE 2015; UNODC 2009). Uma das atribuições da Polícia Federal (PF) é encontrar e erradicar plantações ilegais de marijuana em todo o território nacional. A metodologia atualmente utilizada pela PF é baseada na inspeção visual de grandes áreas verdes feita por peritos experientes a bordo de helicópteros. Após a identificação de plantações suspeitas, um time de peritos utiliza drones equipados com câmeras digitais para obter imagens mais precisas e essas imagens são avaliadas por um perito especialista em *Cannabis sativa* L. para confirmar a identidade das plantas. A identificação das plantações é relativamente demorada e as operações de erradicação demandam muitos recursos humanos e financeiros (POLICE 2015). Faz-se necessário desenvolver uma metodologia capaz de superar essas limitações e capaz de identificar as plantações de forma mais rápida, barata, simples e acurada.

Atualmente a *Cannabis sativa* L. tem sido bastante estudada acerca de seus diferentes aspectos. Os componentes psicoativos da marijuana, que podem ser encontrados em concentrações significativas fazem parte da classe de compostos canabinoides (ZULFIQAR *et al.*, 2012). Alguns estudos relatam o uso das espectroscopias de Infravermelho por Transformada de Fourier (FT-IR: *Fourier Transform Infrared*) e de Infravermelho Próximo (NIR) para análises semi-quantitativa e qualitativa dos canabinoides presentes na marijuana (BORILLE *et al.*, 2017; CALLADO *et al.*, 2018; HAZEKAMP *et al.*, 2005). A espetroscopia FT-IR é largamente utilizada para identificação e elucidação das estruturas química de compostos orgânicos e por isso têm sido uma ferramenta indispensável para a caracterização dos canabinoides e outros compostos extraídos das plantas de *Cannabis sativa* L. (HAZEKAMP *et al.*, 2005). Com relação a espectroscopia NIR, a interpretação dos espectros é mais complexa e em geral demandam a utilização de ferramentas quimiométricas para extração e utilização da informação química que os espectros carregam (PASQUINI, 2003).

Em comparação com os equipamentos FT-IR, os espectrômetros NIR são mais simples, robustos e versáteis. Os recentes avanços tecnológicos possibilitaram a criação de sistemas de mapeamento químico altamente eficientes e rápidos. Esses sistemas de mapeamento capturam a informação química e criam uma imagem hiperespectral (HSI: *Hyperespectral Images*) que carrega a informação das variáveis espectrais em cada pixel da imagem (VIDAL; AMIGO, 2012). As imagens na região do NIR (HSI-NIR) são bastante interessantes para uso forense pois possibilitam a interpretação visual dos resultados, como por exemplo a identificação de um determinado analito na imagem por meio de técnicas quimiométricas de reconhecimento de padrão (GOWEN *et al.*, 2015).

4.1.1 Objetivo Geral para identificação de *Cannabis sativa* L.

Nesse contexto, o objetivo principal deste trabalho foi desenvolver um método automático, econômico e direto baseado em HSI-NIR e classificação pelo método de modelagem de classes SIMCA (SIMCA: *Soft Independent Class Analogy*) para identificação de *Cannabis sativa* L. em condições naturais e discriminá-la de plantas comumente encontradas próximas das plantações ilegais.

4.1.1.1 Objetivos Específicos para identificação de *Cannabis sativa L.*

- Desenvolver modelos de PCA para identificar os padrões de distribuição dos componentes das amostras.

- Definir variáveis espectrais importantes para caracterização de *Cannabis sativa* L.

- Determinar o número mínimo de variáveis responsáveis pela identificação e discriminação de *Cannabis sativa* L. através do método de seleção de variáveis *sparse*-PCA.

- Construir modelos *sparse* SIMCA de classe única utilizando informações espectrais da *Cannabis sativa* L.

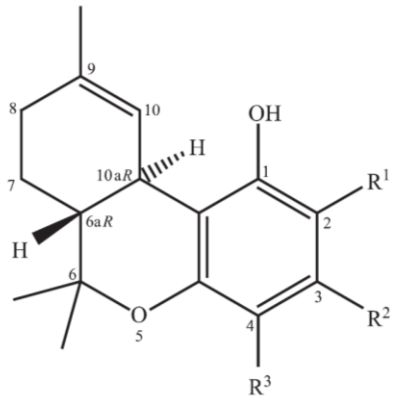
- Aplicar o modelo SIMCA de classe única em imagens/amostras de diferentes plantas, incluindo *Cannabis sativa* L.

4.2 REVISÃO DA LITERATURA SOBRE *CANNABIS SATIVA* LINNAEUS

Cannabis, da família Cannabaceae, tem origem botânica relatada à mais de 5000 anos atrás em algum lugar com localização exata incerta entre o leste da China e a Ásia central (RUSSO, 1998). Existem três espécies de *Cannabis* conhecidas na literatura, a *Cannabis sativa,* a *Cannabis índica* e a *Cannabis ruderalis* (EVANS *et al.*, 1974). A espécie *Cannabis sativa* Linnaeus que originalmente foi utilizada como erva medicinal no preparo de chás para combater dor de cabeça e fadiga, também é a mais comum e conhecida entre as três espécies. Além de seu uso como erva medicinal, a *Cannabis sativa* L. tem sido consumida como droga de recreação desde muito tempo, seja na forma de cigarros, cachimbos ou em receitas (ELSOHLY; SLADE, 2005; RUSSO, 1998). Isso provavelmente decorre da sua fácil obtenção, rápida absorção das toxinas pelo organismo dos usuários e o efeito psicológico dos tetra-hidro-cannabidiol (THC: *tetrahydrocannabinol*) presentes nas folhas com elevadas concentrações (LEITE *et al.*, 2018; WILLIAMSON; EVANS, 2000).

A planta *Cannabis sativa* L. tem uma composição bastante complexa, contendo mais de 535 constituintes identificados pertencentes a quase todas as classes de compostos químicos naturais (ZULFIQAR *et al.*, 2012). Cada classe química presente é constituída por vários componentes, mas apenas os compostos do grupo dos canabinoides, ou grupo C21, apresentam atividade psicoativa significativa. Dentro dessa classe, os Δ9-*trans*-tetrahidrocannabinoides (Δ9-THC) são os compostos com maior potencial psicoativo e encontrados em maior quantidade na folhas das plantas (DAYANANDAN; KAUFMAN, 1976; ELSOHLY; SLADE, 2005; RUSSO, 1998). A Figura 30 apresenta a estrutura química genérica dos Δ9-*trans*-tetrahidrocannabinoides.

Figura 30 – Estrutura química dos compostos Δ9-*trans*-tetrahidrocannabinoides



Fonte: ZULFIQAR *et al.* (2012).

Usualmente, a *Cannabis sativa* L. é consumida em cigarros ou em alimentos, sendo o Δ9-THC rapidamente liberado na corrente sanguínea e distribuído por todo o corpo. O excesso dos compostos derivados do Δ9-THC se aprisionam nos tecidos mais gordurosos e podem permanecer sendo librados aos pouco por até 4 semanas após consumo (JOHANSSON *et al.*, 1989; MCDANIEL *et al.*, 2018). A metabolização da droga pelo organismo é lenta e pode afetar o sistema cognitivo durante todo o processo de eliminação das toxinas. Existem outros efeitos que podem ser causados pelo consumo de *Cannabis sativa* L. em excesso, como por exemplo taquicardia, hipotensão e ansiedade. No entanto, desordens psiquiátricas que podem ser geradas ou fortalecidas pelo consumo da droga são os principais problemas (DIERKER *et al.*, 2018; WILLIAMSON; EVANS, 2000).

Estudos sobre os constituintes da *Cannabis sativa* L. têm demonstrado diferentes propriedades farmacológicas, tais como analgésico, anticonvulcional e estimulantes de apetite para pacientes com AIDS e esclerose múltipla (AZEKAMP; HOI; ERPOORTE, 2004; FRIEDMAN; DEVINSKY, 2015; THOMAS *et al.*, 2010). Assim, os extratos da *Cannabis sativa* L. revelam grande potencial a ser explorado, o que tem motivado inúmeras pesquisas farmacológicas visando diferentes aplicações. Muitos pesquisadores têm desenvolvido metodologias para análise das *Cannabis sativa* L. com o objetivo de identificar, classificar e caracterizar os compostos presentes nas plantas (CALLADO *et al.*, 2018; CITTI *et al.*, 2018; JOHANSSON *et al.*, 1989; LEITE *et al.*, 2018; MCDANIEL *et al.*, 2018). Algumas pesquisadores entendem que os canabinoides são os compostos mais importantes da planta pois apresentam propriedades farmacológicas mais

acentuadas e, portanto, buscam desenvolver metodologias para identificação das plantas e espécies baseado na presença e concentração dos THC's (JOHANSSON *et al.*, 1989).

Os métodos clássicos para extração e quantificação de compostos naturais, tais como técnicas cromatográficas e espectroscópicas, são também os mais comuns para análise de plantas de *Cannabis sativa* L. A maior parte dos métodos quantitativos ou semi-quantitativos para determinação e quantificação de canabinoides em *Cannabis sativa* L. empregam a Cromatografia Líquida de Alta Eficiência (HPLC: *High Performance Liquid Chromatogra*phy) para fazer a extração e a Cromatografia Gasosa hifenada a Espectroscopia de Massas (GC-MS: *Gas Chromatography coupled to Mass spectroscopy*), ou a Ressonância Magnética Nuclear (NMR: *Nuclear Magnetic Resonance Spectroscopy*) ou ainda a Espectroscopia de Massas por Mobilidade Iônica (IMS: *Ion Mobility Mass Spectrometry*) para fazer a quantificação (CONTRERAS *et al.*, 2018; POLITI *et al.*, 2008; RUSSO, 1998; THOMAS *et al.*, 2010). Os métodos qualitativos são muito mais comuns em aplicações forense, que fazem uso contínuo de testes presuntivos para identificação de substâncias ilícitas. A reação de Fast Corinth V e o teste Duquenois-Levine são os mais utilizados para identificação de *Cannabis sativa* L. com objetivos forenses (UNODC, 2009). Contudo, esses métodos clássicos são em geral destrutivos ou inviabilizam análises posteriores, além de serem métodos que se aplicam em amostras já processadas.

Atualmente, poucos estudos utilizam métodos alternativos que preservam a integridade das amostras É o caso de Callado *et al*. (2018) que investigaram a capacidade para quantificação de canabinoides por meio da espectroscopia na região do Infravermelho Próximo e métodos multivariados de análise de dados. Nesse estudo, foi feita a quantificação de 8 canabinoides presentes nas folhas e talos de marijuana em diferentes estágios de maturação e condições de cultivo. A técnica PLS foi utilizada para estabelecer a correlação entre a quantidade de cada componente e os espectros NIR (CALLADO *et al.*, 2018).

Outro grupo de pesquisadores também utilizou a espectroscopia NIR para avaliar a concentração de canabinoides nas plantas, porém o objetivo foi determinar os diferentes estágios de maturação de acordo com a concentração de canabinoides (BORILLE *et al.*, 2017). Nesse estudo foram utilizadas plantas em 3 estágios de maturação (5,5, 7,5 e 10 semanas) para obter espectros NIR das folhas secas e trituradas. Os espectros foram utilizados para construir modelos

multivariados de classificação empregando PLS-DA e Máquinas de Vetores de Suporte (SVM-DA: *Support Vector Machine*s - *Discriminant Analysis*).

As imagens hiperespectrais Vis-NIR também já foram utilizadas em um estudo que visava discriminar plantas de *Cannabis sativa* L. de outras plantas. Nesse trabalho, PCA foi utilizada para discriminar plantas de *Cannabis* e capim verde a partir de conjuntos de espectros médios obtidos de imagens hiperespectrais de cada tipo de planta (AZARIA; GOLDSCHLEGER; BEN-DOR, 2012). Esse estudo se destaca dos demais por seu objetivo de discriminar a *Cannabis* de uma outra espécie de planta. Contudo, vale ressaltar que as imagens hiperespectrais foram utilizadas apenas para obter espectros médios de cada plantação e que a PCA não é uma ferramenta discriminante. O ideal seria a implementação de uma técnica de reconhecimento de padrão supervisionada. Em se tratando da discriminação das plantas de *Cannabis sativa* de outras plantas, a técnica mais adequada deve ser capaz de reconhecer apenas as plantas de *Cannabis*, uma vez que as demais plantas podem formar uma classe heterogênea de amostras. Nesse sentido, a modelagem SIMCA de classe única se apresentam como uma alternativa promissora, pois permite criar um modelo de classe única (*Cannabis sativa* L.).

4.3 METODOLOGIA PARA IDENTIFICAÇÃO DE *CANNABIS SATIVA* L.

4.3.1 Preparo de amostras de *Cannabis sativa* L.

Plantas de *Cannabis sativa* L. foram coletadas diretamente de plantações ilegais durante uma operação de erradicação das plantas conduzida pela Polícia Federal brasileira. Foram coletadas plantas no último estágio de desenvolvimento (plantas maduras) e plantas em desenvolvimento (não completamente maduras). A coleta das amostras foi realizada cuidadosamente para não danificar as raízes e também para manter uma certa quantidade de terra nas raízes. Adicionalmente, seis plantas de diferentes espécies que são comumente encontrados no entorno das plantações de *Cannabis sativa* L. também foram coletadas. Todas as amostras de plantas foram preservadas dentro de baldes com terra úmida ao redor das plantas e em suas raízes.

Para simular uma plantação de *Cannabis sativa* L.*,* no momento da aquisição das imagens hiperespectrais parte do solo coletado foi colocado sobre a bandeja de análise do equipamento

cobrindo toda a área para representar o plano de fundo das plantações. Em seguida, folhas de *Cannabis sativa* L. e das outras plantas coletadas foram removidas das plantas e colocadas, de modo aleatório e com sobreposições, sobre o solo que cobria a bandeja. As folhas de *Cannabis* rodeada por outras plantas simulou a vista superior de uma plantação comum. Essas imagens/amostras foram utilizadas para avaliar o potencial da HSI-NIR para identificação de folhas de *Cannabis sativa* L. e discriminá-las de outras plantas simulando um caso real para aplicação da técnica. Cinco imagens resultantes de combinação de *Cannabis sativa* L. e outras plantas, sendo uma delas escolhida aleatoriamente para otimizar as configurações do pré-processamento.

4.3.2 Instrumentação e aquisição das imagens

A câmera SisuChema da Specim foi utilizada para aquisição das HSI-NIR. A lente usada produzia imagens com dimensões aproximadas de 19 × 15 cm e pixels com 600 × 600 μm. A faixa espectral do equipamento foi de 928 a 2524 nm, com velocidade de amostragem de 60 mm/s, resolução espacial de 3 nm e resolução espectral de 6 nm.

4.3.3 Pré-processamento das imagens de *Cannabis sativa* L.

É importante destacar que estas amostras apresentam uma superfície bastante irregular devido ao formato, curvas e disposição das folhas. Essas características geralmente influenciam o espalhamento da radiação quando técnicas de imagens são adotadas. Com o objetivo de corrigir esses efeitos físicos indesejáveis, que resultam em ruído espectral e espalhamento do sinal, diferentes pré-processamentos foram avaliados. Dentre os pré-processamentos utilizados incluem a variação normal padrão (SNV), correção de sinal multiplicativo (MSC), suavização Savitzky-Golay (polinômio de 2º grau e janela de 5 a 15 pontos) e 1ª e 2ª derivadas de Savitzky-Golay (polinômio de 2º grau e janela de 5 a 15 pontos). Diferentes faixas espectrais de trabalho foram avaliadas (RINNAN; VAN DEN BERG; ENGELSEN, 2009; VIDAL; AMIGO, 2012).

4.3.4 Análise das imagens

PCA foi inicialmente executada em cada imagem para explorar as características das imagens e identificar possíveis clusters relacionados com a *Cannabis sativa* L. Uma imagem foi escolhida para selecionar pixels correspondentes a *Cannabis* fazendo uma seleção de RoI

visualmente através da imagem dos *scores* da PCA. Um conjunto de treinamento consistindo de pixels de *Cannabis* e pixels de outras plantas foi utilizado para construir diferentes modelos de sPCA. Como os número de variáveis originais (comprimentos de onda) é 256, os níveis de compressão/*sparse* testados variou de c_{min} = 1 até c_{max}=16. Para cada nível de *sparse*, quatro modelos de PCA foram construídos, fazendo o número de componentes principais variar de 2 a 5 PCs. A separação obtida para o conjunto de treinamento foi monitorada e utilizada como critério para escolher o melhor grau de restrição de variáveis nos *loadings* e o número de PCs mais adequado.

Uma nova seleção de RoI foi realizada em duas folhas de *Cannabis* da mesma amostra anterior, apenas para coletar pixels correspondentes às principais partes que compõem as folhas, como por exemplo, margens, pontas, caule central, veios e centro das folhas. Esse novo conjunto de dados foi utilizado para construir os modelos SIMCA de classe única, mas agora considerando apenas as variáveis selecionadas pela sPCA.

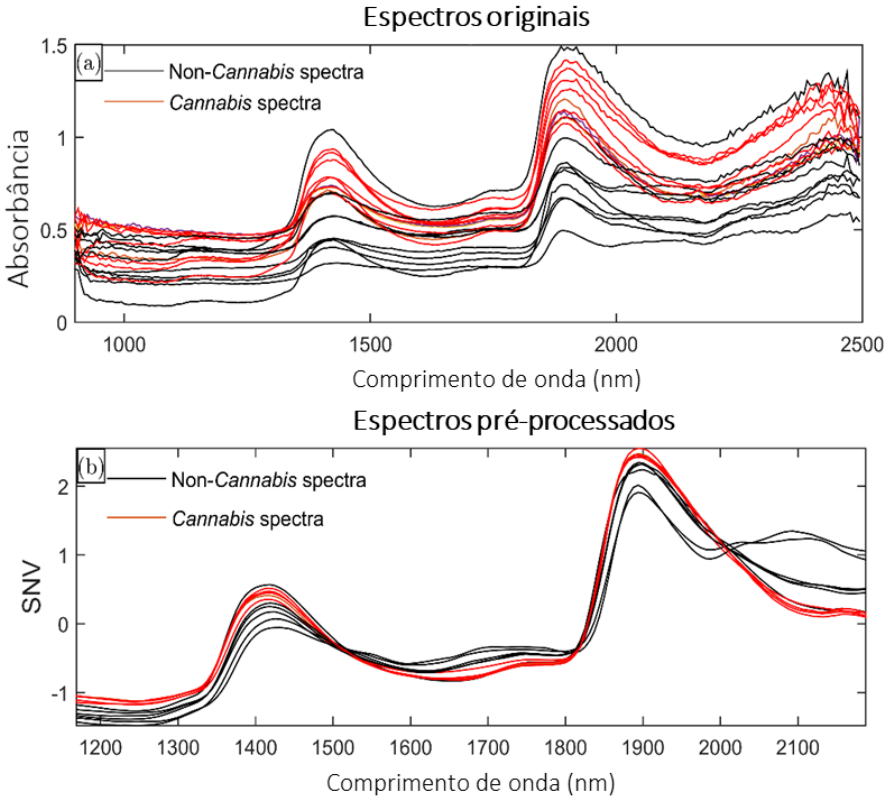
A seleção RoI foi executada mais uma vez para criar um verdadeiro positivo visualmente, em que todas as folhas de *Cannabis sativa* L. foram devidamente identificadas e utilizadas para comparação posteriormente. Todos os modelos e tratamentos quimiométricos foram realizados utilizando rotinas implementadas no MATLAB (MATLAB® R2010a 7.10.0.499, MathWorks), PLS_Toolbox (Eigenvector Research Inc., USA) e HYPER-Tools toolbox disponível para download gratuito em (www.hypertools.org) (MOBARAKI; AMIGO, 2018).

4.4 RESULTADOS E DISCUSSÃO SOBRE IDENTIFICAÇÃO DE *CANNABIS SATIVA* L.

O primeiro passo do pré-processamento foi remover os comprimentos de onda que apresentavam excesso de ruído espectral. A Figura 31.a mostra que a primeira e a última partes dos espectros além de apresentarem muito ruído espectral, também não são muito informativas e por isso foram removidas. Desse modo, a faixa espectral de trabalho foi reduzida para 1262 – 2187nm. O ruído espectral e o espalhamento da radiação observados na faixa espectral resultante foram atenuados em todos os dados utilizando a suavização com filtro Savitzky–Golay (polinômio de segunda ordem e janela de 13 pontos) e SNV (Figura 31.b). As absorções características dos compostos presentes na *Cannabis sativa* L (principalmente canabinoides nas folhas) podem ser

identificadas em toda a faixa espectral, no entanto as outras plantas também apresentam bandas na mesma região correspondentes a compostos diferentes. As diferenças sutis entre as bandas observadas na Figura 31.b indicam que existem pequenas diferenças entre os compostos que compõem a *Cannabis sativa* L. e as outras plantas. A Figura 31.a apresenta bandas intensas em 1450 e 1930 nm correspondentes as absorções do primeiro e do segundo sobretons do grupo −OH mais provavelmente correspondente a moléculas de água presentes nas folhas, uma vez que as plantas foram mantidas frescas até o momento da análise. As bandas mais sutis em 1730 nm e 1760 nm estão relacionadas as vibrações do primeiro sobretom de estiramento de hidrogênios ligados a carbono dos grupos −CH2 e −CH3 presentes em todas as plantas (CALLADO *et al.*, 2018). Entre 2030 e 2180 nm estão presentes bandas bastante sobrepostas relacionadas com proteínas e fibras que podem ser encontradas em todas as espécies de plantas, mas com diferentes concentrações (WORKMAN JR; WEYER, 2012).

Figura 31 – gráfico dos espectros originais (a) e dos espectros pré-processados (b) correspondentes à 20 pixels randômicos de *Cannabis* em vermelho e outras plantas em preto.

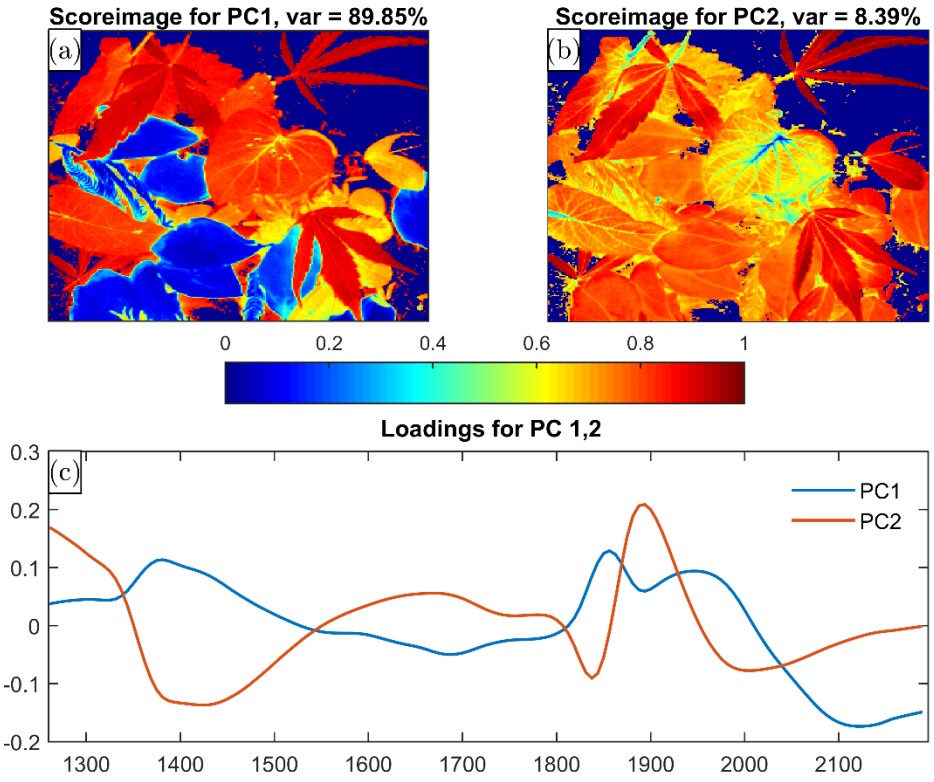


Fonte: Do autor (2019)

Após aplicar o pré-processamento outro modelo de PCA foi construído para uma das imagens. As 10 primeiras PCs foram avaliadas, porém apenas as duas primeiras PCs apresentam as informações mais significativas. Observando a Figura 32 é possível presumir que as folhas de *Cannabis* são similares as outras folhas. As imagens reconstruídas com os escores da PC1 e da PC2 mostraram que os espectros da *Cannabis sativa* L. estão correlacionados com valores positivos.

As bandas em torno de 1450 nm, 1860 nm e 1930 nm nos loadings da PC1 apresentam uma forte correlação com os pixels referentes a *Cannabis*, como pode ser visto nas imagens dos escores, mas essas bandas também têm uma significante correlação com algumas outras folhas. Adicionalmente, os loadings da PC2 mostram que as bandas em 1450 e 1930 nm (provavelmente água) são um pouco mais relevantes para distinguir as folhas de *Cannabis* em relação às demais plantas. O modelo de PCA indica que é possível fazer certa distinção entre *Cannabis sativa* L. e outras plantas similares, porém existem muitas variáveis que estão correlacionadas com ambas as espécies de plantas.

Figura 32 – Imagens reconstruídas a partir dos *scores* da PC1 (a), dos *scores* da PC2 (b) e os loadings das duas primeiras PCs (c).



Fonte: Do autor (2019).

O método de seleção de variáveis *sparse*-PCA foi utilizado visando melhorar a classificação, em decorrência da alta correlação observada para as variáveis responsáveis para distinguir as folhas de *Cannabis* e as demais plantas. Inicialmente, 56 pixels de *Cannabis sativa* L. e 348 pixels do restante da imagem foram selecionados. Uma área correspondente a folhas de *Cannabis sativa* L. foi utilizada para selecionar aleatoriamente 56 pixels e outra área que não continha pixels de *Cannabis* foi utilizada para selecionar 348 pixels dos demais componentes da imagem (incluindo solo). Como mencionado anteriormente, sPCA foi implementada com 16 níveis de *sparse*, considerando os espaços de projeção com 2, 3, 4 e 5 PCs.

A Figura 33 mostra que é possível separar razoavelmente o grupo de pixels da *Cannabis* e das outras plantas utilizando o parâmetro de *sparse* $c = 3$ e 4 *sparse*-PCs (Figura 33.a). Avaliando os loadings para essa sPCA (Figura 33.b), foi possível identificar as variáveis espectrais mais

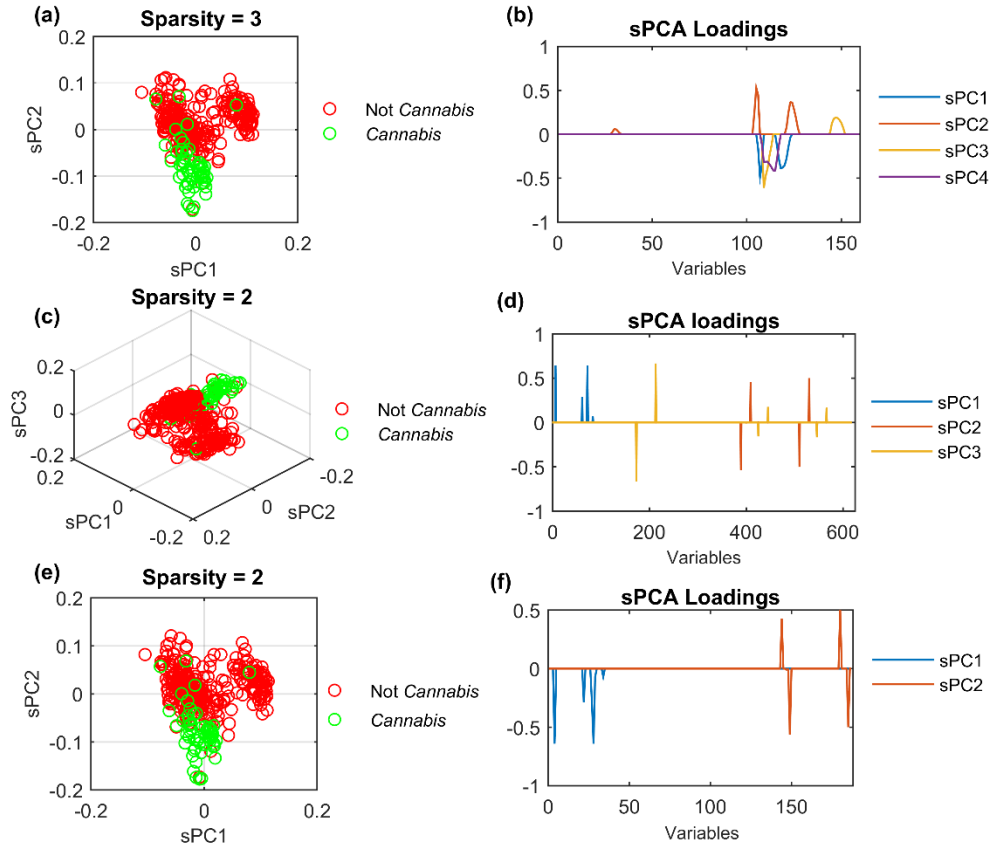
influente para a separação dos conjuntos observada no gráfico de escores. Dessa forma, através dos picos apresentados na Figura 33.b foram selecionados 11 comprimentos de onda (28, 31, 104, 106, 108, 111, 117, 123, 127, 145, 150) correspondentes às bandas mais importantes.

As variáveis selecionadas poderiam ser utilizadas para construir modelos SIMCA para classificar as *Cannabis*, porém algumas delas estão localizadas na mesma banda quando se compara os *loadings* e os espectros brutos das amostras. Partindo do mesmo princípio de redução de variáveis, outra sPCA foi utilizada para avaliar a discriminação dos grupos e selecionar quais das variáveis restantes são essenciais para isso. Para tanto, além das 11 variáveis originais, foram utilizadas variáveis resultantes de operações matemáticas (soma, subtração, produto, divisão e derivada) combinadas com estas. Aplicando este método de combinação de variáveis, foi obtida uma matriz com 405 amostras/pixels e 616 variáveis, e em seguida sPCA foi aplicada com as mesmas configurações iniciais.

A Figura 33.c apresenta o gráfico de escores em que novamente uma sutil separação entre *Cannabis* e o as demais planta pode ser observada. Avaliando-se os loadings (Figure 33.d) para essa segunda sPCA, foi possível identificar 12 variáveis importantes para a separação observada. Entre as variáveis selecionadas, 1 delas é uma variável original e 11 são variáveis resultantes de combinação de primeira ordem de 5 outras variáveis originais. As 6 variáveis originais utilizadas para construção desse modelo sPCA foram 31ª, 106ª, 108ª, 111ª, 117ª e 127ª.

Novamente as variáveis resultantes foram submetidas às combinações matemáticas para gerar uma matriz com mais variáveis e em seguida um último modelo sPCA foi construído. Como resultado, foi observada novamente o mesmo padrão de separação das amostras de *Cannabis* das demais amostras utilizando $c = 2$ e 2sPC (ver Figura 33.e.f). Neste último modelo as variáveis mais importantes observadas no gráfico de loadings continham apenas 4 variáveis originais (1.875 nm, 1.894 nm, 1.931 nm, 1.994 nm). Este foi o modelo mais simples que apresentou uma separação razoável dos pixels de *Cannabis* no gráfico de escores. Assim, as variáveis originais resultantes e suas combinações matemáticas foram utilizadas para compor uma matriz contendo apenas pixels de *Cannabis* para construir um modelo *sparse*-SIMCA de classe única.

Figura 33 – Análise de Componentes Principais *Sparse* construída para um conjunto de espectros de *Cannabis sativa* L. e espectros de outras plantas: (a) *scores* e (b) *loadings* para o 1º modelo *sparse*-PCA; (c) *score* e (d) *loadings* para a 2º modelo *sparse*-PCA; (e) *scores* e (f) *loadings* para o 3º modelo *sparse*-PCA.



Fonte: Do autor (2019).

O modelo de classe única *sparse*-SIMCA foi construído utilizando um conjunto de treinamento composto de 401 pixels/espectros selecionados em diferentes partes das folhas de *Cannabis sativa* L. de acordo com o descrito na metodologia. As diferentes partes das folhas carregam quantidades distintas dos compostos presentes na *Cannabis sativa* L. (ELSOHLY; SLADE, 2005) e, exatamente por isso, foram consideradas para construção do modelo *sparse*-SIMCA trazendo maior variabilidade para os espectros de *Cannabis sativa* L.

Após seleção das amostras, foram selecionados os comprimentos de onda mais importantes determinados por meio do sPCA e o auto-escalonamento das variáveis foi aplicado no conjunto de dados final. A validação cruzada pelo método *venetian blind* utilizando 20 subconjuntos e 15 PCs foi aplicada para validar o modelo *sparse*-SIMCA. Esse método separa o conjunto de validação em

20 subgrupos, e em cada subgrupo um conjunto de amostras (10% neste caso) é utilizado para validação do subgrupo, sendo que as figuras de mérito são as médias dos subgrupos. A validação interna do *sparse*-SIMCA obteve especificidade e sensibilidade iguais a 90,7% e 92,8%, respectivamente.

Além da validação interna, também foi executada a validação externa do modelo utilizando um conjunto de teste. Outra seleção de RoI foi efetuada na imagem e o conjunto de testes foi construído a partir de 256 pixels de *Cannabis sativa* L. selecionados no centro de uma folha e 1044 pixels selecionados randomicamente correspondentes a outras plantas e ao solo. Para validação externa foram obtidas especificidade e sensibilidade iguais a 97,6% e 89,4%, respectivamente. A especificidade para o conjunto de teste foi maior provavelmente pelo fato de que na validação interna havia muitos pixels correspondentes às bordas e pontas das folhas que são regiões de fronteira e difíceis para definir uma RoI adequada. A validação externa continha pixels selecionados apenas na parte mais central de uma das folhas. Esses resultados comprovam a eficiência do modelo *sparse*-SIMCA para a correta identificação de *Cannabis sativa* L. com uma baixa taxa de falso-positivos.

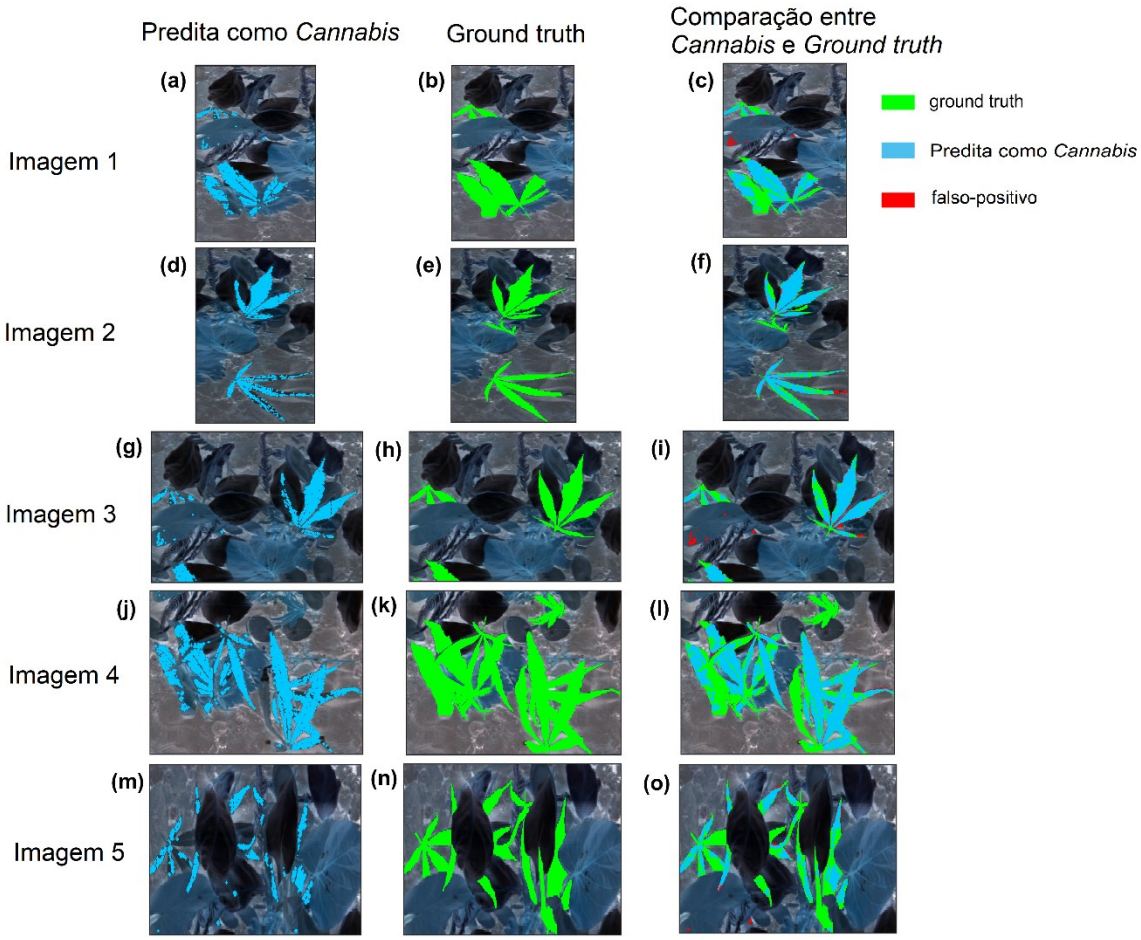
Confirmada a eficácia da metodologia desenvolvida, outro conjunto de validação externa foi construído com todas as imagens que simularam plantações ilegais de *Cannabis sativa* L. As cinco imagens adquiridas foram pré-processadas e projetadas no modelo *sparse*-SIMCA. As imagens remontadas utilizando os *scores* da classificação realizada com o *sparse*-SIMCA estão apresentadas na Figura 34 (a, d, g, j, m). Para evitar falsos-positivos nas imagens de predição, agrupamentos com menos de 5 pixels foram suprimidos.

Uma seleção de RoI foi realizada em todas as imagens fazendo uma identificação visual de todas as folhas de *Cannabis sativa* L., como pode ser visto na Figura 34 (b, e, h, k, n), e essas imagens foram utilizadas como padrão (denominada aqui como *ground truth*) para fazer a comparação com os resultados de predição. A comparação entre as imagens preditas e o *ground truth* pode ser vista na Figura 34 (c, f, i, l, o), onde os pixels vermelhos representam os espectros classificados como *Cannabis sativa* L., mas que não correspondem ao *ground truth*. Vale ressaltar que os pixels presentes no *ground truth* não podem ser considerados como verdadeiro-positivo definitivo, uma vez que a seleção dessa área foi realizada manualmente e pode ter excluído

pequenas áreas correspondentes às extremidades das folhas de *Cannabis sativa* L. ou ter incluído pixels de outras espécies presentes nessas áreas de fronteiras entre folhas.

A predição para todas as imagens demonstrou ainda mais a acurácia do modelo s-SIMCA desenvolvido para identificação de *Cannabis sativa* L., uma vez que a maior parte dos pixels de *Cannabis* foram corretamente classificados e claramente evidenciam as formas reais de folhas de *Cannabis* que é bastante importante do ponto de vista prático para aplicação forense. Adicionalmente, apenas um pequeno número de pixels foi incorretamente predito como sendo *Cannabis,* o que implica em dizer que o modelo é bastante específico. Os resultados de predição para a 5ª imagem (Figura 34.m), representa uma situação em que o modelo s-SIMCA teve baixa sensibilidade, porém as projeções dos resultados de predição mostram claramente o formato de folhas de *Cannabis sativa* L.

Figura 34 – Projeção dos pixels preditos como sendo *Cannabis sativa* L. pelo modelo *sparse*-SIMCA em uma imagem falsa reconstruída a partir da HSI-NIR original (a, d, g, j, m); projeção do *ground truth* na mesma imagem falsa reconstruída (b, e, h, k, n); sobreposição das áreas preditas como *Cannabis sativa* (ciano) sobre o *ground truth* (verde) e os falsos-positivos resultantes (c, f, I, l, o).



Fonte: Do autor (2019).

Observando a Figura 34 (c, f, i, l, o), fica evidente que o modelo s-SIMCA classificou a maior parte dos pixels de Cannabis sativa L. e o baixo número de falsos-positivos confirma a alta especificidade da técnica. Esses resultados também demonstram o potencial das HSI-NIR para identificação de *Cannabis*, mesmo utilizando um pequeno número de comprimentos de onda.

4.5 CONCLUSÃO SOBRE A IDENTIFICAÇÃO DE *CANNABIS SATIVA* L.

Este estudo propõe uma metodologia rápida, objetiva e eficiente baseada em *sparse*-PCA e SIMCA para identificação de *Cannabis sativa* L. utilizando imagens multiespectrais na região do

NIR. Amostras contendo folhas de *Cannabis sativa* L., folhas de plantas similares e solo da região da coleta foram utilizadas para simular uma plantação ilegal em laboratório sob condições ambientais. Pixels de uma das imagens hiperespectrais adquirida de uma amostra foram utilizados para construir um modelo de classe única *sparse*-SIMCA contendo apenas 4 variáveis espectrais no NIR. Este modelo foi aplicado em todas as imagens pré-processadas para identificar *Cannabis sativa* L.

A abordagem quimiométrica utilizada para caracterizar folhas de *Cannabis sativa* L. resultou em uma seleção bastante eficiente de comprimentos de onda, reduzindo de 256 para apenas 4 canais espectrais. O modelo sparse-SIMCA construído com essas 4 variáveis e suas combinações matemáticas demonstrou alta especificidade e sensibilidade para classificar *Cannabis sativa* L. As folhas de *Cannabis sativa* L. de todas as imagens foram corretamente identificadas pelo modelo, sendo possível também reconhecer os padrões das folhas de *Cannabis* em todas as imagens remontadas com os escores do modelo. Além disso, esses resultados podem ser bastante úteis para o desenvolvimento de um sistema de imageamento por drone e um câmera utilizando apenas os quatro comprimentos de onda mais específicos para identificação da *Cannabis sativa* Linnaeu.

5 PERSPECTIVAS FUTURAS

5.1 IDENTIFICAÇÃO DE MANCHAS DE SANGUE

Os próximos passos da pesquisa estão sendo direcionados para a construção de um modelo único que possa ser aplicado para amostras encontradas em todos os tipos de piso estudados. O desafio a ser vencido está relacionado com a minimização da influência dos substratos, para isso é necessário ampliar a base de dados. A metodologia utilizada neste trabalho já está sendo testada para otimização de um modelo, considerando a diferença entre os pisos e outras variáveis, como o tempo de secagem e cor dos pisos. Além disso, outras abordagens que se apresentam como alternativas viáveis estão sendo exploradas, como por exemplo a utilização de modelos hierárquicos de classificação para a identificação das manchas de sangue em pisos.

A identificação de manchas de sangue em tecidos introduz outro desafio relacionado com a variabilidade na composição das amostras, coloração dos tecidos e com a textura das fibras, especialmente para os tecidos de algodão. Com relação aos tecidos sintéticos, a perspectiva é aumentar a variedade dos tecidos estampados. Assim, trabalhos futuros visam ampliar a base de tecidos utilizados para considerar diferentes texturas e mais tipos de pigmentos, para assim tornar os modelos hierárquicos ainda mais robustos e avaliar novas condições para construção dos mesmos.

5.2 IDENTIFICAÇÃO DE *CANNABIS SATIVA* L.

O próximo passo para o desenvolvimento de uma metodologia para a identificação de plantações de *Cannabis sativa* L. é utilizar imagens aéreas obtidas através de câmeras no NIR para testar o modelo desenvolvido. Considerando a dificuldade de obtenção de imagens aéreas dentro da faixa NIR estudada, outra opção é construir um modelo análogo utilizando imagens aéreas em outras faixas espectrais. Outra alternativa interessante é a obtenção de uma câmera construída utilizando os comprimentos de onda selecionados nesta etapa da pesquisa com *Cannabis sativa* L.

REFERÊNCIAS

ARAÚJO, Mário César Ugulino *et al.* The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, v. 57, p. 65–73, 2001.

AZARIA, Ilan; GOLDSCHLEGER, Naftali; BEN-DOR, Eyal. Identification of Cannabis plantations using hyperspectral technology. *Israel Journal of Plant Sciences*, v. 60, n. 1–2, p. 77–83, 2012.

AZEKAMP, Arno H; HOI, Young Hae C; ERPOORTE, Robert V. Quantitative Analysis of Cannabinoids from Cannabis sativa Using H-NMR. *Chemical and Pharmaceutical Bulletin*, v. 52, n. 6, p. 718–721, 2004.

BARNI, Filippo *et al.* Forensic application of the luminol reaction as a presumptive test for latent blood detection. *Talanta*, v. 72, p. 896–913, 2007.

BEEBE, Kenneth R.; PELL, Randy J.; SEASHOLS, Sarah J. *Chemometrics: a simple guide*. 1. ed. New York: Wiley-Blackwell, 1998.

BORILLE, Bruna Tassi *et al.* Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, v. 173, p. 318–323, 2017. Disponível em: <http://dx.doi.org/10.1016/j.saa.2016.09.040>.

BOTELHO, Bruno G *et al.* 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.

BOTONJIC- SEHIC, E. *et al.* Forensic application of near-infrares spectroscopy: agind of bloodstains. *Spectroscopy*, v. 24, n. 2, p. 42–48, 2009.

BREMMER, Rolf H.; NADORT, Annemarie; *et al.* Age estimation of blood stains by hemoglobin derivative determination using reflectance spectroscopy. *Forensic Science International*, v. 206, p. 166–171, 2011. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2010.07.034>.

BREMMER, Rolf H. *et al.* Forensic quest for age determination of bloodstains. *Forensic Science International*, v. 216, n. 1–3, p. 1–11, 2012. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2011.07.027>.

BREMMER, Rolf H.; EDELMAN, Gerda; *et al.* Remote spectroscopic identification of bloodstains. *Journal of Forensic Sciences*, v. 56, n. 6, p. 1471–1475, 2011.

BRERETON, Richard G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant.* [S.l.]: John Wiley & Sons, Ltd, 2003a. v. 8.

BRERETON, Richard G. *Chemometrics: Data Analysis for the Llaboratory and Chemical Plant*. Chichester: John Wiley & Sons, Ltd, 2003b.

BRERETON, Richard G; LLOYD, Gavin R. Partial least squares discriminant analysis : taking the magic away. *Journal of Chemometrics*, v. 28, n. Tutorial, p. 213–225, 2014.

BRO, Rasmus; SMILDE, K. Age. Principal component analysis. *Analytical Methods*, v. 6, p. 2812–

2831, 2014.

BROWN, Steven D.; WALCZAK, Beata; TAULER, Romá. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. 1. ed. Amsterdam: Elsevier, 2009.

CALLADO, C Sánchez-carnerero *et al.* The potential of near infrared spectroscopy to estimate the content of cannabinoids in Cannabis sativa L.: A comparative study. *Talanta*, v. 190, p. 147–157, 2018.

CHALMERS, J.M.; EDWARDS, H.G.M.; HARGREAVES, M.D. *Infrared and Raman Spectroscopy in Forensic Science*. Chichester: John Wiley & Sons, Ltd, 2012.

CHEMELLO, E. Ciência Forense: manchas de sangue. *Química Virtual*, n. 2007, p. 1–11, 2007.

CITTI, Cinzia *et al.* Pharmaceutical and biomedical analysis of cannabinoids: A critical review. *Journal of Pharmaceutical and Biomedical Analysis*, v. 147, p. 565–579, 2018.

CONTRERAS, María del Mar *et al.* Thermal desorption-ion mobility spectrometry: A rapid sensor for thedetection of cannabinoids and discrimination of Cannabis sativa L. chemotypes. *Sensors and Actuators B: Chemical*, v. 273, p. 1413–1424, 2018.

DAYANANDAN, P; KAUFMAN, Peter B. TRICHOMES OF CANNABIS SATIVA L. (CANNABACEAE). *American Journal of Botany*, v. 63, n. 5, p. 578–591, 1976.

DE JUAN, Anna *et al.* Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. *Trends in Analytical Chemistry*, v. 23, n. 1, p. 70–79, 2004.

DIERKER, Lisa *et al.* Addictive Behaviorsmarijuana use disorder symptoms among current marijuana users. *Addictive Behaviors*, v. 76, n. July 2017, p. 161–168, 2018. Disponível em: <http://dx.doi.org/10.1016/j.addbeh.2017.08.013>.

DOTY, Kyle C; LEDNEV, Igor K. Differentiation of human blood from animal blood using Raman spectroscopy: A survey of forensically relavant species. *Forensic Science International*, v. 282, p. 204–210, 2018.

EDELMAN, G. J.; VAN LEEUWEN, T. G.; AALDERS, M. C. G. Hyperspectral imaging of the crime scene for detection and identification of blood stains. *Proceedings of the SPIE*, v. 8743, p. 87430A, 2013. Disponível em: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2021509>.

EDELMAN, G J *et al.* Hyperspectral imaging for non-contact analysis of forensic traces. *Forensic Science International*, v. 223, p. 28–39, 2012.

EDELMAN, Gerda *et al.* Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy. *Forensic Science International*, v. 220, p. 239–244, 2012.

EDELMAN, Gerda; VAN LEEUWEN, Ton G.; AALDERS, Maurice C G. Hyperspectral imaging for the age estimation of blood stains at the crime scene. *Forensic Science International*, v. 223, n. 1–3, p. 72–77, 2012.

ELKINS, Kelly M. Rapid presumptive “fingerprinting” of body fluids and materials by atr ft-ir spectroscopy. *Journal of Forensic Sciences*, v. 56, n. 6, p. 1580–1587, 2011.

ELSOHLY, Mahmoud A.; SLADE, Desmond. Chemical constituents of marijuana The complex

mixture of natural cannabinoids. *Life Sciences*, v. 78, p. 539–548, 2005.

EVANS, Richard *et al.* CANNABIS : AN EXAMPLE OF TAXONOMIC NEGLECT. *Botanical Museum Leaflets*, v. 23, n. 9, p. 337–367, 1974. Disponível em: <https://www.jstor.org/stable/41762285>.

FRIEDMAN, Daniel; DEVINSKY, Orrin. Cannabinoids in the Treatment of Epilepsy From. *The New England Journal of Medicine*, v. 373, n. 11, p. 1048–1058, 2015.

GONZALEZ, Rafael C; WOODS, Richard E. *Digital Image Processing*. 2. ed. New Jersey: Prentice-Hall, Inc, 2002.

GOWEN, Aoife A *et al.* Recent applications of hyperspectral imaging in microbiology. *Talanta*, v. 137, p. 43–54, 2015. Disponível em: <http://dx.doi.org/10.1016/j.talanta.2015.01.012>.

GROBÉRIO, Tatiane S. *et al.* Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR. *Forensic Science International*, v. 257, p. 297–306, 2015.

HAZEKAMP, Arno *et al.* Chromatographic and Spectroscopic Data of Cannabinoids from Cannabis sativa L. *Journal of Liquid Chromatography & Related Technologies*, v. 28, p. 2361–2382, 2005.

JAMES, Stuart H.; KISH, Paul E.; SUTTON, T. Paulette. *Principles of Bloodstain Pattern Analysis: theory and practice*. 3rd revise ed. [S.l.]: CRC Press, Boca Raton, 2005.

JOHANSSON, Eva *et al.* Determination of delta-1-tetrahydrocannabinol in Human Fat Biopsies from Marihuana Users by Gas Chromatography-Mass Spectrometry. *Biomedical Chromatography*, v. 3, n. 1, p. 35–38, 1989.

KJELDAHL, Karin; BRO, Rasmus. Some common misunderstandings in chemometrics. *Journal of Chemometrics*, v. 24, n. Especial Isssue, p. 558–564, 2010.

KUENSTNER, J Todd; NORRIS, Karl H; MCCARTHY, William F. Measurement of Hemoglobin in Unlysed Blood by Near-Infrared Spectroscopy. *Applied spectroscopy*, v. 48, n. 4, p. 484–488, 1994.

LARKIN, Tony; GANNICLIFFE, Chris. Illuminating the health and safety of luminol. *Science and Justice*, v. 48, n. 2, p. 71–75, 2008.

LEARDI, Riccardo; GONZÁLES, Amparo L. Genetic algorithms applied to feature selection in PLS regression how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, v. 41, p. 195–207, 1998.

LEITE, Júlia De A *et al.* Extraction and isolation of cannabinoids from marijuana seizures and characterization by 1H NMR allied to chemometric tools. *Science & Justice*, v. 58, p. 355–365, 2018.

LUTZ, Oliver M. D. *et al.* Reproducible quantification of ethanol in gasoline via a customized mobile near-infrared spectrometer. *Analytica Chimica Acta*, v. 826, p. 61–68, 2014.

MARQUES, Emanuel José Nascimento *et al.* Rapid and non-destructive determination of quality parameters in the “Tommy Atkins” mango using a novel handheld near infrared spectrometer. *Food*

Chemistry, v. 197, p. 1207–1214, 2016.

MCDANIEL, Austin *et al.* Toward the Identification of Marijuana Varieties by Headspace Chemical Forensics. *Forensic Chemistry*, v. 11, p. 22–31, 2018.

MCLAUGHLIN, Gregory; DOTY, Kyle C; LEDNEV, Igor K. Discrimination of human and animal blood traces via Raman spectroscopy. *Forensic Science International*, v. 238, p. 91–95, 2014.

MISTEK, Ewelina; LEDNEV, Igor K. Identification of species' blood by attenuated total reflection (ATR) Fourier transform infrared (FT-IR) spectroscopy. *Analytical and Bioanalytical Chemistry*, v. 407, p. 7435–7442, 2015.

MOBARAKI, Nabiollah; AMIGO, José Manuel. HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis. *Chemometrics and Intelligent Laboratory Systems*, v. 172, p. 174–187, 2018.

MORILLAS, Alvaro Varela; GOOCH, James; FRASCIONE, Nunzianda. Feasibility of a handheld near infrared device for the qualitative analysis of bloodstains. *Talanta*, v. 184, p. 1–6, 2018.

ORPHANOU, Charlotte-Maria. The detection and discrimination of human body fluids using ATR FT-IR spectroscopy. *Forensic Science International*, v. 252, p. e10–e16, 2015. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0379073815001668>.

PASQUINI, Celio. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. *Analytica Chimica Acta*, v. 1026, p. 8–36, 2018.

PASQUINI, Celio. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *Journal of Brazillian Chemistry Society*, v. 14, n. 2, p. 198–219, 2003.

PAVIA, Donald L *et al. Introduction to spectroscopy*. 4. ed. Belmont: Brooks/cole, 2009.

PEREIRA, José F Q *et al.* Evaluation and identi fi cation of blood stains with handheld NIR spectrometer. *Microchemical Journal*, v. 133, p. 561–566, 2017.

POLICE, Brazilian Federal. *GPRE bulletins, Suppression GB-GCotPN, DPF Newsletter*. Disponível em: <http://www.pf.gov.br/agencia/noticias/2015/03/pf-erradica-mais-de-62-mil-pes-de-maconha-em-terras-indigenas-no-ma>.

POLITI, M *et al.* Direct NMR analysis of cannabis water extracts and tinctures and semi-quantitative data on D9-THC and D9-THC-acid. *Phytochemistry*, v. 69, p. 562–570, 2008.

PONTES, Márcio José Coelho *et al.* The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics and Intelligent Laboratory Systems*, v. 78, p. 11–18, 2005.

PRATS-MONTALBÁN, J M; DE JUAN, Ana; FERRER, A. Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems*, v. 107, n. 1, p. 1–23, 2011. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2011.03.002>.

RINNAN, Åsmund; VAN DEN BERG, Frans; ENGELSEN, Søren. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, v. 28, n. 10, p. 1201–1222, 2009.

RUSSO, Ethan. Cannabis for migraine treatment : the once and future prescription? An historical and scientific review. *Pain*, v. 76, n. 1–2, p. 3–8, 1998.

SEIDL, S; HAUSMANN, R; BETZ, P. Comparison of laser and mercury-arc lamp for the detection of body fluids on different substrates. *International Journal of Legal Medicine*, v. 122, p. 241–244, 2008.

SIKIRZHYTSKAYA, Aliaksandra; SIKIRZHYTSKI, Vitali; LEDNEV, Igor K. Raman spectroscopic signature of vaginal fluid and its potential application in forensic body fluid identification. *Forensic Science International*, v. 216, n. 1–3, p. 44–48, 2012. Disponível em: <http://dx.doi.org/10.1016/j.forsciint.2011.08.015>.

SIKIRZHYTSKI, Vitali; SIKIRZHYTSKAYA, Aliaksandra; LEDNEV, Igor K. Advanced statistical analysis of Raman spectroscopic data for the identification of body fluid traces: Semen and blood mixtures. *Forensic Science International*, v. 222, n. 1–3, p. 259–265, 2012.

SIKIRZHYTSKI, Vitali; VIRKLER, Kelly; LEDNEV, Igor K. Discriminant analysis of Raman spectra for body fluid identification for forensic purposes. *Sensors*, v. 10, n. 4, p. 2869–2884, 2010a.

SIKIRZHYTSKI, Vitali; VIRKLER, Kelly; LEDNEV, Igor K. Discriminant Analysis of Raman Spectra for Body Fluid Identification for Forensic Purposes. *Sensors*, v. 10, p. 2869–2884, 2010b.

SILVA, Carolina S *et al.* Near infrared hyperspectral imaging for forensic analysis of document forgery. *Analyst*, v. 139, p. 5176–5184, 2014.

SILVA, Carolina Santos; BRAZ, André; PIMENTEL, Maria Fernanda. Vibrational Spectroscopy and Chemometrics in Forensic Chemistry: Critical Review, Current Trends and Challenges. *Journal of Brazillian Chemistry Society*, v. 30, n. 11, p. 2259–2290, 2019.

SILVA, Carolina Santos; PIMENTEL, Fernanda. *IDENTIFICAR FRAUDES EM DOCUMENTOS*. 2013. 61 f. Universidade Federal de Pernambuco, 2013.

SILVA, Neirivaldo Cavalcante *et al.* Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters. *Analytica Chimica Acta*, v. 954, p. 32–42, 2017.

SILVERSTEIN, Robert M; WEBSTER, Francis X; KIEMLE, David J. *Spectrometric identification of organic compunds*. 7. ed. Danvers: John Wiley & Sons, Inc., 2005.

SKOOG, Douglas A.; HOLLER, F. James; NIEMAN, Timothy A. *Princípios de Análise Instrumental*. 6. ed. Porto Alegre: Bookman, 2009.

STOILOVIC, Milutin. Detection of semen and blood stains using polilight light source as a. *Forensic Science International*, v. 51, p. 289–296, 1991.

TAKAMURA, Ayari *et al.* Soft and Robust Identification of Body Fluid Using Fourier Transform Infrared Spectroscopy and Chemometric Strategies for Forensic Analysis. *Scientific Reports*, n. February, p. 1–10, 2018.

THOMAS, Justin Thomas *et al.* Phytochemistry Metabolic fingerprinting of Cannabis sativa L ., cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. *Phytochemistry*, v. 71, n. 17–18, p. 2058–2073, 2010. Disponível em:

<http://dx.doi.org/10.1016/j.phytochem.2010.10.001>.

TROMBKA, Jacob I *et al.* Crime scene investigations using portable , non-destructive space exploration technology. *Forensic Science International*, v. 129, p. 1–9, 2002.

UNODC. *Recommended methods for the identification and analysis of cannabis and cannabis products*. . Vienna: UNITED NATIONS PUBLICATION. Disponível em: <https://www.unodc.org/documents/scientific/ST-NAR-40-Ebook_1.pdf>. , 2009

VIDAL, Maider; AMIGO, José Manuel. Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems*, v. 117, p. 138–148, 2012. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2012.05.009>.

VIRKLER, Kelly; LEDNEV, Igor K. Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene. *Forensic Science International*, v. 188, n. 1–3, p. 1–17, 2009a.

VIRKLER, Kelly; LEDNEV, Igor K. Forensic body fluid identification: The Raman spectroscopic signature of saliva. *The Analyst*, v. 135, n. 3, p. 512–517, 2010. Disponível em: <http://xlink.rsc.org/?DOI=B919393F>.

VIRKLER, Kelly; LEDNEV, Igor K. Raman spectroscopy offers great potential for the nondestructive confirmatory identification of body fluids. *Forensic Science International*, v. 181, p. e1–e5, 2008.

VIRKLER, Kelly; LEDNEV, Igor K. Blood Species Identification for Forensic Purposes Using Raman Spectroscopy Combined with Advanced Statistical Analysis. *Analytical Chemistry*, v. 81, n. 18, p. 7773–7777, 2009b.

WILLIAMSON, Elizabeth M; EVANS, Fred J. Cannabinoids in Clinical Practice. *Drugs*, v. 60, n. 6, p. 1303–1314, 2000.

WOLD, Svante. Chemometrics ; what do we mean with it , and what do we want from it ? *Chemometrics and Intelligent Laboratory Systems*, v. 30, p. 109–115, 1995.

WOLD, Svante; SJOSTROM, Michael. PLS-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 109–130, 2001.

WORKMAN JR, Jerry; WEYER, Lois. *Pratical guide and spectral atlas for interpretative near-infrared spectroscopy*. 2. ed. New York: CRC Press, 2012.

ZAPATA, Féix; FERNÃNDEZ DE LA OSSA, Ma Ángeles; GARCÍA-RUIZ, Carmen. Emerging spectrometric techniques for the forensic analysis of body fluids. *TrAC - Trends in Analytical Chemistry*, v. 64, p. 53–63, 2015. Disponível em: <http://dx.doi.org/10.1016/j.trac.2014.08.011>.

ZAPATA, Félix; OSSA, M Ángeles Fernández de La; GARCÍA-RUIZ, Carmen. Emerging spectrometric techniques for the forensic analysis of body fluids. *Trends in Analytical Chemistry*, v. 64, p. 53–63, 2015.

ZHANG, Linna *et al.* Blood species identification using Near-Infrared diffuse transmitted spectra and PLS-DA method q. *Infrared Physics & Technology*, v. 76, p. 587–591, 2016.

ZULFIQAR, Fazila *et al.* Cannabisol, a novel Δ9-THC dimer possessing a unique methylene

bridge, isolated from Cannabis sativa. *Tetrahedron Letters*, v. 53, p. 3560–3562, 2012.

APÊNDICE A – ARTIGO PUBLICADO

Microchemical Journal 133 (2017) 561–566

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Evaluation and identification of blood stains with handheld NIR spectrometer

José F.Q. Pereira [a], Carolina S. Silva [a], Maria Júlia L. Vieira [b], Maria Fernanda Pimentel [b,*], André Braz [a], Ricardo S. Honorato [c]

[a] Universidade Federal de Pernambuco, Departamento de Química Fundamental, Recife, PE, Brazil
[b] Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife, PE, Brazil
[c] Polícia Federal, Recife, PE, Brazil

ARTICLE INFO

Article history:
Received 15 November 2016
Received in revised form 12 April 2017
Accepted 24 April 2017
Available online 25 April 2017

Keywords:
Blood stains
Near infrared
Supervised classification
Forensic
Handheld spectrometer

ABSTRACT

This work describes a noninvasive, nondestructive methodology for the confirmatory and in situ identification of dry blood stains deposited in different substrates using a hand-held near infrared (NIR) spectrometer. Different supervised pattern recognition methods were evaluated and compared for the correct identification of human blood. Human and animal blood stains and stains from different commercial products (perceived as common false positives) that had been deposited on different substrates were analyzed directly by NIR spectroscopy. The best pre-processing used was Standard Normal Variate (SNV) and normalization by range, which was obtained by evaluating the best classification for human blood with Soft Independent Modeling of Class Analogy (SIMCA). SIMCA showed 100% correct classification for porcelain and glass substrates but 80% for metal and 90% for ceramic substrates. Genetic Algorithm - Linear Discriminant Analysis (GA-LDA) showed better classification performance (100%) than Successive Projection Algorithm - Linear Discriminant Analysis (SPA-LDA), in which one false positive and one false negative were obtained. Partial-Least Squares Discriminant Analysis (PLS-DA) correctly classified the human blood and other stains in all substrates.

1. Introduction

Blood is perhaps the most important type of forensic trace that can be found at crime scenes [1] and it can confirm the presence of a person, providing this way information related to the activity or action. It can also be used for DNA analysis, which may eventually lead to the identification of a possible suspect [2]. However, DNA analysis takes time and resources, and successful results usually depend on a careful selection of blood traces. This is not always easy to accomplish because the presence of animal blood or other substances of similar aspect (like pepper sauce for example) can cause the erroneous collection of evidences (also known as false positives). Thus, in order to have more cost-effective and efficient forensic examinations, it is important to identify correctly the nature of possible blood traces found at crime scenes.

According to a recent review [3], there are several tests that can be used to detect or confirm the presence of blood traces directly at the crime scene. These include visual examinations using microscopy; chemical examinations like the commonly known Luminol or Kastle-Meyer tests and the Takayama or Teichmann microcrystal tests; spectroscopic examinations like alternating light sources or UV–vis absorption; and immunological examinations like the antibodies test kits RSIDTM or ABA Card®Hematrace®. Generally speaking, these examinations vary greatly in specificity for human blood. Additionally, they suffer several disadvantages, for example: their performance may be affected by several interferences; some require dark conditions or swab collection of traces; others use reagents that invalidate subsequent DNA analysis or can directly destroy the evidence; some can suffer from cross-reactivity with other biological fluids; and many commonly found-substances can give false positives or negatives [4].

Therefore, there is a great need for, not only presumptive but also confirmatory methods that are simultaneously nondestructive and specific for human blood. Additionally, current trends suggest the paramount need for field-portable, easy-to-use and reliable instruments that provide rapid identification of human blood directly at the crime scene [5].

In this sense, vibrational spectroscopic techniques such as Raman and Infrared (IR) have been extensively applied in Forensic Science [6], particularly for the identification of human blood traces [3] due to the fact that human blood has characteristic chemical signatures. Additionally, these techniques are noninvasive, nondestructive and permit rapid analysis. Raman spectroscopy has been reported to a greater extent in the literature when compared to IR (for example, see [7,8]), chiefly because the presence of water in blood has very little influence on the blood's Raman spectra, unlike IR. Nevertheless, IR has the

* Corresponding author.
E-mail address: mfp@ufpe.br (M.F. Pimentel).

http://dx.doi.org/10.1016/j.microc.2017.04.038
0026-265X/

practical advantage over Raman that fluorescence or external ambient light does not influence the quality of spectral acquisitions, which is particularly relevant for examinations at the crime scene.

Developments in technology and instrumentation in recent years have led to considerable miniaturization of equipment used in many analytical techniques, especially IR and Raman spectroscopy [9]. Portable and hand-held devices are now important assets for distinct sectors of research and investigation [10], including Forensic Science, where analysis has become faster, more effective and adaptable to different crime scenes, thus contributing to a reduction in the amount of time for decision-making and the need for fewer laboratory resources.

Up to now, works published on the use of IR for discrimination of blood stains consist mainly of studies in the middle infrared region (MIR) and comparison of spectral fingerprinting from different body fluids, using benchtop equipment [11,12]. Also, thermal infrared images have been evaluated for the identification of blood stains in black substrates [13]. Edelman et al. [14] used Near IR (NIR) spectroscopy combined with visible reflectance and Partial Least Squares (PLS) regression to analyze blood stains in cotton fabric as well as pure blood as a function of time, for the purpose of determining age. They used two benchtop instruments working in the ranges of 400–2500 nm and 800–2778 nm. The authors reported that bloodstains could be distinguished from other red-like substances and colored cotton backgrounds with 100% sensitivity and specificity. NIR spectroscopy proved to be suitable for short-term age estimation but limitations, including problems identifying spectral contributions from the different components present in blood and materials that strongly absorb in the NIR range, could complicate analyses.

Multivariate data analysis can overcome these limitations and extract the maximum useful chemical information from NIR spectra [10,11,15,16]. In this study, the potential of different strategies for multivariate classification are described: two hard modeling approaches, PLS Discriminant Analysis (PLS-DA) and Linear Discriminant Analysis (LDA), and one soft modeling approach, Soft Independent Modeling of Class Analogy (SIMCA). The difference between the two approaches is that with hard modeling, it is mandatory that a sample be classified in one of the modeled classes. The limitation of this constraint is that a sample may in fact not belong to any of the classes. In soft modeling approaches, samples can be assigned or excluded from a certain class or even assigned or excluded to all classes. This way makes it possible to identify outliers, unexpected samples that belong to classes that have not been included in the training set, or even erroneous data [17].

PLS-DA has become a widely used chemometric tool in many analytical fields, and also in forensic applications using IR spectroscopy (for example, see [15,16,18,19]) owing to its availability in common statistical software. LDA is another classification method that has shown usefulness in forensic contexts (for example, see [20]). Frequently, this method requires a variable reduction step, such as simple PCA or even variable selection algorithms, for example, the Successive Projections Algorithm (SPA) and the Genetic Algorithm (GA), which were used in this study [21]. Class-modeling SIMCA has also been used in forensic applications of NIR spectroscopy [22,23].

This work describes a novel application of a handheld NIR instrument for the identification of human blood directly from stains deposited on different surfaces. The performance of different supervised pattern recognition methods, such as PLS-DA, SPA-LDA, GA-LDA and SIMCA, was tested and compared as to their correct identification of human blood species against animal blood and other common substances that usually give false-positive results.

2. Material and methods

2.1. Samples and sample preparation

Thirty-one volunteers (14 men and 17 women) donated blood samples which, were collected directly from their fingers using individual and sterilized needles. The animal blood samples, from a cat and a dog, were collected by a veterinary hospital. Different red-colored commercially available products that could be commonly perceived as blood were used as the common false positives samples, such as red lipstick, pepper sauce, soy sauce, red wine and balsamic vinegar. Stains were applied directly or with a Pasteur pipette (2 drops) on 4 different substrates, such as beige floor porcelain tile, white ceramic tile, glass and the metallic part of a knife. The volume of human blood stains was not controlled. Table 1 shows a summary of the samples and number of stains deposited on each substrate. The number of stains applied differed for each substrate due to the varied volume of blood extracted from each donor. All 220 stains were allowed to dry for three days under ambient conditions before analysis.

2.2. Instrumentation and data acquisition

All spectra were acquired directly from the stains using a MicroNir 1700 spectrophotometer (from Viavi). This instrument has a linear-variable filter (LVF), which is directly attached to a linear Indium Gallium Arsenide (InGaAs) array detector. The system has two small tungsten light bulbs and a USB interface coupled for power and data transfer. The spectrometer's dimensions are 45 mm in diameter and 42 mm in height, weighing about 60 g. The instrument spectral range is 908 to 1676 nm, with a spectral resolution of 12.5 nm. Each spectrum was an average of 64 scans, with an integration time of 5 ms. In order to increase spectral variability and improve representativeness for the chemometric models, three spectra were acquired in different positions for each stain by performing small displacements over the stain. For stains that dispersed in the substrate, up to five spectra were taken. Each spectrum acquired was considered as an individual sample, in order to account for differences in the amount of blood per stain as a result of the spreading and drying process. Due to the high transparency of glass, a white Teflon board was placed underneath the glass in order to avoid absorption from the countertop.

2.3. Data pre-processing

Different pre-processing techniques were tested to correct undesirable effects related to noise, scattering of radiation and to normalize differences related to the amount of sample deposited: Standard Normal Variate (SNV), Multiplicative Signal Correction (MSC), Savitzky-Golay smoothing filter, Savitzky-Golay 1st and 2nd derivatives (5 to 13 point window) and normalization by maximum, range and mean [24,25]. The best combination of pre-processing techniques was chosen based on visual inspection of the spectra and on the best results obtained with SIMCA modeling for samples on the porcelain tiles.

2.4. Data analysis

All chemometric treatment was performed with Matlab software (MATLAB® R2010a 7.10.0.499, MathWorks) and the PLS-DA models were performed with PLS Toolbox (Eigenvector Research, Inc). The SPA and GA algorithms used in the LDA models were those described by [26]. The different treatments were performed on the data sets for each substrate. Three different classes were considered: Human Blood

Table 1
Samples and number of stains added to each substrate.

Substrates	Number of stains		
	Human blood	Animal blood	False-positive
Porcelain	27 (7 donors)	10	15
Ceramic	36 (9 donors)	10	15
Glass	33 (9 donors)	10	15
Metal	25 (6 donors)	9	15
Total	121	39	60

(HB), Animal Blood (AB) and Common False-Positive (CFP). For SIMCA, the animal blood and the false positive samples on each substrate were divided into a training set (comprising 70% of samples) and an external validation set (comprising 30% of samples). Regarding the human blood, the number of samples for the training set varied. PCA was employed for splitting the human blood samples into the training and validation set because we wanted to put all the samples from two volunteers in the external validation set. The training set should span the maximum variability and the samples from the two volunteers of the validation set could not be an extrapolation. The score plots were also examined to identify possible grouping patterns and outliers (based on Hotelling's T^2 and Q residual values). Due to experimental limitations (the volume of blood extracted from each donor varied; the amount of blood was limited and it was not possible to apply equal number of stains per substrate; the metal substrate available could not accommodate as many stains as the other substrates) the ratio of samples selected for the training set differed for each substrate and model. The percentage of samples selected for the training set for each substrate was: 63% in porcelain tile, 75% in ceramic tile, 77% in glass and 69% in metal. Leave-one-out-cross-validation method was employed to select the number of components. The same samples that were used in the training and external validation sets for SIMCA were also used to build the PLS-DA, SPA-LDA and GA-LDA models. Furthermore, for SPA-LDA and GA-LDA, 66% of the samples of the training set were then used for calibration and the rest were used for internal validation (test set), as required by the algorithms employed [21]. The SPA and GA variable selection methods were employed to reduce the dimensionality of data. The GA was carried out over 100 generations with 200 chromosomes each generation. The rates of mutation and crossover were set to 10% and 60% respectively. The variable selection process was repeated 10 times starting from random populations, and the best result of cost function defined the variables to be used.

3. Results and discussion

Initially, spectra from human blood, animal blood and false positive samples on the porcelain tile substrate were used to evaluate the best combination of preprocessing methods. For this, the raw and the preprocessed spectra were constantly compared in order to observe the spectral effects of each preprocessing. Also, PCA was performed after each preprocessing, in order to observe the distribution of the data. The score plots were used to visualize the pattern and the few outliers identified were removed. Then, SIMCA models were used to evaluate the preprocessing that could most effectively discriminate among the samples in the three pre-established classes: Human Blood (HB), Animal Blood (AB) and Common False-Positive (CFP). Following the methodology described previously, it was found that SNV combined with normalization by range was the most efficient in minimizing scattering effects and classifying the samples correctly (95% of confidence) 100% of the HB and CFP samples and only 3 AB samples were not classified in any of the classes. Fig. 1 shows the PC1 versus PC2 score plot and loadings from HB, AB and CFP samples deposited on the porcelain substrate after SNV correction and normalization by range. As depicted, each class formed distinct clusters from the other classes but the CFP samples were much more dispersed. The lipstick samples had negative scores in PC1 while the other CFP samples had high positive scores. The lipstick samples had been influenced by the region from 900 to 1300 nm, as shown in the loadings of PC1. The other CFP had been influenced by the band around 1460 nm (high positive loadings). This band may be attributed to the first overtone of OH from alcohol and acid. The samples dispersion in the PC2 scores is due mainly to a water band around 970 nm and the band at 1210 nm from the second overtone of C—H stretching [27].

The same method was applied to the spectra of the stains on the other substrates. Fig. 2 shows the raw and preprocessed spectra of HB, AB and CFP blood stains on different substrates.

As depicted, the absorptions in the spectra of HB, AB and CFP vary significantly depending on the substrate. This indicates that there is a strong inference from the substrate, especially in the spectral region 900–1300 nm. The bands observed at 900–1150 nm and 1450 nm relate to the H—O—H third overtone and stretching vibrations of water, respectively [27]. Although identification and interpretation of specific blood compounds are difficult with NIR spectra, the —CH third overtone stretching vibrations of oxyhemoglobin around 930 nm and the N—H stretching vibrations of the amide groups from albumin protein around 1570 nm can be observed [28]. The weak absorption band at 1550 nm, can either correlate to N—H first overtone stretching vibrations of albumin or the stretching and O—H first overtone vibrations of water [27,28]. Nevertheless, the region 1300–1676 nm of the spectra clearly exhibit the spectral information most related to blood because this is where HB and AB spectra are similar to each other and different from the CFP.

The SIMCA models built for the other substrates were able to classify correctly 80%, 90% and 100% of the HB samples on metal, ceramic tile and glass substrates, respectively (95% of confidence). Furthermore, the samples that were not classified correctly were not classified in any other class, which indicates specificity equal to 1. SIMCA, as a soft modeling method, allows flexible ways of classifying samples that may not belong to any class included in the training set. From a forensic investigative perspective, this is very important because blood evidence that has not been clearly identified as blood will not be rejected as evidence and needs to be analyzed by other confirmatory techniques. It is also important to remark that only 62.5% to 70% correct classification results were obtained from AB samples on almost every substrate with

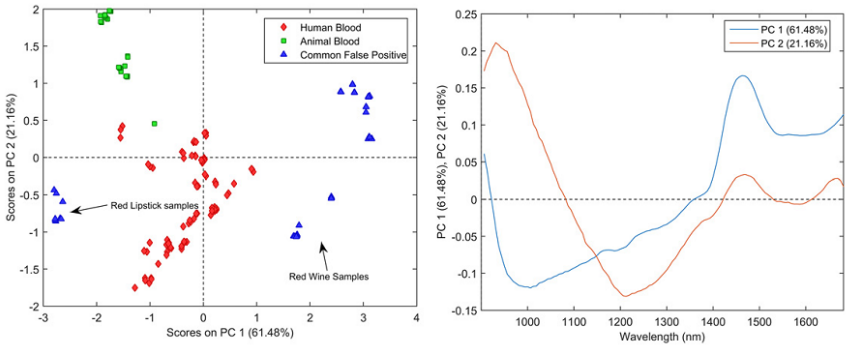


Fig. 1. PCA score plot and loadings of human blood (red), animal blood (green) and common false positive (blue) samples deposited on porcelain substrate after preprocessing with SNV and normalization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

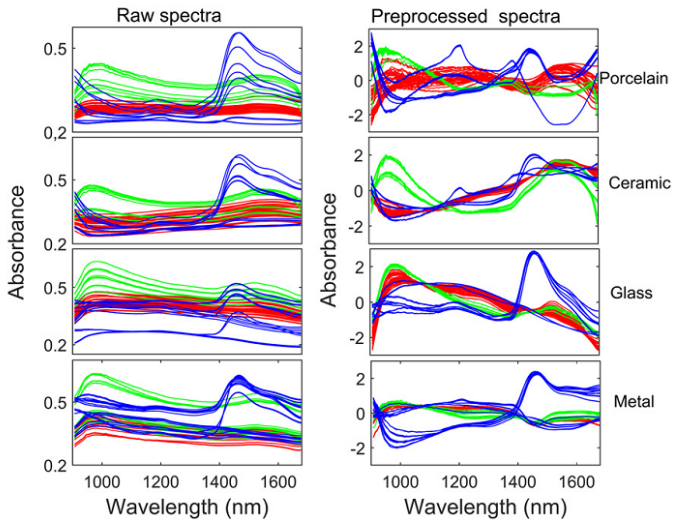


Fig. 2. Raw spectra (left column) and pre-processed spectra (right column), with SNV and normalization by range, of all training samples on different substrates. Human blood (red), animal blood (green) and common false positives (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

SIMCA, with the exception of the model for the glass substrate, where a 100% correct classification was obtained. This result may be due to the fact that the number of AB samples to build the model and make predictions was small. Increasing the number of AB samples might solve this issue. However, the purpose of the classification in this study was to prevent HB from being confused with AB samples rather than classifying AB samples correctly.

Table 2 shows the classification results in more detail and features of the models. As depicted, models were built using up to 3 PCs for each class in every substrate and the explained variance was always above 90%. These aspects characterize the simplicity and efficiency of the SIMCA models.

The same datasets used in SIMCA analyses were applied for LDA. The only difference was that the training set was divided into training and test sets (66% and 34%, respectively), as required by the algorithms employed [26,29].

Table 2
SIMCA Classification results and structure features from all models build using 5% of significance.

Substrates	Classes	Datasets			Classification		
		Training		External validation	Models		
		No. of spectra	No. de PC's	No. of spectra	HB	AB	CFP
Porcelain	HB	57 (5 donors)	3	24 (2 donors)	100.0%	–	–
	AB	20	3	10	–	70.0%	–
	CFP	30	2	15	–	–	100.0%
Ceramic	HB	81 (7 donors)	3	27 (2 donors)	88.9%	–	–
	AB	20	2	8	–	62.5%	–
	CFP	30	3	15	–	–	93.0%
Glass	HB	81 (7 donors)	2	18 (2 donors)	100.0%	–	–
	AB	20	3	7	–	100.0%	–
	CFP	30	3	15	–	–	80.0%
Metal	HB	45 (4 donors)	3	15 (2 donors)	80.0%	–	–
	AB	20	3	10	–	70.0%	–
	CFP	30	2	15	–	–	100.0%

The SPA algorithm selected different number of variables from the samples on different substrates. The 8 variables that were selected for the porcelain tile models were in the region of 900–1150 nm, corresponding to H—O—H third overtone vibration of water and —CH and —CH2 third stretching overtone vibration of oxyhemoglobin (930 nm); and at 1577 nm corresponding to N—H stretching vibration of the amide group from albumin protein [27,28]. The 16 variables that were selected for the ceramic tile models were in the region of 900–1150 nm, at 1430 nm corresponding to H—O—H stretching vibration of water and at 1530 nm either corresponding to N—H stretching vibration of the amide group from albumin protein or to stretching and O—H first overtone vibration of water [28]. From the 4 variables that were selected for glass models, 2 variables corresponded to informative bands at 1460 nm stretching vibration of water and 1570 nm that either corresponded to N—H stretching vibration of the amide group from

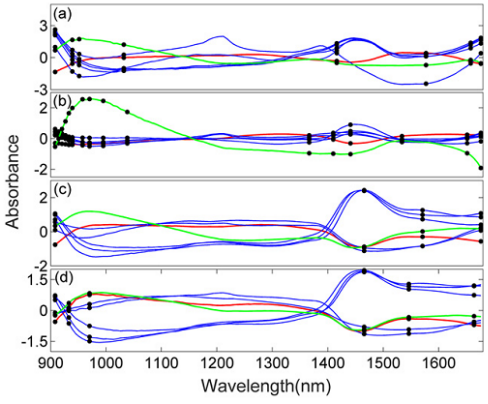


Fig. 3. Mean-centered spectra for all samples and the variables selected by SPA algorithm for samples on (a) porcelain, (b) ceramic, (c) glass and (d) metallic substrates. Blue spectra for CFP; red spectra for HB; green spectra for AB and black dots for selected variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

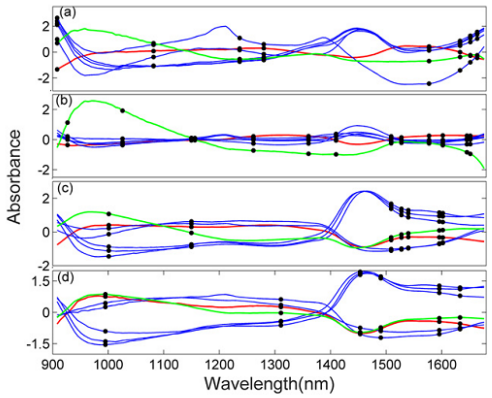


Fig. 4. Mean-centered spectra for all samples and the variables selected by GA algorithm for samples on (a) porcelain, (b) ceramic, (c) glass and (d) metallic substrates. Blue spectra for CFP; red spectra for HB; green spectra for AB and black dots for selected variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

albumin protein or to stretching and O—H first overtone vibration of water [28]. The 6 variables that were selected for metal models were in the region 900–1150 nm, at 1550 nm, either corresponding to a N—H stretching vibration of the amide group from albumin protein or to stretching and O—H first overtone vibration of water [28]; and at 1466 nm corresponding to O—H vibrations of alcohol. Fig. 3 show the mean spectra for all samples and the vibrations assigned to the variables selected with the SPA algorithm. The spectra were mean centered to better visualize the differences among them.

The variable selection by the GA algorithm was chosen based on the lowest value of cost function. Each time the GA algorithm is used for the same data set, different variables may be selected, for that reason the models were run 10 times. The most representative variables with the lowest value of cost function were selected for each GA-LDA model. The GA algorithm selected 8, 13, 7 and 6 variables to build each model, for porcelain tile, ceramic tile, glass and metal, respectively. For porcelain, the 8 variables selected were at 1080 nm and 1570 nm, at 1280 nm, corresponding to the O—H vibrations, possibly from alcohol, and at 1650 nm corresponding to the C—H first overtone vibration [28]. The 13 variables selected for the samples on ceramic tile included all the variables that had already been identified for this substrate. The 7 variables selected for the glass substrate were at 1000 nm, at 1150 nm corresponding to the C—H second overtone vibration and 6 other variables in the region 1500–1676 nm, possibly corresponding to the N—H stretching vibrations [28]. The 6 variables selected for the metal substrate were at 1000 nm corresponding to the C—H second overtone vibration; 1450 nm corresponding to the stretching vibration of water and 1600 nm either corresponding to the N—H stretching vibration of the amide group from albumin protein or to stretching and O—H first overtone vibration of water [28]. Fig. 4 shows the mean spectra for all samples and the vibrations assigned from the variables selected by the GA algorithm.

PLS-DA and LDA models, with the variables selected using SPA and GA algorithms, were applied to the external validation set of all substrates. The performances of PLS-DA, SPA-LDA and GA-LDA are summarized in Table 3. As can be seen, for SPA-LDA models, one CFP sample (red wine) on ceramic was incorrectly classified as belonging to the HB class. This classification error (false positive) could be a consequence of the use of many variables selected in the region 900–1150 nm, where the spectral information is very similar for all samples. Additionally, one HB sample used on the metal substrate was incorrectly classified as belonging to the AB class. This occurred because most of the variables selected did not show spectral differences between human and animal blood, with exception of one variable around 1660 nm. This classification error (false negative) represents a serious problem for forensic application because it could cause an important piece of evidence to be discarded.

The GA-LDA models achieved the best correct classification results, 100% of correct classification for all models built. Unlike SPA-LDA, the GA algorithm selected variables from the entire spectral region and therefore included most of the informative bands, as shown in Fig. 4.

PLS-DA models were performed using the leave-one-out cross-validation. Different threshold and number of Latent Variables (LV) were automatically chosen for each model by the PLS-Toolbox, which was calculated using a Bayesian approach [30,31]. The numbers of LVs selected for each model were 3, 2, 5 and 5 for porcelain, ceramic, glass and metallic substrates, respectively. The high threshold of 0.8 obtained for prediction of human blood samples on the ceramic tiles may have resulted from the wide variations within the AB and CFP samples in this case.

For the external validation set, the PLS-DA models correctly classified 100% of the samples on all substrates, similar to the results obtained with GA-LDA. As shown in Table 3, the sensitivity (Sn) and specificity (Sp) for PLS-DA and GA-LDA were equal to 1. In this respect, the SPA-LDA models of samples on ceramic tiles and metal substrates showed a small decrease of sensitivity and specificity values, due to the classification errors already mentioned previously. In the specific case of the hard modeling approach, the decrease of specificity is related to the decrease of sensitivity.

4. Conclusion

This work has presented a methodology for the identification of blood stains directly on different surfaces using a handheld NIR

Table 3
Results of classification for external validation sets using SPA-LDA, GA-LDA and PLS-DA.

			SPA-LDA			GA-LDA			PLS-DA		
Classes per model		No. samples	Classification (%)	Sn	Sp	Classfication (%)	Sn	Sp	Classfication (%)	Sn	Sp
Porcelain	HB	24	100	1	1	100	1	1	100	1	1
	AB	10	100	1	1	100	1	1	100	1	1
	CFP	15	100	1	1	100	1	1	100	1	1
Ceramic	HB	27	100	1	0.958	100	1	1	100	1	1
	AB	8	100	1	1	100	1	1	100	1	1
	CFP	15	93.4	0.937	1	100	1	1	100	1	1
Glass	HB	24	100	1	1	100	1	1	100	1	1
	AB	9	100	1	1	100	1	1	100	1	1
	CFP	15	100	1	1	100	1	1	100	1	1
Metal	HB	15	93.4	0.937	1	100	1	1	100	1	1
	AB	11	100	1	0.967	100	1	1	100	1	1
	CFP	15	100	1	1	100	1	1	100	1	1

Sn: sensitivity; Sp: specificity.

spectrometer. The possible uses of this instrument at the scene of a crime are numerous: it is handheld and easily carried to any crime scene, it is simple to use, analyses are rapid, noncontact, nondestructive and made directly over the stain, which avoids sample extraction or treatment and does not destroy the evidence. However, a limitation that is inherent in the technique is the need for the sample to be dry since water has strong IR absorption. From the different models evaluated to identify the presence of human blood against the presence of animal blood and possible false positives in different substrates, SIMCA showed sensitivity and specificity values of 1 for human blood on porcelain and glass substrates. The sensitivity value decreased, however, with the samples on ceramic and metal substrates. Similarly SPA-LDA showed a decrease in the sensitivity and also in specificity values for human blood samples on ceramic and metal substrates. On the other hand, PLS-DA and GA-LDA showed sensitivity and specificity values of 1 for all substrates, which demonstrates the efficiency of both models for correctly classifying human blood stains. Understanding that more research is required, we should emphasize the great potential of the MicroNIR portable spectrometer associated with efficient classification models for in situ confirmatory identification of blood stains in real crime scenarios.

Acknowledgements

INCTAA (Processes no.: CNPq 573894/2008-6; FAPESP 2008/57808-1), NUQAAPE –FACEPE (APQ-0346-1.06/14), CNPq, FACEPE, CAPES, Núcleo de Estudos em Química Forense – NEQUIFOR (CAPES AUXPE 3509/2014, Edital PROFORENSE 2014). The English text of this paper has been revised by Sidney Pratt, Canadian, MAT (The Johns Hopkins University), RSAdip - TESL (Cambridge University).

References

[1] S.H. James, P.E. Kish, T.P. Sutton, Principles of Bloodstain Pattern Analysis: Theory and Practice (Practical Aspects of Criminal and Forensic Investigations Series), 3rd revised ed. CRC Press, Boca Raton, 2005.
[2] R.H. Bremmer, A. Nadort, T.G. Leeuwen, M.J.C. Gemert, M.C.G. Aalders, Age estimation of blood stains by hemoglobin derivative determination using reflectance spectroscopy, Forensic Sci. Int. 206 (2011) 166–171.
[3] F. Zapata, M.Á. Fernández de la Ossa, C. García-Ruiz, Emerging spectrometric techniques for the forensic analysis of body fluids, Trends Anal. Chem. 64 (2015) 53–63.
[4] K. Virkler, I.K. Lednev, Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene, Forensic Sci. Int. 188 (2009) 1–17.
[5] R.H. Bremmer, G. Edelman, T.D. Vegter, T. Bijvoets, M.C.G. Aalders, Remote spectroscopic identification of bloodstains, J. Forensic Sci. 56 (2011) 1471–1475.
[6] J.M. Chalmers, H.G.M. Edwards, M.D. Hargreaves, Infrared and Raman Spectroscopy in Forensic Science, John Wiley & Sons, Chichester, 2012.
[7] K. Virkler, I.K. Lednev, Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis, Anal. Chem. 81 (2009) 7773–7777.
[8] K. Virkler, I.K. Lednev, Raman spectroscopic signature of blood and its potential application to forensic body fluid identification, Anal. Bioanal. Chem. 396 (2010) 525–534.
[9] O.M.D. Lutz, G.K. Bonn, B.M. Rode, C.W. Huck, Reproducible quantification of ethanol in gasoline via a customized mobile near-infrared spectrometer, Anal. Chim. Acta 826 (2014) 61–68.
[10] E.J.N. Marques, S.T. Freitas, M.F. Pimentel, C. Pasquini, Rapid and non-destructive determination of quality parameters in the 'Tommy Atkins' mango using a novel handheld near infrared spectrometer, Food Chem. 197 (2016) 1207–1214.
[11] C.-M. Orphanou, The detection and discrimination of human body fluids using ATR FT-IR spectroscopy, Forensic Sci. Int. 258 (2015) e10–e16.
[12] K.M. Elkins, Rapid presumptive "fingerprinting" of body fluids and materials by ATR FT-IR spectroscopy, J. Forensic Sci. 56 (2011) 1580–1587.
[13] G.J. Edelman, R.J.M. Hoveling, M. Roos, T.G. Leeuwen, M.C.G. Aalders, Infrared imaging of the crime scene: possibilities and pitfalls, J. Forensic Sci. 58 (2013) 1156–1162.
[14] G.J. Edelman, V. Manti, S.M. Ruth, T.G. Leeuwen, M. Aalders, Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy, Forensic Sci. Int. 220 (2012) 239–244.
[15] E. Mistek, I.K. Lednev, Identification of species' blood by Attenuated Total Reflection (ATR) Fourier Transform Infrared (FT-IR) spectroscopy, Anal. Bioanal. Chem. 407 (2015) 7435–7442.
[16] L. Zhang, S. Zhang, M. Sun, Z. Wang, H. Li, Y. Li, G. Li, L. Lin, Blood species identification using Near-Infrared diffuse transmitted spectra and PLS-DA method, Infrared Phys. Technol. 76 (2016) 587–591.
[17] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213–225.
[18] T.S. Grobério, J.J. Zacca, É.D. Botelho, M. Talhavini, J.W.B. Braga, Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR, Forensic Sci. Int. 257 (2015) 297–306.
[19] M.Á. Fernández de la Ossa, J.M. Amigo, C. García-Ruiz, Detection of residues from explosive manipulation by Near Infrared hyperspectral imaging: a promising forensic tool, Forensic Sci. Int. 242 (2014) 228–235.
[20] C.S. Silva, F.S.L. Borba, M.F. Pimentel, M.J.C. Pontes, R.S. Honorato, C. Pasquini, Classification of blue pen ink using Infrared spectroscopy and Linear Discriminant Analysis, Microchem. J. 109 (2013) 122–127.
[21] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. lab. Syst. 57 (2001) 65–73.
[22] N.C. Silva, M.F. Pimentel, R.S. Honorato, M. Talhavini, A.O. Maldaner, F.A. Honorato, Classification of Brazilian and foreign gasolines adulterated with alcohol using Infrared spectroscopy, Forensic Sci. Int. 253 (2015) 33–42.
[23] C. Muehlethaler, G. Massonnet, P. Esseiva, Discrimination and classification of FTIR spectra of red, blue and green spray paints using a multivariate statistical approach, Forensic Sci. Int. 244 (2014) 170–178.
[24] L.R. Brito, M.P.F. Silva, J.J.R. Rohwedder, C. Pasquini, F.A. Honorato, M.F. Pimentel, Determination of detergent and dispersant additives in gasoline by ring-oven and near infrared hyperspectral imaging, Anal. Chim. Acta 863 (2015) 9–19.
[25] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: a practical guide, Wiley-Interscience Series on Laboratory Automation, John Wiley & Sons Ltd, New York, 1998.
[26] E.D. Moreira, M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection, Talanta 79 (2009) 1260–1264.
[27] J.T. Kuenstner, K.H. Norris, W.F. Mccarthy, Measurement of hemoglobin in unlysed blood by near-infrared spectroscopy, Appl. Spectrosc. 48 (1994) 484–488.
[28] J.J. Workman, L. Weyer, Pratical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy, second ed. CRC Press, Boca Raton, 2012.
[29] M.J.C. Pontes, R.K.H. Galvao, M.C.U. Araújo, P.N.T. Moreira, O.D. Pessoa Neto, G.E. José, et al., The successive projections algorithm for spectral variable selection in classification problems, Chemom. Intell. Lab. Syst. 78 (2005) 11–18.
[30] T.C.M. Pastore, J.W.B. Braga, V.T.R. Coradin, W.L.E. Magalhães, E.Y.A. Okino, J.A.A. Camargos, G.I.B. de Muniz, O.A. Bressan, F. Davrieux, Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: discrimination of true mahogany, cedar, andiroba, and curupixa, Holzforschung 65 (2011) 73–80.
[31] N. Armstrong, D.B. Hibbert, An introduction to Bayesian methods for analyzing chemistry data, Chemom. Intell. Lab. Syst. 97 (2009) 194–210.

APÊNDICE B – TCLE

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA FUNDAMENTAL

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO
(PARA MAIORES DE 18 ANOS OU EMANCIPADOS - Resolução 466/12)

Convidamos o (a) Sr. (a) para participar como voluntário (a) da pesquisa **"Desenvolvimento de métodos analíticos baseados em imagens hiperespectrais para fins forenses"**, que está sob a responsabilidade da pesquisadora **Carolina Santos Silva** (Rua Dr. Genaro Guimarães, 12, Casa Amarela, Recife-PE, CEP: 52070-040; tel: (81)92456456, (81)32680145, carolinasantosilva@gmail.com) e está sob a orientação de: Mª Fernanda Pimentel, Telefones para contato: (81)21267233; (81)21267235, e-mail (mfp@ufpe.br). Também participam também desta pesquisa:(Ricardo S. Honorato) Telefones: (21373969).

Este Termo de Consentimento pode conter informações que o/a senhor/a não entenda. Caso haja alguma dúvida, pergunte à pessoa que está lhe entrevistando para que o/a senhor/a esteja bem esclarecido (a) sobre sua participação na pesquisa. Após ser esclarecido (a) sobre as informações a seguir, caso aceite em fazer parte do estudo, rubrique as folhas e assine ao final deste documento, que está em duas vias. Uma delas é sua e a outra é do pesquisador responsável. Em caso de recusa o (a) Sr. (a) não será penalizado (a) de forma alguma. Também garantimos que o (a) Senhor (a) tem o direito de retirar o consentimento da sua participação em qualquer fase da pesquisa, sem qualquer penalidade.

INFORMAÇÕES SOBRE A PESQUISA:

Essa pesquisa tem como objetivo desenvolver métodos confiáveis, rápidos e não destrutivos para identificação e datação de vestígios de fluidos corporais em cenas de crimes utilizando imagens químicas das manchas dos fluidos doados em diferentes tecidos. Amostras de material biológico (sangue e/ou sêmen) serão depositadas em diferentes superfícies para identificação. A coleta dos materiais será realizada:

1. Sêmen:
    a. A coleta do material será realizada no centro de Pesquisa Vidas.
    b. O procedimento de coleta consiste na masturbação/autoestimulação voluntária do doador, não podendo fazer uso de água, lubrificante, saliva, sabonete ou qualquer outro produto químico durante o procedimento. O doador deve depositar o material no recipiente fornecido pela equipe.
    c. O período de participação do voluntário na pesquisa consiste em apenas um dia de visita para a doação do material.
    d. Serão colhidos aproximadamente 2,0 ml de sêmen por doador.
    e. Para os doadores voluntários de sêmen, não há riscos associados ao procedimento de doação.
2. Sangue:
    a. A coleta será realizada utilizando material descartável em ambiente limpo e apropriado. A coleta será realizada pelo farmacêutico participante da equipe.
    b. O período de participação do voluntário na pesquisa consiste em apenas um dia de visita para a doação do material.
    c. Serão colhidos aproximadamente 0,05 ml de sangue (equivalente a uma gota) por punção na falange distal do 4º dedo da mão direita.
    d. Para os doadores de sangue, o risco associado ao procedimento de coleta é um leve desconforto no local puncionado.

Não há benefícios diretos à saúde dos voluntários. Os benefícios dessa pesquisa estão associados a questões de segurança pública.

As informações desta pesquisa serão confidenciais e serão divulgadas apenas em eventos ou publicações científicas, não havendo identificação dos voluntários, a não ser entre os responsáveis pelo estudo, sendo assegurado o sigilo sobre a sua participação. Os dados coletados nesta pesquisa (material biológico), ficarão armazenados em computador pessoal, sob a responsabilidade do pesquisador, no endereço acima informado, pelo período de mínimo 5 anos.

O (a) senhor (a) não pagará nada para participar desta pesquisa. Se houver necessidade, as despesas para a sua participação serão assumidos pelos pesquisadores (ressarcimento de transporte e alimentação). Fica também garantida indenização em casos de danos, comprovadamente decorrentes da participação na pesquisa, conforme decisão judicial ou extra-judicial.

Em caso de dúvidas relacionadas aos aspectos éticos deste estudo, você poderá consultar o Comitê de Ética em Pesquisa Envolvendo Seres Humanos da UFPE no endereço: (Avenida da Engenharia s/n – 1º Andar, sala 4 - Cidade Universitária, Recife-PE, CEP: 50740-600, Tel.: (81) 2126.8588 – e-mail: cepccs@ufpe.br).

_____

(assinatura do pesquisador)

CONSENTIMENTO DA PARTICIPAÇÃO DA PESSOA COMO VOLUNTÁRIO (A)

Eu, _____, CPF _____, abaixo assinado, após a leitura (ou a escuta da leitura) deste documento e de ter tido a oportunidade de conversar e ter esclarecido as minhas dúvidas com o pesquisador responsável, concordo em participar do estudo **"Desenvolvimento**