



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO QUÍMICA FUNDAMENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

JULIETH PATRICIA GONZALEZ HERRERA

**DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS METABONÔMICOS PARA
ESTADIAMENTO DE FIBROSE PERIPORTAL EM PACIENTES COM
ESQUISTOSSOMOSE MANSONI USANDO CROMATOGRAFIA À LÍQUIDO E
QUIMIOMETRÍA**

Recife

2022

JULIETH PATRICIA GONZALEZ HERRERA

**DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS METABONÔMICOS PARA
ESTADIAMENTO DE FIBROSE PERIPORTAL EM PACIENTES COM
ESQUISTOSSOMOSE MANSONI USANDO CROMATOGRAFIA À LÍQUIDO E
QUIMIOMETRÍA**

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Química.

Área de concentração: Química Analítica

Orientador: Prof. Dr. José Licarion Pinto Segundo Neto

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

G643d Gonzalez Herrera, Julieth Patricia

Desenvolvimento e avaliação de modelos metabonômicos para estadiamento de fibrose periportal em pacientes com esquistossomose mansoni usando cromatografia à líquido e quimiometria / Julieth Patricia Gonzalez Herrera. – 2022.

82 f.: il., fig., tab.

Orientador: José Licarion Pinto Segundo Neto.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Química, Recife, 2022.

Inclui referências e anexos.

1. Química analítica. 2. Fibrose. 3. Esquistossomose. 4. Metabonômica. 5. HPLC-DAD. I. Segundo Neto, José Licarion Pinto (orientador). II. Título.

543

CDD (23. ed.)

UFPE- CCEN 2022 - 37

JULIETH PATRICIA GONZÁLEZ HERRERA

DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS METABONÔMICOS PARA ESTADIAMENTO DE FIBROSE PERIPORTAL EM PACIENTES COM ESQUISTOSSOMOSE MANSONI USANDO CROMATOGRAFIA À LÍQUIDO E QUIMIOMETRÍA

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Química.

Área de concentração: Química Analítica

Aprovada em: 28/01/2022

BANCA EXAMINADORA:

Prof. Dr. José Licarion Pinto Segundo Neto
Departamento de Química Analítica – UERJ
Orientador

Prof. Dr. Ricardo Oliveira da Silva
Departamento de Química Fundamental – UFPE
Examinador Interno

Prof. Dr. Edmundo Pessoa de Almeida Lopes
Departamento de Medicina Tropical - UFPE
Examinador Externo

AGRADECIMENTOS

A Deus que concedeu o desejo de meu coração de sair de meu país para continuar com os estudos e durante todo o processo há cuidado de cada detalhe, sem Ele nada poderia ter feito.

A meu professor e orientador Licarion Pinto que me apresentou o maravilhoso mundo da quimiometria e quem desde o primeiro semestre estava ensinando-me e apoiando-me, sempre disposto e paciente.

Aos meus pais, Osíris e Alejandro, minha irmã Kleitin, meu irmão Alejandro David, que mesmo na distância sempre estiveram me incentivando e apoiando.

A minha família brasileira do coração da IPRVarzea, os pastores João Luiz e Silvia, todos os irmãos e irmãs e minhas tias, pelo cuidado, carinho e ensino de português.

Aos meus colegas e amigos do LabMeQ, especialmente a Abinoan, Marcelino, Helder, Milena, Tati e Luísa pela companhia, acolhimento e por tornar esta etapa bem mais leve.

Ao meu colega Ian pelos conhecimentos passados e pela grande ajuda na operação do equipamento na central analítica no DQF.

Aos técnicos Júlio e Patrícia pela ajuda na reserva e operação do equipamento no CENAPESQ.

À central analítica do DQF/UFPE e ao CENAPESQ/UFRPE pela estrutura e equipamentos disponibilizados.

À CAPES pelo fomento da bolsa concedida e a todos os professores do DQF com os que tive o prazer de aprender mais de química durante o mestrado.

Minha gratidão a todos que de alguma forma contribuíram para realização do mestrado e deste trabalho.

“Portanto, se vocês comem, ou bebem ou fazem qualquer outra coisa, façam tudo para a glória de Deus. Porque dele, e por meio dele, e para ele são todas as coisas”(BÍBLIA, 1Coríntios, 10,31; Romanos, 11,36).

RESUMO

A esquistossomose é uma das doenças de maior prevalência entre aquelas veiculadas às coleções hídricas, sendo principalmente um risco nas populações de área rurais e das periferias urbanas. No Brasil estima-se que cerca de 1,5 milhões de pessoas vivem em áreas sob o risco de contrair a doença. Portanto a determinação da presença e gravidade da fibrose hepática é essencial em especial para a população de difícil acesso a exames existentes apenas em grande centro. Para contemplar o estadiamento da doença nesta população, estudos estão surgindo na tentativa de desenvolver métodos que sejam capazes de identificar e avaliar a fibrose periportal por meio de marcadores biológicos presentes no sangue. Por essa razão, este trabalho objetivou desenvolver e avaliar modelos metabonômicos para estadiamento de fibrose periportal em pacientes com esquistossomose mansônica usando cromatografia à líquido e quimiometria, tendo a ultrassonografia como padrão de referência. O estudo foi realizado com 94 amostras de soro, sendo 19 de pacientes com o estágio leve da doença, 30 intermediário e 45 avançado. As condições de análise das amostras foram previamente otimizadas. No preparo de amostra foram testados os métodos DLLME e QuEChERS. Já para a fase móvel da análise cromatográfica, foram testados gradientes de acetonitrila e metanol. A otimização revelou que o melhor preparo de amostra foi o DLLME e o melhor gradiente da fase móvel foi o metanol. Com as condições de trabalho definidas foram analisadas as 94 amostras utilizando um HPLC-DAD. Os dados obtidos foram pré-processados, usando o algoritmo Icoshift para correção de deslocamento de pico e posteriormente foram submetidos a análise multivariada exploratória e classificatória através do software MATLAB R2010a. A análise por ROBPCA não indicou a presença de amostras anômalas. Dentre os modelos metabonômicos construídos, o melhor foi o GA-LDA que obteve 84% de sensibilidade, 83% de especificidade e 83% de exatidão para discriminação de pacientes com FPP leve daqueles com FPP avançada. Outro modelo que apresentou bons resultados foi o DD-SIMCA que obteve 92% de sensibilidade, 63% de especificidade e 75% de exatidão. Desta forma o presente trabalho desenvolveu modelos que podem ser utilizados para auxiliar o diagnóstico e controle dinâmico da doença utilizando amostras de soro. O diagnóstico usando marcadores biológicos no sangue possibilita um acompanhamento mais frequente da doença e o monitoramento de pacientes com dificuldade de se deslocar para realizar a ultrassonografia.

Palavras-chave: fibrose; esquistossomose; metabonômica; HPLC-DAD; quimiometria.

ABSTRACT

Schistosomiasis is one of the most prevalent diseases among those transmitted to water collections, being mainly a risk in populations in rural areas and urban peripheries. In Brazil it is estimated that around 1.5 million people live in areas at risk of contracting the disease. Therefore, determining the presence and severity of liver fibrosis is essential. Studies are emerging in an attempt to develop non-invasive methods that are able to identify and assess periportal fibrosis through biological markers present in the blood. For this reason, this study aimed to develop and evaluate metabonomic models for staging periportal fibrosis in patients with schistosomiasis mansoni using liquid chromatography and chemometry, using ultrasound as the reference standard. The study was carried out with 94 serum samples, 19 from patients with mild stage of the disease, 30 intermediate and 45 advanced. The analysis conditions of the samples were previously optimized. In preparing the sample, the DLLME and QuEChERS methods were tested. For the mobile phase of the chromatographic analysis, gradients of acetonitrile and methanol were tested. Optimization revealed that the best sample preparation was DLLME and the best mobile phase gradient was methanol. With the defined working conditions, the 94 samples were analyzed using an HPLC-DAD. The data obtained were pre-processed, using the Icoshift algorithm for peak displacement correction, and later submitted to exploratory and classificatory multivariate analysis using the MATLAB R2010a software. ROBPCA analysis did not indicate the presence of anomalous samples. Among the constructed metabonomic models, the best was the GA-LDA, which obtained 84% sensitivity, 83% specificity and 83% accuracy for discriminating patients with mild PPF from those with advanced PPF. Another model that showed good results was the DD-SIMCA, which obtained 92% sensitivity, 63% specificity and 75% accuracy. Thus, the present work has developed models that can be used to aid the diagnosis and dynamic control of the disease using serum samples. Diagnosis using biological markers in the blood allows for more frequent follow-up of the disease and monitoring of patients who have difficulty traveling for ultrasound.

Keywords: fibrosis; schistosomiasis; metabonomics; HPLC-DAD; chemometrics.

LISTA DE FIGURAS

Figura 1 –	Representação da aparência do fígado saudável, com fibrose e com cirrose	19
Figura 2 –	Ciclo de vida e estágios de desenvolvimento do Schistosoma mansoni. (A) Vermes adultos. (B) Ovos. (C) Miracídios. (D) Cercárias	21
Figura 3 –	Classificação ultrassonográfica de FPP – Niamey	23
Figura 4 –	Principais etapas do método QuEChERS	27
Figura 5 –	Etapas envolvidas no método DLLME	28
Figura 6 –	O princípio do conjunto de dados HPLC-DAD	30
Figura 7 –	Ilustração da PCA: (esquerda) tipos de outlier; (direita) mapa de outlier: gráfico de distâncias ortogonais versus distâncias de escores	33
Figura 8 –	Esquema da organização dos dados para a construção do modelo de classificação usando PLS-DA	35
Figura 9 –	Esquema de aplicação do DD-SIMCA na autenticação e identificação de adulterantes em óleo de coco	37
Figura 10 –	Esquema do QuEChERS nas amostras da otimização	40
Figura 11 –	Esquema do DLLME nas amostras da otimização	41
Figura 12 –	Gradientes de concentração do solvente orgânico, linear (esquerda) e curvo (direita)	42
Figura 13 –	HPLC-DAD utilizado nas análises das amostras do estudo	42
Figura 14 –	Cromatograma, com os comprimentos de onda somados, das amostras controle e hepatite nas duas versões do método QuEChERS, Q1 vs Q2. Usando como fase móvel, gradiente	44
Figura 15 –	Gráficos em 3D da amostra controle em dois tipos de gradientes de concentração da fase orgânica (ACN) na versão Q1 do método QuEChERS	45
Figura 16 –	Comparação dos métodos DLLME vs QuEChERS na amostra controle após subtração do branco.	45

Figura 17 –	Comparação dos gradientes MeOH vs ACN no método DLLME	46
Figura 18 –	Cromatogramas com os dados brutos (a) 210 nm e (c) somando os comprimentos de onda. E cromatogramas após alinhamento (b) a 210 nm e (d) somando os comprimentos de onda.	47
Figura 19 –	Cores apresentadas nas amostras do estudo. (a) amostra original, (b) após precipitação de proteínas e (c) extração.	48
Figura 20 –	Cromatogramas e gráficos de escores dos dados brutos somados entre 200-800 nm (a) e (b). E entre 200-400 nm (c) e (d)	49
Figura 21 –	respectivamente. Escores dos dados (a) 210 nm e (b) somando os comprimentos de onda. Foram selecionadas 5 PC's com uma variância explicada de 78% para os dados a 210 nm e 96,1% para os dados somados 200-400nm.	49
Figura 22 -	Dados contendo amostras da classe C e classe EF, com os comprimentos de onda somados. (a) Cromatograma. (b) gráfico de escores.	50
Figura 23 –	Analise de componentes principais robusto e clássico	51
Figura 24 –	Gráfico de escores (a) e pesos (b) e (c) no algoritmo PLS-DA	52
Figura 25 –	Gráficos do modelo DD-SIMCA para o grupo de treinamento da classe C (a) e o grupo teste (b). Em verde as amostras da classe de treinamento, em vermelho as amostras anômalas preditas pelo modelo.	53
Figura 26 –	Gráficos do modelo DD-SIMCA para o grupo de treinamento da classe EF (a) e o grupo teste (b). em verde as amostras da classe de treinamento, em laranja as amostras classificadas como extremos, em vermelho as amostras anômalas preditas pelo modelo.	54
Figura 27 –	Gráficos do modelo GA-LDA. (a) escores. (b) Cromatograma com as variáveis selecionadas pelo modelo. (c) Zoom do cromatograma indicando os tempos de retenção selecionados.	56

- Figura 28 – Dados contendo amostras da classe C e classe EF, com o sinal 58 entre 5,53 – 6,34 min, removido. (a) Cromatograma. (b) gráfico de escores.
- Figura 29 – Gráfico de escores (a) e pesos (b) e (c) do modelo PLS-DA nos 58 dados reduzidos.
- Figura 30 – Gráficos do modelo DD-SIMCA para o grupo de treinamento da 60 classe C (a) e o grupo teste (b). em verde as amostras da classe de treinamento, em vermelho as amostras anômalas preditas pelo modelo.
- Figura 31 - Gráficos do modelo DD-SIMCA para o grupo de treinamento da 61 classe EF (a) e o grupo teste (b). em verde as amostras da classe
- Figura 32 – Dados contendo amostras da classe C, classe D e classe EF, com 64 o sinal entre 5,53 – 6,34 min, removido. (a) Cromatograma. (b) gráfico de escores.
- Figura 33 – Predição das amostras da classe D no modelo DD-SIMCA da 66 classe EF.

LISTA DE TABELAS

Tabela 1 –	Classificação da Cromatografia pelas formas físicas das fases moveis e estacionarias	29
Tabela 2 –	Exemplo de matriz de contingência	38
Tabela 3 –	Distribuição das amostras segundo classificação de Niamey com US	39
Tabela 4 –	Matriz de contingência e métricas de desempenho obtido do algoritmo PLS-DA	52
Tabela 5 –	Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelagem da classe C	53
Tabela 6 –	Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe EF	55
Tabela 7 –	Matriz de contingência e métricas de desempenho obtido do algoritmo GA-LDA	57
Tabela 8 –	Matriz de contingência e métricas de desempenho do modelo PLS-DA obtidos nos dados com sinal removido	59
Tabela 9 –	Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe C	60
Tabela 10 –	Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe EF	61
Tabela 11 –	Comparação de resultados dos modelos construídos para as classes C e EF	62
Tabela 12 –	Predição das amostras da classe D no modelo GA-LDA construído para as classes C e EF. Na coluna de predição 1 corresponde a classe C e 2 a classe EF	65
Tabela 13 –	Predição das amostras da classe D no modelo DD-SIMCA da classe EF. Na coluna de predição os critérios pertencem e não pertencem corresponde a classe modelada (EF).	66

LISTA DE ABREVIATURAS E SIGLAS

CEP	Comitê de Ética em Pesquisa
DAD	Detetor de Arranjo de Diodos
DD-SIMCA	do inglês <i>Data Driven Soft Independent Modeling of Class Analogy</i>
DLMME	Microextração Líquido-Líquido Dispersiva
EHI	Esquistossomose Hepatointestinal
EH	Esquistossomose Hepática
EHE	Esquistossomose Hepatoesplênica
ER	Erro
EX	Exatidão
FPP	Fibrose Periportal
FP	Falso Positivo
FN	Falso Negativo
GA	Algoritmo Genético, do inglês <i>Genetic Algorithm</i>
HPLC	Cromatografia Líquida de Alta Eficiência, do inglês <i>High Performance Liquid Chromatography</i>
LDA	Análise Discriminante Linear, do inglês <i>Linear Discriminant Analysis</i>
PC	Componente Principal, do inglês <i>Principal Component</i>
PCA	Análise de Componentes Principais, do inglês <i>Principal Component Analysis</i>
PLS-DA	Análise discriminante por Mínimos Quadrados Parciais, do inglês <i>Partial Least Squares Discriminant Analysis</i>
QuEChERS	do inglês, <i>Quick, Easy, Cheap, Effective, Rugged, Safe</i>
ROBPCA	PCA Robusta, do inglês <i>Robust PCA</i>
US	Ultrassonografia
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
VL	Variable Latente

SUMÁRIO

1	INTRODUÇÃO	15
1.1	OBJETIVO GERAL	17
1.1.1	Objetivos Específicos	17
2	REVISÃO DE LITERATURA	18
2.1	FIBROSE	18
2.2	ESQUISTOSSOMOSE	19
2.3	DIAGNOSTICO	22
2.4	METABONÔMICA	24
2.5	PREPARO DE AMOSTRA	25
2.6	CROMATOGRAFIA À LÍQUIDO	28
2.7	QUIMIOMETRIA	31
2.7.1	Métodos Não Supervisionados	31
2.7.2	Métodos Supervisionados	34
2.7.3	Figuras de Mérito	37
3	METODOLOGIA	39
3.1	POPULAÇÃO DE ESTUDO	39
3.2	OTIMIZAÇÃO	40
3.2.1	Preparo de Amostra	40
3.2.1.1	QuEChERS	40
3.2.1.2	DLLME	41
3.2.2	Cromatografia	41
3.3	MODELOS QUIMIOMÉTRICOS	43
4	RESULTADOS	44
4.1	OTIMIZAÇÃO	44
4.2	ESTUDO METABONÔMICO	46
4.2.1	Pré-processamento dos dados cromatográficos	46
4.2.2	Análise exploratória	48
4.2.3	Análise classificatória	51
5	CONCLUSÕES E PERSPECTIVAS FUTURAS	68
	REFERÊNCIAS	69

ANEXO A – PARECER CONSUBSTANCIADO DO CEP	78
ANEXO B – CROMATOGRAMAS 2D E 3D DAS AMOSTRAS	82
CONTROLE E HEPATITE NOS 4 GRADIENTES	
ANALISADOS - DLLME	

1 INTRODUÇÃO

O fígado tem um grande número de funções digestivas e excretórias, bem como armazenamento de nutrientes e funções metabólicas, síntese de novas moléculas e purificação de produtos químicos tóxicos; é capaz de desintoxicar o corpo humano de várias substâncias por meio de sua alteração estrutural, tornando-as menos tóxicas ou mais fáceis de excretar (GUERRA *et al.*, 2016). Quando é atingido repetidamente por um amplo espectro de lesões hepáticas, o fígado produz uma resposta de cicatrização nas feridas conhecida como fibrose. A fibrose hepática resulta de um desequilíbrio entre a síntese e a degradação da matriz extracelular (CAVIGLIA *et al.*, 2017).

Quando a fibrose é causada pelo *Schistosoma mansoni* refere-se a esquistossomose da qual a fibrose periportal (FPP) é congênita (BRASIL, 2014). A transmissão da esquistossomose foi relatada em 78 países. É considerada uma doença negligenciada, conhecida como doença da pobreza, e a incidência da infecção concentra-se em comunidades particularmente pobres com dependência de águas superficiais, muitas vezes contaminadas com fezes de indivíduos infectados e colonizadas por caramujos que agem como hospedeiros intermediários do esquistosomo (LAGO *et al.*, 2018).

Entre as parasitoses que afetam o homem, a esquistossomose é uma das mais disseminadas no mundo. De acordo com a Organização Mundial de Saúde – OMS, ocupa o segundo lugar depois da malária, pela sua importância e repercussão socioeconômica (BRASIL, 2014).

No Brasil, a Secretaria de Vigilância em Saúde, do Ministério da Saúde, estimou 1,5 milhão de pessoas infectadas pelo *Schistosoma mansoni* em 2019. O agravo está presente, de forma endêmica, em nove estados, persistindo como foco em outras nove Unidades da Federação (UFs) onde se observa a manutenção de suas características epidemiológicas (SANTOS, I. G. De A. *et al.*, 2021).

A doença ainda se configura como problema de saúde pública, do total de pessoas afetadas no país, 80% delas são da região Nordeste. Acrescentem-se as graves formas clínicas de expressão da esquistossomose, o que contribui para sua magnitude e transcendência. Pernambuco é um dos estados nordestinos com maior prevalência de esquistossomose (BRITO; SILVA, M. B. A.; QUININO, 2020).

As infecções por esquistossomose são diagnosticadas por uma história clínica de contato com uma fonte de água doce, como rios ou riachos de áreas endêmicas, seguida pela detecção de ovos nas fezes (GOUVEIA *et al.*, 2017). Já para a presença de fibrose, a biópsia hepática é

atualmente considerada o padrão ouro para seu diagnóstico e avaliação histológica. No entanto, devido à sua natureza invasiva, uma biópsia não é adequada para fins de triagem e não pode ser implementada no início do diagnóstico de pacientes em potencial. É principalmente reservada para pacientes com alto risco de doença hepática e para determinar sua extensão (HEYENS *et al.*, 2021). O risco de complicações graves inclui, sangramento, erros de amostragem, variabilidade na interpretação histopatológica e custos econômicos (TADOKORO, 2021).

Em virtude das complicações da biópsia, o diagnóstico de FPP é realizado por métodos de imagem como ultrassonografia (US), tomografia computadorizada e ressonância magnética. Destes, o US é o mais utilizado no Brasil, devido ao seu baixo custo e por ser tão sensível quanto a biópsia. Apesar do uso generalizado da US para o diagnóstico e monitoramento das alterações causadas pela fibrose, seu uso apresenta algumas limitações, como a baixa sensibilidade nas formas iniciais da doença; a necessidade de examinador treinado e o fato de não estar disponível em todos os centros, principalmente nos localizados na zona rural (BARRETO, 2017).

Em decorrência dessas limitações da US e da biópsia, estudos estão surgindo na tentativa de desenvolver métodos não invasivos capazes de identificar e avaliar a FPP por meio de marcadores biológicos. Marcadores séricos, proteínas responsáveis pela fibrogênese, ácido hialurônico, plaquetas, laminina e colágeno tipo IV, tem sido estudada na avaliação da fibrose. Entretanto estudos sobre a medida de FPP por biomarcadores em pacientes com esquistossomose é relativamente pequena (BARRETO, 2017; MARIA SOARES TOJAL DE BARROS LIMA *et al.*, 2008; MEDEIROS *et al.*, 2014).

Dentro deste cenário, é preciso desenvolver ferramentas para o diagnóstico diferencial da doença utilizando métodos não invasivos que além de discriminá-lo, possam ser empregados para o monitoramento terapêutico.

Em 2020 Lima Rodrigues, do nosso grupo de pesquisa, apresentou um estudo no qual construiu modelos metabonômicos, para diagnóstico e classificação de FPP, baseados na técnica de RMN de ^1H . Nesse contexto e com o objetivo de continuar com a pesquisa na área propõe-se aqui um estudo metabonômico utilizando dados de cromatografia líquida de alta eficiência. Visto que na cromatografia é possível identificar marcadores biológicos em concentração menor que no RMN. Portanto, sendo esta técnica mais sensível, pode render resultados mais interessantes ao permitir enxergar metabólitos em maior quantidade (CANUTO *et al.*, 2018). Por outro lado, o HPLC-DAD é fácil de obter na maioria dos laboratórios e

apresenta menor custo associado, o que o torna uma ferramenta acessível para hospitais e clínicas.

1.1 OBJETIVO GERAL

Desenvolver e avaliar modelos metabonômicos, a partir de dados de cromatografia líquida de alta eficiência com detector por arranjo de diodos de amostras de soro, para estadiamento de fibrose periportal em pacientes com esquistossomose, tendo a classificação ultrassonográfica de Niamey como padrão de referência.

1.1.1 Objetivos Específicos

- Otimizar o preparo e as análises cromatográficas das amostras de soro;
- Realizar análise exploratória utilizando a PCA Robusta para detecção de amostras anômalas;
- Construir e validar modelos metabonômicos, a partir dos dados obtidos das análises cromatográficas, que permitam classificar os padrões de fibrose periportal;
- Avaliar a eficiência de modelos de classe única para classificação de fibrose periportal;
- Avaliar a eficiência de modelos discriminantes para classificação de fibrose periportal;
- Avaliar a eficiência de modelos não lineares para classificação de fibrose periportal;
- Comparar os resultados de eficiência de todos os modelos e definir o mais adequado para o caso em estudo;

2 REVISÃO DA LITERATURA

2.1 FIBROSE

O fígado é o maior órgão do corpo humano, depois da pele, representando 2,5 a 4,5% da massa corporal. É um órgão muito complexo que realiza várias funções vitais, essenciais à manutenção da homeostasia corporal. Dentre estas funções destacam-se a regulação do metabolismo de diversos nutrientes, papel imunológico, síntese proteica e de outras moléculas, armazenamento de vitaminas e ferro, degradação hormonal e a inativação e excreção de drogas e toxinas (ORIÁ *et al.*, 2016).

De forma aproximada, o fígado exibe uma forma triangular e uma coloração rosa a marrom. É uma estrutura altamente vascular que: apresenta uma consistência macia; está posicionado no quadrante superior direito da cavidade abdominal, abaixo do diafragma; é protegido pela caixa torácica; mantém sua posição com o apoio de vários ligamentos (ABDEL-MISIH; BLOOMSTON, 2010; GUERRA *et al.*, 2016; TAHARA; SHIBATA, 2016).

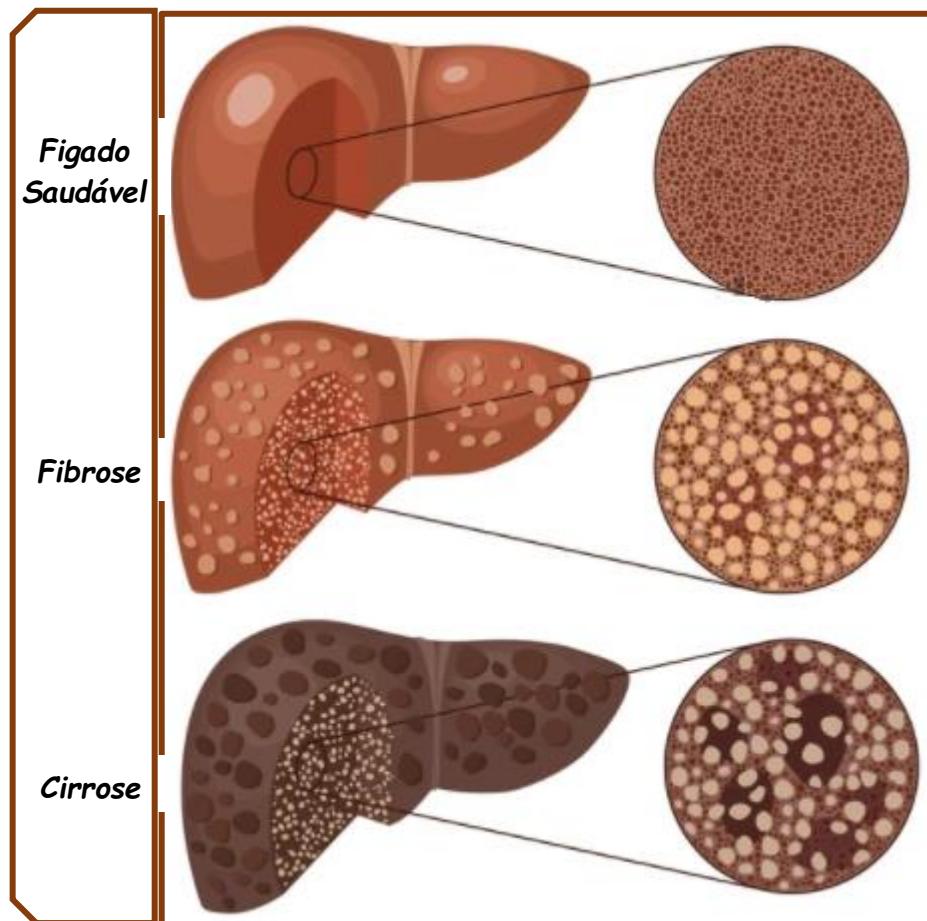
Lesões crônicas produzidas por uma variedade de agressões, como doenças metabólicas, infecções virais, abuso na ingestão de álcool, drogas e ataque autoimune aos hepatócitos ou defeitos congênitos, produzem uma resposta comum no fígado chamada fibrose (CEQUERA; GARCÍA LEÓN MÉNDEZ, DE, 2014).

A fibrose hepática envolve múltiplos eventos celulares e moleculares que induzem a deposição excessiva de proteínas da matriz extracelular que distorcem a arquitetura do parênquima hepático, cujo estágio final é conhecido como cirrose (SENTÍES-GÓMEZ *et al.*, 2005). A fibrose hepática é considerada um processo de reparo de tecidos altamente complexo que aparece após dano hepatocelular sustentado. Durante esse processo reparador, ocorre uma reação inflamatória, juntamente com um depósito extracelular da matriz composto de colágeno não fibrilar, que contribui para recompor o lóbulo hepático estragado. No entanto, se a necrose hepática for mantida, esse processo reparador continuará. Com o tempo, a capacidade de regeneração hepática diminui, enquanto a produção de matriz extracelular, que é depositada de maneira desorganizada no lóbulo hepático, aumenta consideravelmente (DOMÍNGUEZ; COLMENERO; BATALLER, 2009).

Como resultado de todas essas alterações hepáticas, uma cicatriz fibrosa é formada ao redor da área lesionada e se mantida pode levar à cirrose (uma consequência do estágio final da fibrose) caracterizada pela presença de nódulos regenerativos circundados por fibrose (FRIEDMAN, 2008; TSUKADA; PARSONS; RIPPE, 2006). A cirrose e sua distorção vascular

associada, resultam em uma grave perda das funções hepáticas (GUERRA *et al.*, 2016; SCHUPPAN; AFDHAL H, 2008).

Figura 1 – Representação da aparência do fígado saudável, com fibrose e com cirrose.



Fonte: (BOLBOT, 2021). Modificado pela autora.

2.2 ESQUISTOSSOMOSE

A esquistossomose mansoni é uma doença infecto parasitária provocada por vermes do gênero *Schistosoma*, que têm como hospedeiros intermediários caramujos de água doce do gênero *Biomphalaria*, e pode progredir de formas assintomáticas até formas clínicas extremamente graves (BRASIL, 2014).

A esquistossomose é uma das doenças de maior prevalência entre aquelas veiculadas as coleções hídricas, sendo principalmente um risco nas populações de área rurais e das periferias urbanas. De acordo com a Organização Mundial de Saúde – OMS, a doença afeta 200 milhões de pessoas, sendo uma ameaça mais elevada para indivíduos residentes em área de risco (BRASIL, 2014). No Brasil estima-se que cerca de 1,5 milhões de pessoas vivem em áreas sob o risco de contrair a doença. Os estados das regiões Nordeste e Sudeste são os mais afetados

sendo a ocorrência diretamente ligada à presença dos moluscos transmissores. Atualmente, a doença é detectada em todas as regiões do país. As áreas endêmicas e focais abrangem os Estados de Alagoas, Bahia, Pernambuco, Rio Grande do Norte (faixa litorânea), Paraíba, Sergipe, Espírito Santo e Minas Gerais (predominantemente no Norte e Nordeste do Estado) (MINISTÉRIO DA SAÚDE, 2017).

Em Pernambuco, a esquistossomose é considerada endêmica. Regiões onde o saneamento básico é precário constituem o cenário perfeito para sua ocorrência e aliado a isso, a migração de pessoas parasitadas tem promovido a expansão da esquistossomose para áreas indenes, como os centros urbanos de cidades do interior e localidades litorâneas do estado de Pernambuco (SOUZA GOMES, DE et al., 2016).

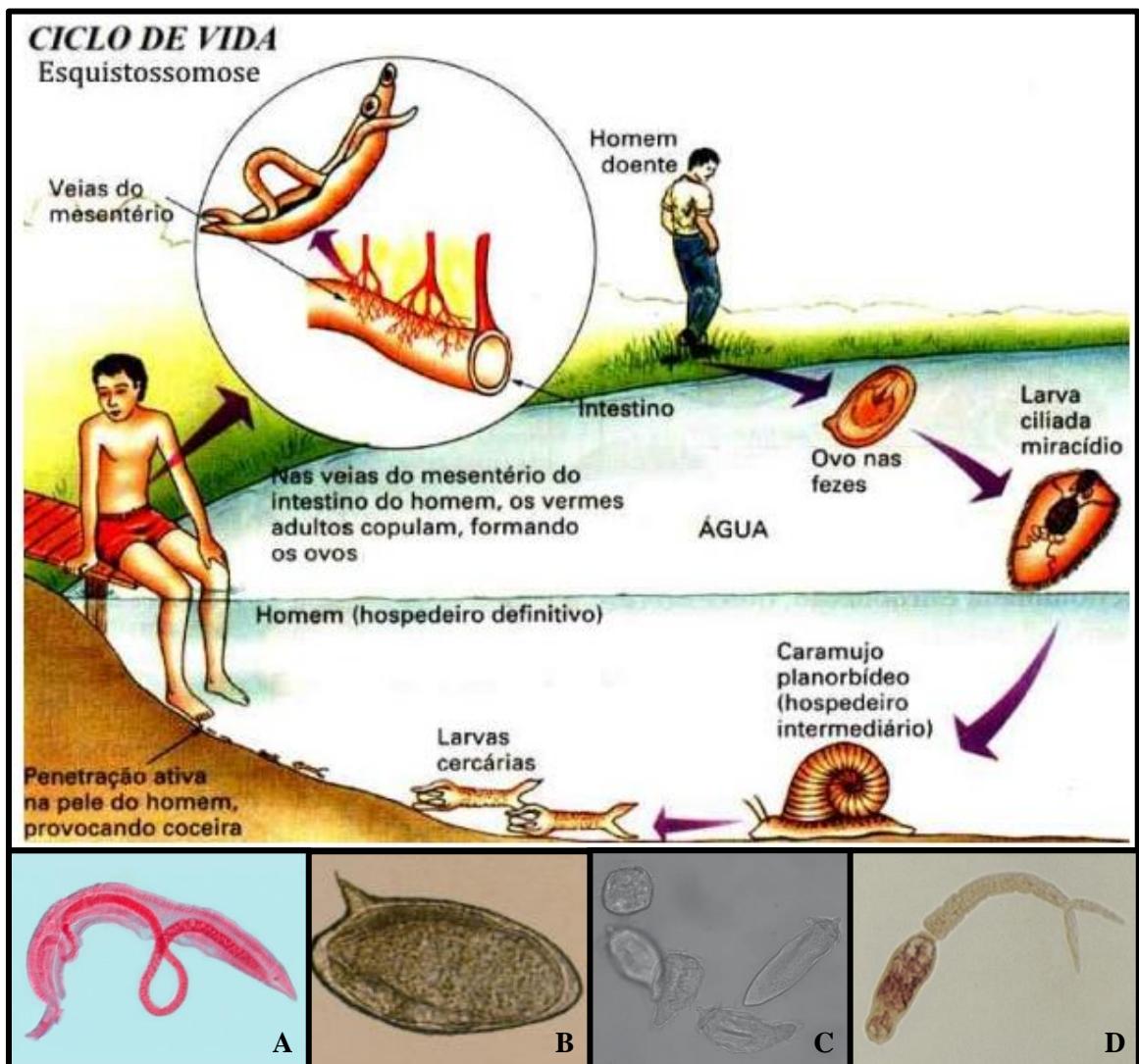
Três espécies principais de esquistossomos infectam os seres humanos, *Schistosoma haematobium*, *Schistosoma japonicum* e *Schistosoma mansoni*. Esta última espécie é a única descrita no Brasil, em virtude da inexistência de moluscos suscetíveis aos demais helmintos. O verme adulto tem seu habitat nas vênulas tributárias do sistema porta, particularmente das veias mesentéricas superiores e inferiores, do plexo hemorroidário e mesmo da porção intra-hepática da veia porta. No interior desses vasos, encontram-se geralmente o macho e a fêmea, acasalados. Os ovos podem ter três destinos diferentes: podem ser liberados pelas fezes; podem permanecer no intestino e causar uma infecção granulomatosa local; ou podem migrar para o fígado e se alojar nos vasos pré-sinusoidais, provocando granuloma e fibrose periportal.

Os ovos que chegam à água doce eclodem, liberando miracídios ciliados de vida livre que infectam um caracol hospedeiro adequado. No caracol, o parasita sofre replicação assexuada através dos estágios de esporocistos de mãe e filha, eventualmente liberando dezenas de milhares de cercárias (a forma infecciosa para os seres humanos) na água. Ao penetrar na pele, as cercárias se transformam em esquistossômulos. Estes secretam enzimas proteolíticas que vão destruir a matriz conjuntiva e penetrar na circulação periférica. Após penetrar a circulação, o verme imaturo vai passar, pelo coração, pelos pulmões e vai migrar para o fígado. Uma vez no sistema intra-hepático, alimentam-se e desenvolvem-se, transformando-se em formas unissexuadas, machos e fêmeas, 28 a 30 dias após a penetração. A partir deste ponto migram, acasalados, via sistema porta, até o território da artéria mesentérica inferior, onde farão a oviposição (BRASIL, 2014; COLLEY et al., 2014; KOSMINSKY, 2020; SOUZA, W. et al., 2011).

A classificação da esquistossomose compreende a forma aguda na fase inicial, e a forma crônica na fase tardia. A forma crônica é dividida e classificada em 3 formas clínicas:

esquistossomose hepatointestinal (EHI), esquistossomose hepática (EH) e esquistossomose hepatoesplênica (EHE) (BRASIL, 2014).

Figura 2 – Ciclo de vida e estágios de desenvolvimento do *Schistosoma mansoni*. (A) Vermes adultos. (B) Ovos. (C) Miracídios. (D) Cercárias.



Fonte: (BRASIL, 2014); (KOSMINSKY, 2020). Modificado pela autora.

A EHI, é a forma mais encontrada nas regiões endêmicas. Caracteriza-se por sintomas digestivos como náuseas, vômitos, flatulências, podendo ocorrer surtos diarreicos ou constipação intestinal crônica. Na forma EH, ocorre o estabelecimento da hepatomegalia, presença de fibrose, mas sem esplenomegalia e sem hipertensão portal. A forma mais severa da doença, a EHE, é caracterizada por ter o fígado e baco aumentados de volume, com presença de fibrose hepática moderada a acentuada. Alguns casos de EHE evoluem para hipertensão portal, resultando em descompensação, varizes gastroesofágicas, as quais ao romper causam hemorragia digestiva alta, podendo evoluir para coma hepático e morte, representando a causa

usual de óbitos pela esquistossomose (BARRETO, 2018; BRASIL, 2014; KOSMINSKY, 2020).

2.3 DIAGNOSTICO

A determinação da presença e gravidade da fibrose hepática é essencial no diagnóstico e estadiamento da doença (YU et al., 2015). Biópsia hepática, exame de ultrassom do abdômen, marcadores séricos, contagem de plaquetas e, medição de rigidez hepática (elastografia) têm sido usados para avaliar a fibrose que se desenvolve na doença hepática crônica (MEDEIROS et al., 2014).

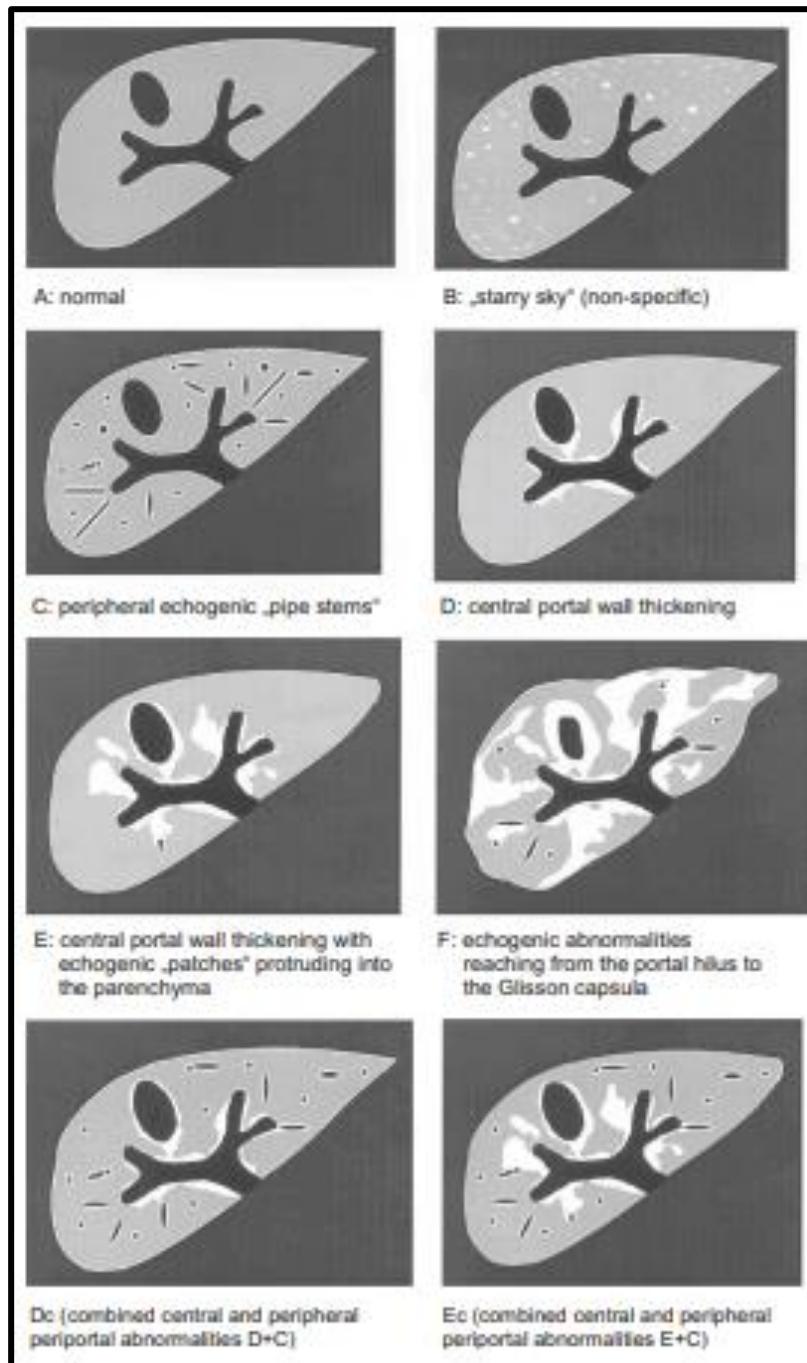
A biópsia hepática é considerada o método padrão ouro para classificar a fibrose, pois permite que os médicos obtenham informações de diagnóstico, não apenas sobre fibrose, mas também sobre outros processos de dano, como necrose, inflamação, esteatose e depósitos de ferro ou cobre, entre outros (CEQUERA; GARCÍA LEÓN MÉNDEZ, DE, 2014).

No entanto, a biópsia é limitada em sua precisão, pois a fibrose hepática é um processo heterogêneo que aumenta as chances de erro de amostragem. Outras desvantagens incluem natureza invasiva, a baixa qualidade da amostra e dependência da experiência do patologista (erro do observador). Existem riscos associados à obtenção de uma biópsia hepática, variando entre dor, hipotensão, sangramento peritoneal, infecção, danos ao sistema biliar e morte. Os resultados gerados tendem a ser frequentemente representativos quando se trata de uma doença relativamente avançada (CEQUERA; GARCÍA LEÓN MÉNDEZ, DE, 2014; YU et al., 2015).

Especialmente em áreas onde a esquistossomose é endêmica, a ultrassonografia abdominal superior (US) tornou-se a ferramenta diagnóstica mais útil para diagnosticar e quantificar FPP. Além disso, é a mais utilizada no Brasil devido a seu baixo custo, natureza não invasiva e por ser tão sensível quanto a biopsia (BARRETO, 2017; DOMINGUES; MEDEIROS, DE; ALMEIDA LOPES, DE, 2011).

A avaliação ultrassonográfica para o diagnóstico da FPP é composta por duas análises: a quantitativa que mede a espessura da FPP, e a qualitativa que compara o fígado examinado com padrões de acometimento pela fibrose através da classificação de Niamey- protocolo padronizado da OMS (NIAMEY WORKING GROUP, 2000). A classificação de Niamey considera seis padrões de FPP denominados: A (ausência de fibrose); B (duvidosa); C (periférica); D (central), E (avançada) e F (muito avançada) (BARRETO, 2018; RICHTER et al., 2001). Além disso, existe uma combinação de padrões que refletem anormalidades periféricas e centrais, sendo possíveis as combinações Dc e Ec (SCHEICH, EL et al., 2014), ver figura 3.

Figura 3 – Classificação ultrassonográfica de FPP - Niamey



Fonte: (SCHEICH, EL *et al.*, 2014)

Apesar da ampla utilização da US para o diagnóstico e monitoramento das alterações causadas pela fibrose, seu uso apresenta algumas limitações, como a baixa sensibilidade nas formas iniciais da doença, necessidade de equipamento disponível, examinadores qualificados e não está disponível em todos os centros, principalmente nos localizados na zona rural

(BARRETO, 2017; DOMINGUES; MEDEIROS, DE; ALMEIDA LOPES, DE, 2011; MEDEIROS et al., 2014)

O estudo de novos métodos de diagnóstico e avaliação de FPP é importante, a fim de fornecer subsídios para o desenho de estratégias de tratamento e prevenção da evolução da doença. Estudos estão surgindo na tentativa de desenvolver métodos não invasivos que sejam capazes de identificar e avaliar FPP por meio de marcadores séricos. Diferentes marcadores biológicos têm sido associados com o desenvolvimento e regulação da fibrose hepática em algumas doenças hepáticas crônicas (BARRETO, 2017).

Existem vários biomarcadores que se destacam por apresentarem correlação com a histologia hepática e pela facilidade do seu uso clínico, como a aspartato aminotransferase (AST), alanina amino- transferase (ALT), gama-GT (GGT), fosfatase alcalina (FA), albumina, bilirrubina e plaquetas. Quando ocorrem lesões ou destruição das células hepáticas, há liberação destas enzimas para a circulação sanguínea, sendo que a sua localização auxilia inferir o diagnóstico e prognóstico de patologias hepáticas (JESUS; SOUSA; BARCELOS, 2014; MARIA SOARES TOJAL DE BARROS LIMA et al., 2008).

Diferentes marcadores biológicos têm sido associados ao desenvolvimento e regulação da fibrose hepática em algumas doenças hepáticas crônicas, especialmente aquelas induzidas pelo vírus da hepatite C. Esses marcadores séricos têm sido usados em combinação, como índice para melhorar sua acurácia, como no caso do Fibro Test27, APRI35, índice de Forns13, Hepascore1 e FIB-432. Para o caso específico de esquistossomose o índice de Coutinho é utilizado, este consegui distinguir pacientes com esquistossomose sem fibrose avançada daqueles com fibrose avançada (BARRETO, 2017).

2.4 METABONÔMICA

A metabonômica pode ser definida como um conjunto de ferramentas, analíticas e de estatística multivariada, utilizadas para identificar mudanças de concentração dos metabólitos em um determinado fluido biológico e associar essas mudanças à perturbação sofrida pelo organismo (COSTA, DA, 2016). O termo tem origem grega “meta” significando mudança e “nomos” regras ou leis e foi criado por Jeremy Nicholson, Elaine Holmes e John Lindon, do *Imperial College of London*, em 1999, para descrever a geração de modelos quimiométricos baseado em classificação de mudanças metabólicas (GODOY, 2015).

A ideia de que vários estados de uma doença podem ser refletidos por mudanças nas concentrações de metabólitos é fundamental para a metabonômica. Desse modo, a sua aplicação

potencial é o diagnóstico aliado à descoberta de biomarcador(es) e compressão biológica (MADSEN; LUNDSTEDT; TRYGG, 2010).

A metabonômica tem sido utilizada com sucesso nos estudos de várias doenças infecciosas como em animais infectados por *Schistosoma Mansoni* (WANG et al., 2004), coinfecção por esquistossomose e hepatite viral B ou C (GOUVEIA, 2017) e complicações hepáticas virais como a fibrose hepática (BATISTA, A. *et al.*, 2018).

Por sua vez, a metabonômica, preocupa-se com a análise não direcionada de biofluidos para obter informações quantitativas ou semiquantitativas sobre o maior número possível de metabólitos. Os biofluidos mais comumente analisados são plasma, soro e urina, embora na literatura também haja relatos de análises de líquido cefalorraquidiano, suor, saliva e outros (ÅBERG; ALM; TORGRIK, 2009). Especificamente, envolve o uso de técnicas espectrométricas com ferramentas estatísticas e matemáticas para elucidar padrões dominantes e tendências diretamente correlacionadas com flutuações metabólicas em conjuntos de dados espectrais (DAHAB; SMITH, 2012).

2.5 PREPARO DE AMOSTRA

O preparo da amostra é considerado a etapa mais crítica e demorada em um fluxo de trabalho de análise química. Além disso, nesta etapa, é necessário ter muito cuidado para garantir o correto isolamento dos analitos de interesse em uma matriz complexa, para desenvolver métodos confiáveis e, consequentemente, obter maior sensibilidade, precisão, exatidão e recuperação nos resultados (SOARES DA SILVA BURATO *et al.*, 2020).

Na metabonômica, o preparo de amostra é uma parte fundamental porque afeta tanto o conteúdo do metabolito quanto a interpretação biológica dos dados. Um método de preparo de amostra ideal para análise metabonômica de amostras biológicas por cromatografia líquida deve ser não seletivo, simples e rápido com um número mínimo de etapas e reproduzível (VUCKOVIC, 2012).

Biofluidos, como soro e plasma, representam um meio ideal para o diagnóstico de doenças devido à facilidade de coleta, que pode ser realizada em todo o mundo, e ao seu envolvimento fundamental na função humana (LOVERGNE *et al.*, 2015). No entanto, trabalhar com soro representa um desafio, por um lado, as amostras são bem diluídas, com quantidade baixa de metabólitos. Na cromatografia é possível identificar marcadores biológicos em concentração menor que no RMN. Portanto, sendo esta técnica mais sensível, pode render resultados mais interessantes ao permitir enxergar metabólitos em maior proporção (CANUTO *et al.*, 2018). E, por outro lado, a quantidade de amostra biológica disponível normalmente é

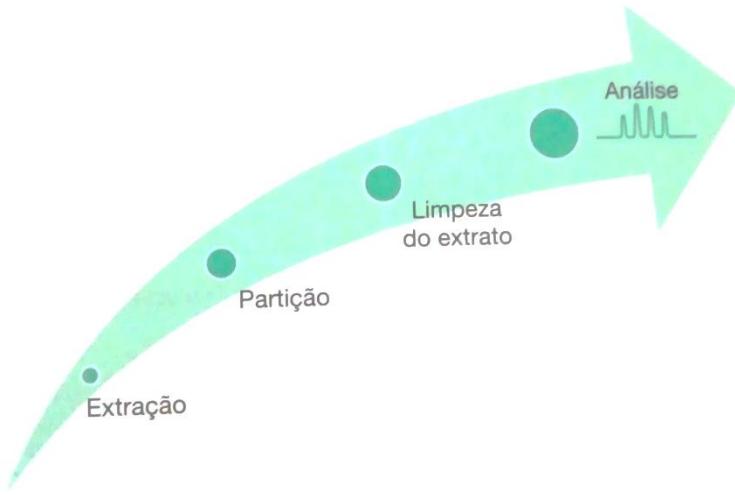
bem reduzida. O uso de técnicas miniaturizadas oferece uma solução, pois minimizam tanto o volume da amostra quanto o consumo de solventes orgânicos, além de reduzir estágios e tempo de análise resultando em métodos mais simples e rápidos (SOARES DA SILVA BURATO *et al.*, 2020). Neste cenário, 2 métodos de preparo de amostra miniaturizado foram avaliados:

1- **QuEChERS** (Quick, Easy, Cheap, Effective, Rugged, Safe), método de preparo de amostra, que tem como vantagem ser rápido, fácil, econômico, efetivo, robusto e seguro, explorando as possibilidades oferecidas pela instrumentação analítica moderna. Foi proposto por Anastassiades et al em 2003 para extração de resíduos de pesticidas em frutas e vegetais, desde então o método vem sofrendo modificações e além deste campo de aplicação, tem sido usado na extração de outros tipos de analitos e em diferentes matrizes, como antibióticos em amostras de mel, multirresíduos de fármacos em amostras de sangue, antibióticos β -lactâmicos em rim de bovinos, entre outras (ANASTASSIADES *et al.*, 2003; PRESTES, Osmar D. *et al.*, 2009).

Esta técnica de preparo de amostra está baseada em 3 etapas principais: (1) extração com acetonitrila, a qual possibilita a extração de uma quantidade menor de interferentes e uma ampla faixa de analitos com diferentes polaridades. (2) partição promovida pela adição de sais, por exemplo sulfato de magnésio ($MgSO_4$) e cloreto de sódio ($NaCl$) e (3) limpeza do extrato empregando a técnica Extração em Fase Solida Dispersiva (*dispersive solid phase extraction*, d-SPE) (ZANELLA; PRESTES, Osmar D.; *et al.*, 2015).

Considerando as vantagens potenciais deste método, como redução dos efeitos da matriz, simplicidade operacional, alta eficiência de extração para muitos analitos diferentes, necessidade de baixos volumes de solvente e baixo custo; a abordagem de extração QuEChERS tem sido avaliada em análises toxicológicas no contexto forense e clínico e aplicada a espécimes biológicos humanos, como urina, soro, sangue e fígado (CAMPÊLO *et al.*, 2021; SILVA, C. P. DA *et al.*, 2021).

Figura 4 – Principais etapas do método QuEChERS.

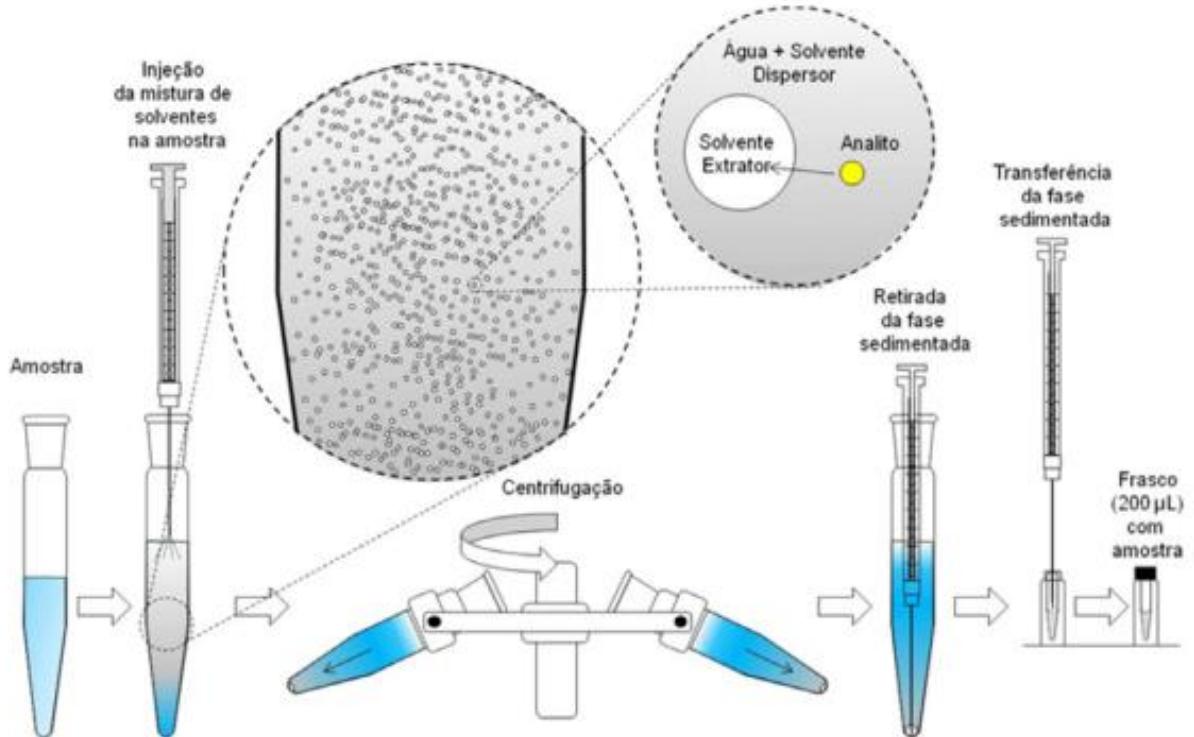


Fonte: (ZANELLA *et al*, 2015)

2- **DLLME** (Dispersive Liquid–Liquid Micro Extraction), em 2006, Assadi *et al* desenvolveram a microextração líquido-líquido dispersiva (DLLME), um método muito simples e rápido para extração e pré-concentração de compostos orgânicos. Este método é baseado em um sistema de solvente de componente ternário no qual os solventes de extração e dispersão são rapidamente injetados na amostra aquosa por seringa. A mistura é então agitada suavemente e uma solução turva (água / solvente dispersor / solvente de extração) é formada. Devido à grande área de superfície de contato das duas fases imiscíveis no DLLME, uma alta eficiência de extração é alcançada em um tempo relativamente curto (MONZÓN *et al.*, 2016; REZAEI *et al.*, 2006).

A DLLME é uma alternativa interessante para o preparo de amostra visando a determinação de compostos orgânicos em diferentes matrizes. O método utiliza a partição dos analitos de interesse empregando pequenos volumes de uma mistura formada por solvente dispersor e solvente extrator para extraer e concentrar os analitos. O solvente dispersor deve ser miscível na amostra (fase aquosa) e no solvente extrator (fase orgânica). Assim irá a promover a concentração dos analitos no solvente extrator, que geralmente com o uso da centrifuga, será separado e recolhido para ser analisado (ZANELLA; ADAIME; *et al.*, 2015).

Figura 5 – Etapas envolvidas no método DLLME.

Fonte: (CALDAS *et al.*, 2011)

A principal vantagem da técnica DLLME em relação à extração líquido / líquido usual ou extração em fase sólida (SPE), geralmente usada para a purificação da amostra, é o uso de quantidades muito baixas de solventes orgânicos, o que torna a técnica muito barata e sem a necessidade de cartuchos comerciais (MERCIECA *et al.*, 2018).

2.6 CROMATOGRAFIA A LÍQUIDO

A cromatografia é um método físico-químico de separação dos componentes de uma mistura, realizada através da distribuição desses componentes em 2 fases, que estão em contato íntimo. Uma das fases permanece estacionária, enquanto a outra se move através dela. Durante a passagem da fase móvel sobre a fase estacionária, os componentes da mistura são distribuídos pelas 2 fases de tal forma que cada uma delas é seletivamente retido pela fase estacionária, o que resulta em migrações diferenciais desses componentes (COLLINS; BRAGA; BONATO, 2006). A tabela 1 mostra as classificações da cromatografia, segundo as formas físicas mais encontradas das fases moveis e estacionarias.

Tabela 1 – Classificação da Cromatografia pelas formas físicas das fases moveis e estacionarias.

Critério de Classificação	Cromatografia									
	Planar			Em Coluna						
Técnica	Líquido			Gás			Fluido supercrítico		Líquido	
Fase estacionária	Líquido	Sólido	Fase ligada	Líquido	Sólido	Fase ligada	Sólido	Fase ligada	Líquido	Sólido
Tipo de cromatografia	CP	CCD	CCD	CGL	CGS	CGFL	CSS	CSFL	CLL	CLS
						CGQ				CLFL
										CLQ
										CTI
										CB

Fonte: Adaptado de (COLLINS; BRAGA; BONATO, 2006).

As técnicas cromatográficas estão entre as principais técnicas de separação, especialmente na análise de substâncias presentes em matrizes complexas, tais como fluidos biológicos, produtos naturais, sedimentos de rio e outras. Isto se deve, principalmente, à sua capacidade de separação dos componentes presentes nas misturas em função da eficiência e do poder de resolução das colunas modernas (LANÇAS, 2009).

As fases estacionarias quimicamente ligadas (CLFL, na tabela 1) são as formas mais representativas da cromatografia líquida em coluna. Essas fases são preparadas reagindo-se alguns grupos hidroxílicos que se encontram na superfície do sólido, normalmente a sílica, com grupos alquilas ou alquilas substituídos. Na maioria dessas fases estacionarias, os grupos disponíveis para interação com os componentes da amostra são cadeias longas de alquila, que funcionam como líquido apolar, e regiões da superfície como pontos ativos interfaciais. Desta maneira, o mecanismo de separação nessas fases é uma mistura de partição e adsorção (COLLINS; BRAGA; BONATO, 2006).

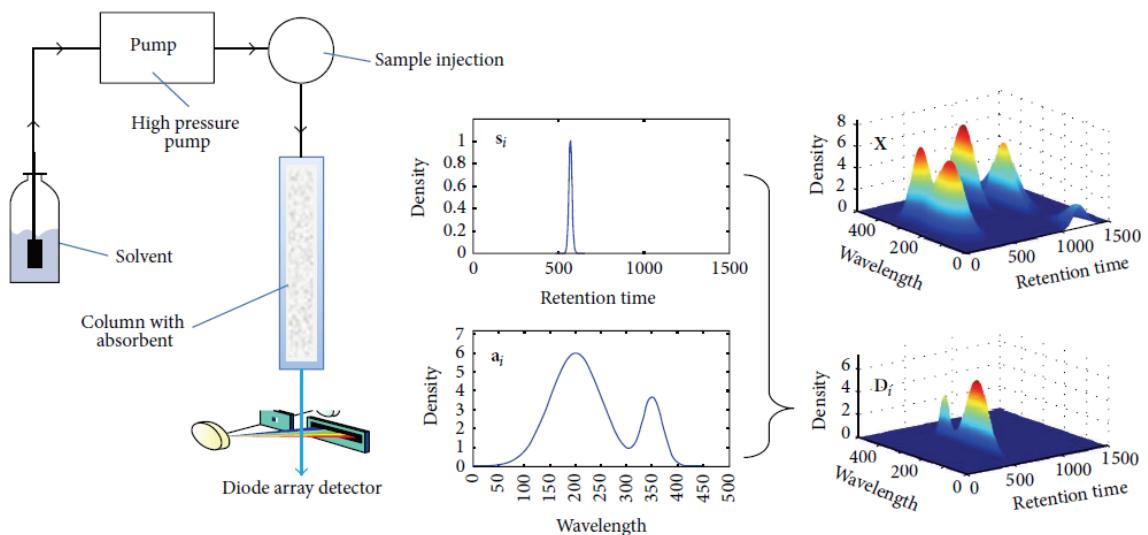
A cromatografia à líquido de alta eficiência (CLAE ou HPLC, do inglês *High Performance Liquid Chromatography*) além de ser uma das principais técnicas utilizadas na análise de compostos não voláteis e/ou termicamente instáveis, tornou-se uma ferramenta indispensável para a determinação de pequenas moléculas em uma ampla gama de aplicações. No campo da medicina, a metabonômica encontra aplicação na busca de biomarcadores precoces de doenças (marcadores diagnósticos) (GIKA *et al.*, 2014; LANÇAS, 2009). Pesquisas atuais estudaram o perfil metabólico em ratos infectados por *Schistosoma Japonicum* usando cromatografia líquida (HUANG *et al.*, 2020) e ratos infectados por *Schistosoma Mansoni* usando cromatografia gasosa (LOYO *et al.*, 2021).

Os detectores de cromatografia líquida mais comuns usados para pesquisas em metabonômica são a espectrometria de massa (MS) e o detector de arranjo de diodos (DAD). O MS pode fornecer grande informação fragmentada sobre metabólitos, cujas estruturas requerem confirmação de compostos padrão. Devido à limitação da base de dados padrão para HPLC-

MS, a identificação de metabolitos é muitas vezes difícil, especialmente para os desconhecidos. O DAD permite o monitoramento de diferentes comprimentos de onda de absorção simultaneamente em um único experimento. Comparativamente, o HPLC-DAD é fácil de obter na maioria dos laboratórios e apresenta menor custo associado, o que o torna adequado para analisar metabólitos globais e buscar biomarcadores de doenças (LIU, Y. *et al.*, 2011). Desta maneira, o HPLC-DAD apresenta-se como uma ferramenta acessível para hospitais e clínicas.

A tecnologia de HPLC-DAD gera um conjunto de dados que contém os picos cromatográficos e espectros de todos os compostos, tal como se apresenta na figura 6 a continuação:

Figura 6 - O princípio do conjunto de dados HPLC-DAD



Fonte: (CUI et al, 2014)

Quando a amostra é injetada, a bomba de alta pressão aciona o solvente para transportá-la pela coluna. Diferentes compostos receberão diferentes resistências quando passarem pela coluna. Dado um detector ultravioleta na parte inferior da coluna, um pico de cromatograma representado por s_i será observado quando um composto sair. A posição e a área do pico podem indicar o nome e a quantidade do composto. Se o detector for um DAD, que possui mais de mil canais para detectar o comprimento de várias ondas simultaneamente, o espectro para o mesmo composto representado por a_i também será gravado (CUI et al., 2014).

2.7 QUIMIOMETRIA

No sentido estrito da palavra, a quimiometria começou formalmente na primeira metade da década de 70, mas só se firmou definitivamente quando o computador se popularizou e se tornou mais presente no laboratório químico; os trabalhos em quimiometria no Brasil podem ser agrupados em três áreas principais: planejamento e otimização de experimentos, reconhecimento de padrões (métodos de análise exploratória e classificação) e calibração multivariada (BARROS NETO; SCARMINIO; BRUNS, 2006).

Em 1974, a quimiometria foi definida por Svante Wold como "*A arte de extrair informações quimicamente relevantes a partir de dados produzidos em experimentos químicos*" (WOLD, 1995). O conteúdo de informações geradas, neste caso pelo HPLC-DAD, é incrivelmente alto e as análises globais apresenta um desafio. Desta forma é necessária a utilização de métodos de estatística multivariada, que sejam capazes de extrair informações dos dados permitindo que os padrões distintos sejam mais facilmente identificados. Esses métodos podem ser classificados como supervisionados e não supervisionados (WORLEY; POWERS, 2013), todos dentre a área de reconhecimento de padrões na quimiometria.

2.7.1 Métodos Não Supervisionados

Os métodos não supervisionados como Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*), não requerem nenhum conhecimento prévio a respeito da classificação das amostras, elas serão agrupadas naturalmente com base na informação contida nos dados experimentais em questão. Foi inicialmente desenvolvida por Karl Pearson no início do século 20, e posteriormente estudada por Harold Hotelling por volta de 1930. A PCA é um método usado para projetar os dados multidimensionais em um espaço de dimensão menor, seu objetivo é reduzir a dimensionalidade dos dados conservando o máximo de informações em poucas dimensões, gerando um novo conjunto de variáveis chamado componentes principais (PC). Cada PC resultante da análise é independente entre si e constituirão uma combinação linear das variáveis originais e são criados de acordo com sua importância em relação à variância total obtida da amostra (ADANIYA HIGA, 2019; FERREIRA, Márcia Miguel Castro, 2015).

Como uma ferramenta de análise exploratória a PCA permite revelar a existência ou não de amostras anômalas, de relações entre as variáveis medidas e de relações ou agrupamentos entre amostras (LYRA *et al.*, 2010). Matematicamente a matriz de dados experimentais, denominada como matriz X, é descomposta em um produto de duas outras matrizes, uma

denominada *scores* ou escores (**T**) e outra denominada *loadings* ou pesos (**P**), mais uma matriz de erro (**E**), conforme equação 1 (VALDERRAMA, L. *et al.*, 2016).

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \quad (1)$$

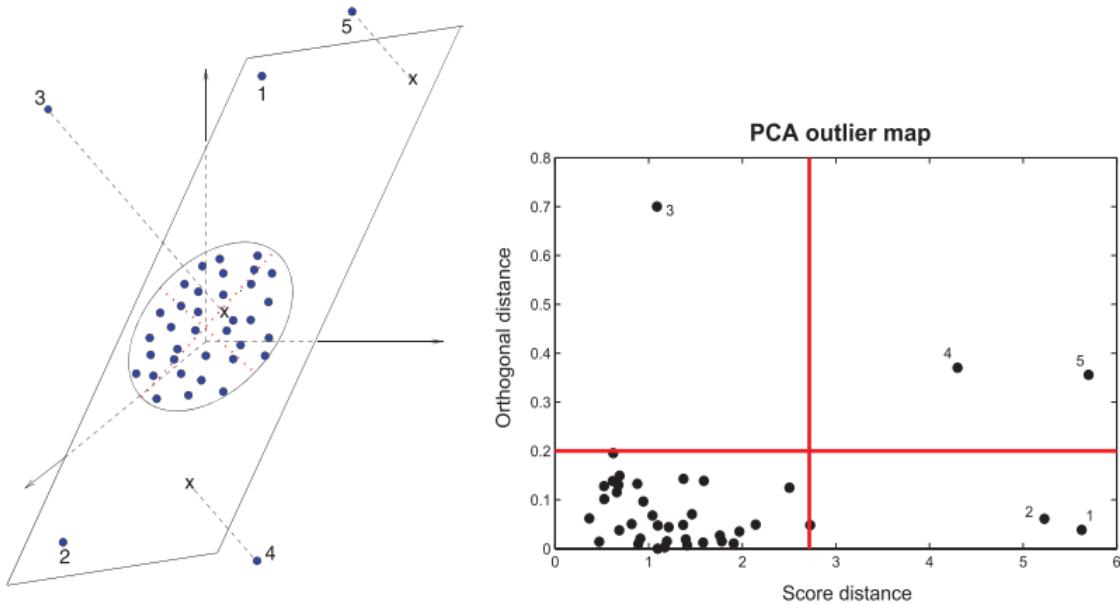
Os escores representam as coordenadas das amostras no sistema de eixos formados pelos componentes principais. Cada componente principal é constituído pela combinação linear das variáveis originais e os coeficientes da combinação são denominados pesos. Levando em conta variáveis discretas podemos dizer que, matematicamente, os pesos são os cossenos dos ângulos entre as variáveis originais e PC, representando, portanto, o quanto cada variável original contribui para uma determinada PC. A primeira componente principal (PC1) é traçada no sentido da maior variação no conjunto de dados; a segunda (PC2) é traçada ortogonalmente à primeira, com o intuito de descrever a maior porcentagem da variação não explicada pela PC1 e assim por diante; enquanto os escores representam as relações de similaridade entre as amostras. A avaliação dos pesos permite entender quais variáveis mais contribuem para os agrupamentos observados no gráfico dos escores. Através da análise conjunta do gráfico de escores e pesos, é possível verificar quais variáveis são responsáveis pelas diferenças observadas entre as amostras. O número de componentes principais a ser utilizado no modelo PCA é determinado pela porcentagem de variância explicada. Assim, seleciona-se um número de componentes de tal maneira que a maior porcentagem da variação presente no conjunto de dados originais seja capturada (SOUZA, A. M. DE; POPPI, 2012).

A PCA é muito sensível a observações anômalas. Consequentemente, os primeiros componentes são frequentemente atraídos para pontos periféricos e podem não capturar a variação das observações regulares. Portanto, a redução de dados com base no PCA clássico (CPCA) torna-se não confiável se amostras anômalas ou outliers estiverem presentes nos dados (HUBERT; ROUSSEEUW; BRANDEN, VANDEN, 2005). Isso é especialmente notório para casos onde a proporção de amostras anômalas no conjunto de dados é elevada. Então em casos onde se tem poucas amostras, se uma ou poucas amostras forem anômalas a probabilidade da PCA ser afetada por estas amostras é elevada, foi com o objetivo de reduzir a influência das amostras anômalas nos resultados de PCA que foi desenvolvido a PCA com critério de análise robusto.

A Análise Robusta da Componente Principal (*Robust PCA* ou ROBPCA) tem como objetivo obter PCs que não são muito influenciadas por amostras anômalas. Existem diferentes

tipos de amostras anômalas que podem influenciar na rotação do modelo. Para diagnosticá-las, podemos desenhar um mapa de outliers para PCA. Um exemplo estilizado de tal mapa é mostrado no lado direito da Figura 7, que corresponde aos dados tridimensionais no lado esquerdo da mesma figura e que é ajustado por dois componentes principais. No eixo vertical do mapa de outlier da PCA, encontramos a distância ortogonal de cada ponto. Esta é apenas a distância euclidiana do ponto de dados para sua projeção. A distância ortogonal é mais alta para os pontos 3, 4 e 5 no exemplo. No eixo horizontal, vemos a distância de escore de cada ponto de dados, que é apenas a distância robusta de sua projeção em relação a todos os pontos de dados projetados. A distância de escore é bastante alta para os pontos 1, 2, 4 e 5 na figura.

Figura 7 – Ilustração da PCA: (esquerda) tipos de outlier; (direita) mapa de outlier: gráfico de distâncias ortogonais versus distâncias de escores.



Fonte: (ROUSSEEUW; HUBERT, 2018)

Ao combinar as duas medidas de distância, o mapa de outliers permite distinguir entre quatro tipos de pontos de dados: (1) O grupo dos pontos que têm uma pequena distância ortogonal e uma pequena distância de escore que são as observações regulares. (2) Pontos com uma distância de escore alta, mas uma pequena distância ortogonal, como os pontos 1 e 2 na Figura 6, são chamados de pontos de boa alavancagem, pois podem melhorar a precisão do subespaço PCA ajustado. (3) Outliers ortogonais têm uma grande distância ortogonal, mas uma pequena distância de escore, como o ponto 3. Pode indicar que existe informação ortogonal a ser explicada com mais uma PC. Caso não seja o caso é uma amostra com indícios de outlier (4) Pontos de alavancagem ruins têm uma grande distância ortogonal e uma grande distância

de escore, como os pontos 4 e 5. Eles estão longe do espaço abrangido pelas componentes principais robustas, e após a projeção naquele espaço, ficam longe da maioria dos outros dados projetados. Eles são chamados 'Ruim' porque normalmente têm uma grande influência na PCA clássica, já que os principais autovetores estarão inclinados em direção a eles (HAMZENEJAD; GHOUSHCHI; BARADARAN, 2021; ROUSSEEUW; HUBERT, 2018). Os limites traçados como valores de distância ortogonal e de escore críticos são calculados com base nos resultados de projeção das amostras, então amostras que estão próximas a estes limites devem ser analisadas com mais cuidado.

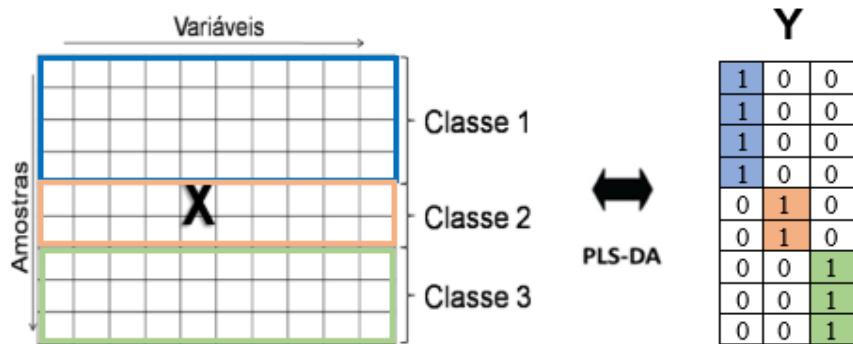
2.7.2 Métodos Supervisionados

A análise exploratória com PCA possibilita investigar tendências e quais variáveis tem mais importância para ressaltar semelhanças e diferenças entre as amostras, porém não pode ser usada para predição de alguma característica desejada, neste caso é necessário usar métodos supervisionados de análise. Nos métodos supervisionados de análise para reconhecimento de padrão, amostras com características conhecidas são usadas para construir modelos de classificação, em seguida, o modelo é utilizado para prever a classe de amostras desconhecidas. Podem ser destacados modelos discriminantes como o PLS-DA (do inglês, *Partial Least Squares - Discriminant Analysis*) e a LDA (do inglês, *Linear Discriminant Analysis*) e os de modelagem de classe como o DD-SIMCA (do inglês, *Data Driven Soft Independent Modeling of Class Analogy*).

O PLS-DA, utiliza a técnica de regressão multivariada por mínimos quadrados parciais (PLS). O PLS é um método de calibração inversa, no qual se busca uma relação direta entre a resposta instrumental (matriz X) e a propriedade de interesse (matriz Y) onde cada coluna desta matriz corresponde a uma classe. O procedimento utilizado para a construção do modelo de classificação é o mesmo utilizado pelo PLS, no entanto, a propriedade de interesse em modelos de classificação é uma variável categórica que descreve a atribuição de classe da amostra. Valores de 1 refere-se as amostras que pertencem aquela classe e valores de 0 refere-se a amostras que não pertencente à classe (LOPES, R. E. C., 2015; SANTANA *et al.*, 2020). A Figura 8 ilustra a organização dos dados utilizados para a construção do modelo de classificação PLS-DA. A matriz X é composta, neste caso, pelos cromatogramas das amostras e a matriz Y é construída com informação binária (1 ou 0) para indicar se a amostra pertence ou não à classe.

Como o método PLS maximiza a relação entre a variável dependente e os escores, pode-se afirmar que as variáveis latentes representam as direções que melhor discriminam as classes e que definem a máxima separação entre elas (FERREIRA, Márcia Miguel Castro, 2015).

Figura 8 – Esquema da organização dos dados para a construção do modelo de classificação usando PLS-DA



Fonte: Adaptado de (LOPES, R. E. C, 2015).

Para uma melhor interpretação dos dados, é possível utilizar um filtro de correção de sinal ortogonal no modelo PLS-DA, permitindo assim, que OPLS-DA separe a variância preditiva da não preditiva (ortogonal). No caso de sobreposição de classes numa direção preditiva, a variação ortogonal pode ser empregada para intensificar as taxas de classificação o que pode levar a uma maior facilidade na visualização dos *loadings* usados na discriminação (SANTOS, J. C., 2018; SOUZA ALMEIDA, 2021).

A LDA, é uma das técnicas usadas para redução de dimensionalidade, aprendizado de máquina e reconhecimento de padrões. Foi proposto por Ronald A. Fisher, em 1936. Através de combinação linear a LDA separa, em duas ou mais classes de interesse, os dados originais. Essa separação em classes ou grupos distintos é feita por meio de uma transformação linear que maximiza a distância entre as classes e minimiza a dispersão dentro de cada classe. No entanto, se o número de amostras for menor ao número de variáveis (situação habitual em dados espectrais), a LDA não pode ser usada devido à necessidade de se inverter a matriz ao calcular os escores discriminantes de Fisher (LI, Y. *et al.*, 2021). Para solucionar este problema, algoritmos de seleção de variáveis podem ser utilizados acoplados à LDA, como por exemplo algoritmo genético (GA-LDA) (COSTA FILHO, DA; POPPI, 1999) e o algoritmo de projeções sucessivas (SPA-LDA) (ARAÚJO, Mário César Ugulino *et al.*, 2001).

Com base na teoria evolutiva das espécies, John H. Holland um pesquisador da Universidade de Michigan, propôs nos anos 60 a construção de um algoritmo matemático para otimização em sistemas complexos, sendo denominado de Algoritmo Genético (GA do inglês, *Genetic Algorithm*). O GA é uma ferramenta de otimização baseado nos processos biológicos da evolução das espécies, simulando matematicamente a teoria de Darwin (COSTA FILHO, DA; POPPI, 1999); e é usada para seleção de variáveis representativas na construção de modelos tanto em calibração como em classificação.

Para aprimorar os resultados do algoritmo genético, em razão da natureza estocástica, cálculos sucessivos podem ser usados para avaliar o conjunto de variáveis que mais se aproximam de um mínimo global. Uma das formas é fazendo uma simulação Monte Carlos que se refere a um tipo de simulação que se baseia em amostragem aleatória repetida e análise estatística para calcular os resultados. Este método de simulação está intimamente relacionado a experimentos aleatórios, experimentos para os quais o resultado específico não é conhecido de antemão (RAYCHAUDHURI, 2008).

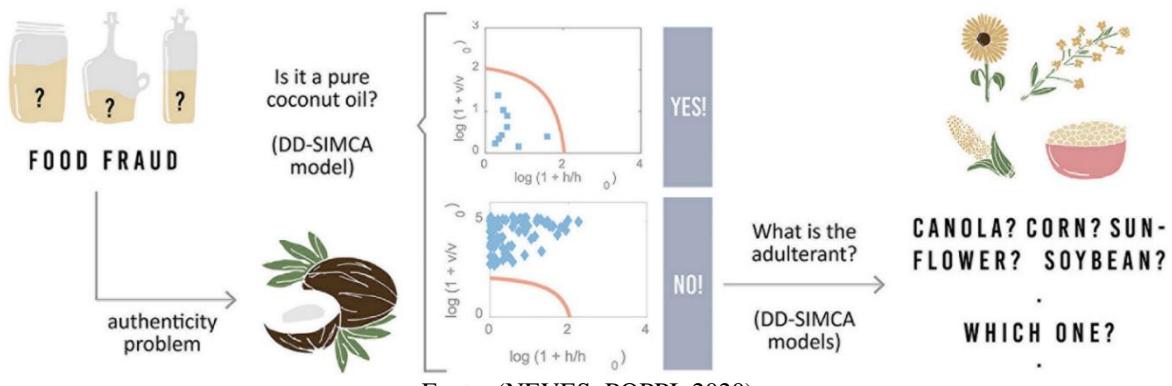
O algoritmo de projeções sucessivas (SPA do inglês, *Successive Projection Algorithm*) foi proposto em 2001 por Araújo *et al*, como um algoritmo de seleção direta para seleção de variáveis em calibração multivariada. O SPA emprega operações simples em um espaço vetorial para obter subconjuntos de variáveis com pequena colinearidade. A diferença do GA, o SPA é de natureza determinístico e exige uma carga de trabalho computacional menor, principalmente quando o número total de variáveis é grande (ARAÚJO, Mário César Ugulino *et al.*, 2001; SOARES *et al.*, 2013).

Os métodos descritos até agora levam em consideração todas as informações das classes na construção do modelo. No entanto, é importante citar também os métodos quimiométricos baseados em uma modelagem de classe ou classificador de uma classe. Classificador de uma classe significa que o modelo é construído para uma classe particular de interesse, usando apenas amostras da classe alvo. Quando testar uma nova amostra, neste tipo de modelos, o resultado via de regra será a indicação de que a nova amostra pertence ou não à classe alvo. O método de modelagem de classe mais usado é o SIMCA (do inglês, *Soft independent modeling of class analogy*). O DD-SIMCA (do inglês, *data driven SIMCA*) é uma extensão ou modificação do SIMCA que usa as informações dos dados para atribuir a amostra a classe alvo com base em cálculos de distância, proposto em 2017 por Zontov et al (FIDELIS *et al.*, 2017; SANTOS, L. O. DOS *et al.*, 2022; ZONTOV *et al.*, 2017).

O DD-SIMCA tem a capacidade de caracterizar os resultados da classificação de uma forma estatisticamente sólida, ou seja, calcular teoricamente os erros de classificação incorreta (MAZIVILA *et al.*, 2020). O método consiste em duas etapas, na primeira, a PCA é aplicada aos dados de treinamento da classe alvo. Na segunda etapa, os resultados da PCA são usados para o cálculo de duas distâncias para cada objeto, distancia de escores (SD) e distancia ortogonal (OD). SD é a distância de Mahalanobis ao quadrado de um objeto até o centro do plano dos escores. Por outro lado, OD é a distância euclidiana quadrada de um objeto ao espaço da PCA. Os resultados da classificação são apresentados graficamente por meio do gráfico de OD versus SD (chamado gráfico de aceitação), que inclui os limites da classe especificados por

um determinado valor alfa. Cada objeto é caracterizado por sua posição neste gráfico de aceitação. Aqueles que estão fora da área de aceitação, não são semelhantes o suficiente a outros membros da classe e são considerados incompatíveis com a classe (ADENAN *et al.*, 2020; MAZIVILA *et al.*, 2020; RODIONOVA, O. Ye *et al.*, 2014; SANTOS, L. O. DOS *et al.*, 2022). Um exemplo de aplicação do DD-SIMCA é ilustrado na figura 9, que descreve o estudo de Neves e Poppi em 2020 sobre a autenticação e identificação de adulterantes em óleo de coco usando FTIR em conjunto com a modelagem de uma classe DD-SIMCA.

Figura 9 – Esquema de aplicação do DD-SIMCA na autenticação e identificação de adulterantes em óleo de coco



Fonte: (NEVES; POPPI, 2020)

2.7.3 Figuras de Mérito

Uma vez construídos os modelos de classificação, é preciso que sejam validados. Algumas métricas são usadas para avaliar a qualidade de um modelo, todas derivam da matriz de contingência, na qual os erros cometidos pelo classificador são organizados na forma de uma tabela de contingência. Nessa tabela, a classe verdadeira (Colunas) é representada *versus* a classe estimada (linhas) pelo classificador ou modelo construído (FERREIRA, Márcia Miguel Castro, 2015).

Na tabela 2 é apresentado um exemplo de matriz de contingência para duas classes onde a classe alvo é a “A”, 4 informações diferentes são geradas. A diagonal verde representa as previsões corretas, verdadeiro positivo (VP) e verdadeiro negativo (VN). Os VP são amostras que pertencendo a classe alvo e são corretamente classificadas naquela classe (amostras positivas). Os VN são amostras que não são da classe alvo e são classificadas corretamente como não pertencentes à classe alvo (amostras negativas). Já a diagonal laranja representa as previsões incorretas, falso negativo (FN) e falso positivo (FP). Os FN são amostras positivas identificadas como negativas, ou erro tipo II. Os FP são amostras negativas identificadas como positivas, ou erro tipo I (THARWAT, 2018).

Tabela 2 – Exemplo de matriz de contingência

		Classe calculada	
		A	B
Classe Verdadeira	A	VP	FN
	B	FP	VN

Fonte: A autora (2022)

Dentre as métricas mais utilizadas para avaliar a eficiência dos modelos de classificação, pode-se citar a sensibilidade, especificidade e acurácia. A sensibilidade é a fração das amostras positivas que foram classificadas corretamente, matematicamente é representada pela equação 2. A especificidade representa a proporção das amostras negativas que foram classificadas corretamente, pode ser representada matematicamente pela equação 3. Geralmente podemos considerar a sensibilidade e especificidade como dois tipos de exatidão, sendo a primeira para amostras positivas e a segunda para amostras negativas. A exatidão é definida como uma razão entre as amostras classificadas corretamente para o número total de amostras. O complemento desta métrica é a taxa de erro ou taxa de classificação incorreta. A exatidão e o erro estão apresentadas matematicamente nas equações 4 e 5 respectivamente (NEVES; POPPI, 2020; THARWAT, 2018).

$$Sensibilidade = \frac{VP}{VP + FN} \quad (2)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (3)$$

$$Exatidão = \frac{VP + VN}{VP + VN + FP + FN} \quad (4)$$

$$Erro = 1 - Exatidão = \frac{FP + FN}{VP + VN + FP + FN} \quad (5)$$

3 METODOLOGIA

3.1 POPULAÇÃO DE ESTUDO

A população do estudo está composta por pacientes residentes no estado de Pernambuco entre 18 e 80 anos, com esquistossomose e FPP evidente, recrutados durante exame de US no ambulatório de Gastroenterologia do Hospital das Clínicas da Universidade Federal de Pernambuco (HC/UFPE), no período de março de 2019 a novembro de 2020.

Os critérios de inclusão no estudo foram pacientes de ambos os sexos com diagnóstico de esquistossomose ou dados clínico-laboratoriais (coproparasitológico positivo, US evidenciando padrão de fibrose periportal) ou dados epidemiológicos (contato com rios e lagos em áreas endêmicas).

Os critérios de exclusão foram pacientes com diagnóstico clínico, laboratorial ou US compatível com doenças hepáticas de outras etiologias (Hepatite B ou C, cirrose hepática ou doença hepática gordurosa), coinfecção com HIV, consumo excessivo de álcool, uso de drogas hepatotóxicas, transplante de fígado e esplenectomia prévia.

No total foram coletadas 99 amostras, dessas 2 foram descartadas por estarem com doenças hepáticas de outras etiologias e 3 pacientes tiveram coleta de difícil acesso, o que comprometeu a qualidade das amostras de soro. As 94 amostras restantes foram catalogadas usando como padrão de referência, na avaliação da fibrose hepática, o US segundo a classificação de Niamey. Essa classificação é apresentada na tabela 3.

Tabela 3 – Distribuição das amostras segundo classificação de Niamey com US

Classe	Número de Amostras
C	19
D	30
E	34
F	11

Fonte: A autora (2022)

Cabe ressaltar que as amostras da classe D foram catalogadas pelo médico de forma distintas, como “D”, outras “DC+”, outras “D+C”, outras “DC”, uma “D+” e outra como “DC+/E-”. estas categorias não se encontram reportadas na literatura (a exceção de D+C) e para fins práticos deste trabalho foram consideradas todas como pertencendo à classe D.

Considerações Éticas

Previamente a manipulação das amostras, o trabalho foi aprovado pelo comitê de ética em pesquisa (CEP) do hospital das clinicas da UFPE/EBSERH, sob o número do parecer 4.844.474 (ANEXO A)

3.2 OTIMIZAÇÃO

Para determinar as melhores condições de trabalho, o preparo de amostra e análises cromatográfica foram otimizados. A otimização foi realizada com 2 amostras diferentes do grupo do estudo, uma amostra controle (C) e uma com hepatite (H). Posteriormente, as melhores condições de analise encontradas foram aplicadas a todas as amostras do estudo.

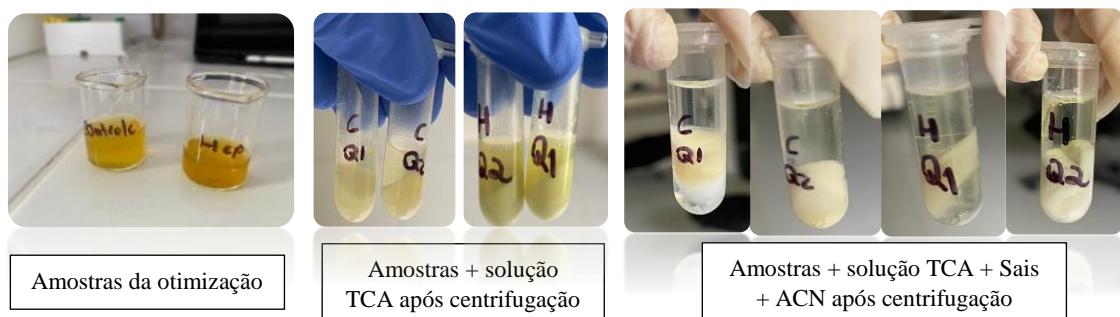
3.2.1 Preparo de amostra

QuEChERS

O método QuEChERS foi avaliado em duas variantes Q1 e Q2, alterando os sais. O Q1 foi feito usando sulfato de magnésio ($MgSO_4$) e acetato de sódio (CH_3COONa). Já o Q2 foi com sulfato de sódio (Na_2SO_4) e cloreto de sódio ($NaCl$).

À 800 μL da amostra de soro foi adicionada 80 μL de uma solução de ácido tricloroacético (TCA) 1:5. A mistura foi agitada vigorosamente por um minuto e levada à centrifuga (Hettich – Mikro 185) por 5 minutos a 3000g (7183 RPM). Seguidamente, foi adicionado ao sobrenadante 800 μL de acetonitrila e a mistura dos sais (Q1 ou Q2) até saturar. O conjunto foi agitado vigorosamente por um minuto e posteriormente centrifugado nas mesmas condições previamente descritas. Finalmente a fase orgânica foi coletada, filtrada em filtro seringa de 0,22 μm e armazenada em vial. Na figura 10 é apresentado a metodologia anteriormente descrita aplicada às amostras da otimização.

Figura 10 – Esquema do QuEChERS nas amostras da otimização.

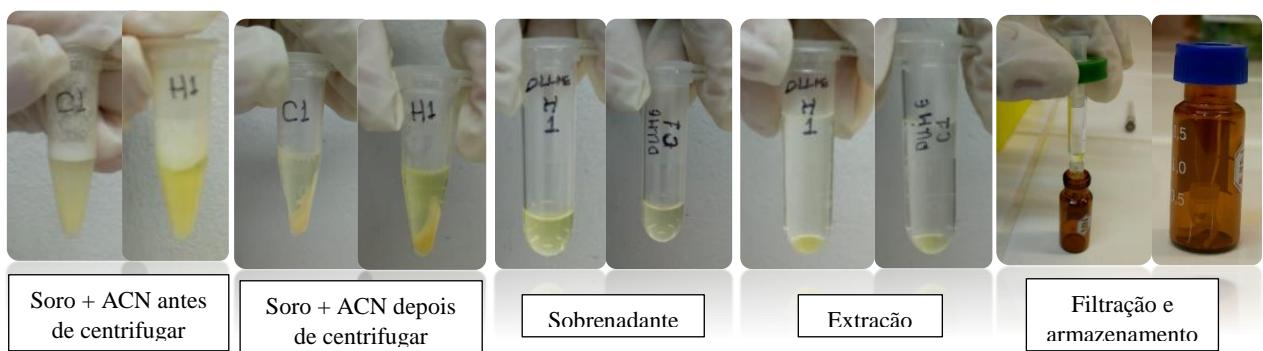


Fonte: A autora (2022)

DLLME

À 200 μ L da amostra de soro foi adicionado 200 μ L de acetonitrila, a mistura foi agitada e então levada ao centrifugador nas mesmas condições do método QuEChERS. Posteriormente o sobrenadante foi coletado e adicionado rapidamente 1,2 mL de uma solução 1:2 de diclorometano/acetona. A solução foi acrescentada usando uma seringa com agulha. Em seguida o sistema amostra + solução foi centrifugado e congelado a -40°C. Logo após a fase orgânica (diclorometano) foi coletada, filtrada em filtro seringa de 0,22 μ m e armazenada em vial com insert. A figura 11 apresenta o esquema deste método aplicado nas amostras da otimização.

Figura 11 – Esquema do DLLME nas amostras da otimização.

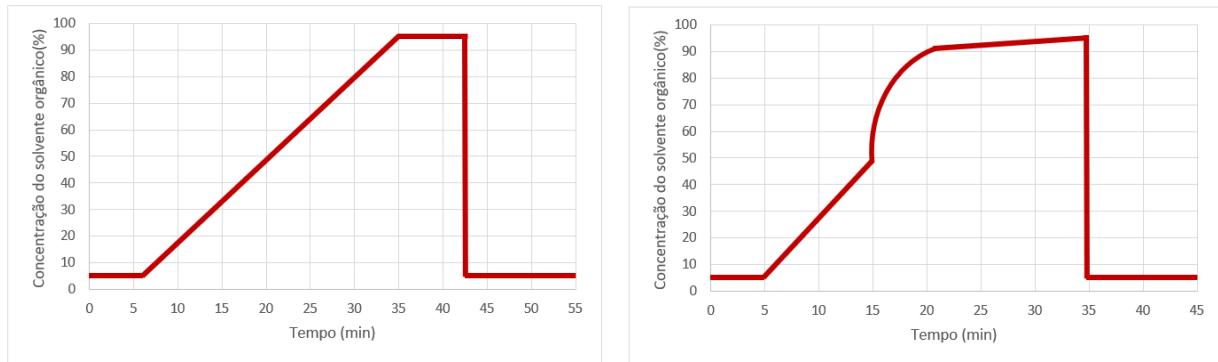


Fonte: A autora (2022)

3.2.2 Cromatografia

As análises cromatográficas da otimização foram realizadas no DQF/UFPE em um HPLC-DAD, da marca Shimadzu, modelo PROEMINENCE LC-20AD com detector DAD SPD-M20A. O tipo de coluna usada foi C18 (Phenomenex, Luna, 250 x 4,6 mm, 100 Å, 5 μ m). A fase móvel estava composta por uma mistura de água Milli-Q e um solvente orgânico (Metanol ou Acetonitrila). Foram avaliados 2 tipos de gradiente de concentração do solvente orgânico (ver figura 12) e 4 classes de fase móvel: ACN:água; MeOH:água; ACN: $3,5 \times 10^{-3}$ mol L $^{-1}$ ácido trifluoroacético (pH aproximadamente igual a 3) e MeOH: $3,5 \times 10^{-3}$ mol L $^{-1}$ ácido trifluoroacético (pH aproximadamente igual a 3). Em todos os casos a vazão foi de 1 mL min $^{-1}$ e temperatura do forno fixa em 30°C. Cada amostra foi injetada em duplicata com um branco injetado entre amostras diferentes e o volume de injeção foi de 20 μ L. O *software* usado pelo instrumento foi o LabSolution e os dados foram exportados em 3D (picos cromatográficos e espectros) no formato txt.

Figura 12 – Gradientes de concentração do solvente orgânico, linear (esquerda) e curvo (direita).



Fonte: A autora (2022)

Devido a problemas com a regulação da temperatura da central analítica no DQF e consequentemente dificuldades para manter a temperatura do forno fixa em 30°C, todas as análises cromatográficas das amostras do estudo foram realizadas no CENAPESQ da UFRPE em um HPLC-DAD modelo PROEMINENCE LC-20AT, da Shimadzu, com o mesmo tipo de detector DAD (SPD-M20A). O instrumento é apresentado na figura 13. A coluna e as condições usadas foram as mesmas otimizadas anteriormente.

Figura 13 – HPLC-DAD utilizado nas análises das amostras do estudo.



Fonte: A autora (2022)

3.3 MODELOS QUIMIOMÉTRICOS

Todos as análises de estatística multivariada foram realizadas no *software* MATLAB R2010a. Dado que a natureza dos dados está na ordem de um tensor (Amostras x Tempo de retenção x Comprimento de onda); para visualização, manipulação e análise dos dados e construção dos modelos metabonômicos, foram utilizadas 3 abordagens: (1) comprimento de onda fixo em 210 nm, (2) somatório de todos os comprimentos de onda (de 200 a 800 nm) e (3) somatório dos comprimentos de onda de 200 a 400 nm.

Os dados foram pré-processados, usando o algoritmo Icoshift para correção de deslocamento de pico e para correção da linha de base no sinal cromatográfico foi subtraído o branco em cada amostra. Após o pré-processamento, os dados foram submetidos a análise exploratório usando os algoritmos PCA e ROBPCA com o auxílio das interfaces gráficas *pca_toolbox_1.5* e LIBrary for Robust Analysis (LIBRA) respectivamente. Já para os métodos de classificação foram aplicados os métodos de PLS-DA, SPA-LDA, GA-LDA e DD-SIMCA. O modelo PLS-DA foi construído usando a interface gráfica *classification_toolbox_5.3*, o DD-SIMCA foi construído usando o *DD_SIMCA_MASTER (2019)* e os modelos LDA com seleção de variáveis foram construídos usando rotinas desenvolvidas pelo Laboratório de automação e instrumentação em Química Analítica e Quimiometria (LAQA) liderado pelo Prof. Dr. Mário Cesar Ugulino de Araújo. Para a seleção de variáveis pelo algoritmo genético, foi usada uma população inicial de 100 indivíduos, 50 gerações, probabilidade de mutação 10% e probabilidade de cruzamento 60%. Dado que a seleção das variáveis é aleatória, o cálculo foi otimizado usando o melhor resultado de 1000 cálculos usando os mesmos parâmetros.

Anteriormente a construção dos modelos classificatórios, as amostras de cada classe, foram pré-processadas e divididas em dois grupos: Treinamento e Teste na proporção 70:30. O algoritmo de Kennard-Stone (GALVÃO *et al.*, 2005) foi utilizado para a divisão e seleção das amostras. Os modelos foram avaliados quanto a sensibilidade, especificidade e exatidão, usando a matriz de contingência.

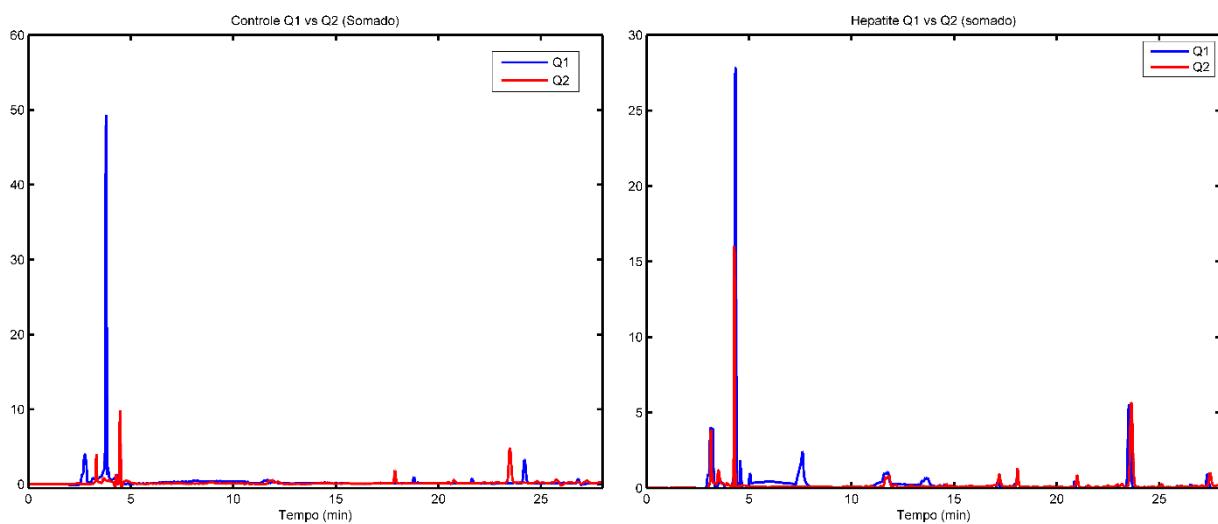
4 RESULTADOS

4.1 OTIMIZAÇÃO

A otimização foi realizada para avaliar o desempenho dos métodos de preparo de amostra e os parâmetros das análises cromatográficas para desta forma encontrar a metodologia mais eficiente para análises das amostras do estudo. Entende-se como metodologia mais eficiente aquela que fornece o maior número de informações, metabolitos ou picos cromatográficos, pois propiciará o desenvolvimento de um modelo mais específico.

No caso do preparo de amostra pelo método QuEChERS, foi avaliado em duas versões, testando 2 pares de sais na fase de partição ou também chamado *salting out*, $\text{MgSO}_4/\text{CH}_3\text{COONa}$ (Q1) e $\text{Na}_2\text{SO}_4/\text{NaCl}$ (Q2). Na figura 14 é apresentado os cromatogramas das amostras de controle e hepatite preparadas na versão Q1 e Q2. É possível observar que a opção de preparo Q1 forneceu o melhor desempenho, apresentando picos mais intensos e em maior quantidade, especialmente na amostra de hepatite.

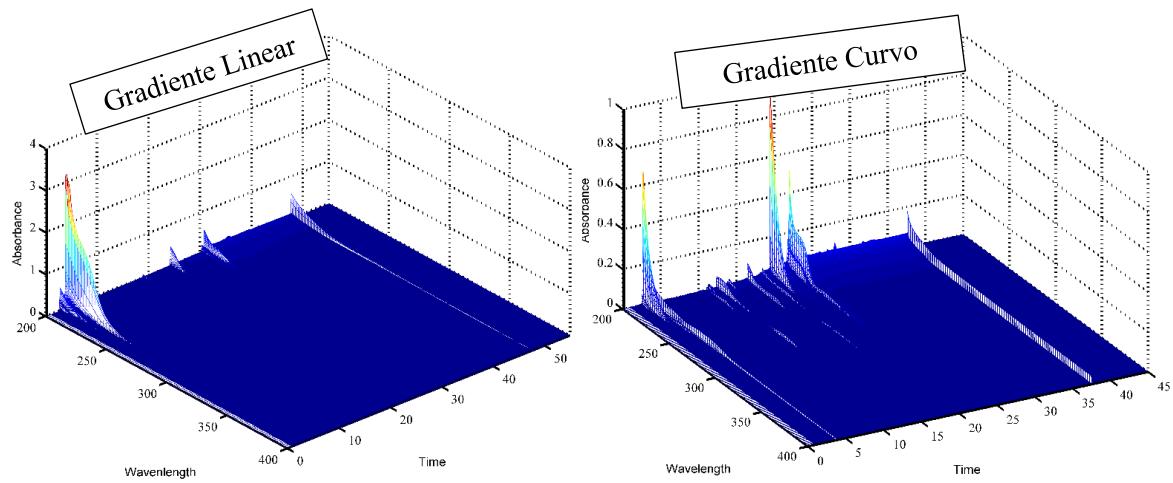
Figura 14 – Cromatograma, com os comprimentos de onda somados, das amostras controle e hepatite nas duas versões do método QuEChERS, Q1 vs Q2. Usando como fase móvel, gradiente linear de ACN.



Fonte: A autora (2022)

Por outro lado, também foi avaliado o gradiente de concentração do solvente orgânico da fase móvel, dois tipos foram testados, denominados neste trabalho como linear e curvo, segundo foi apresentado na figura 12 da secção 3.2.2 (página 42). O gradiente curvo mostrou ser mais eficiente pois forneceu maior número de picos e em menos tempo, quando comparado ao linear, tal e como se apresenta na figura 15.

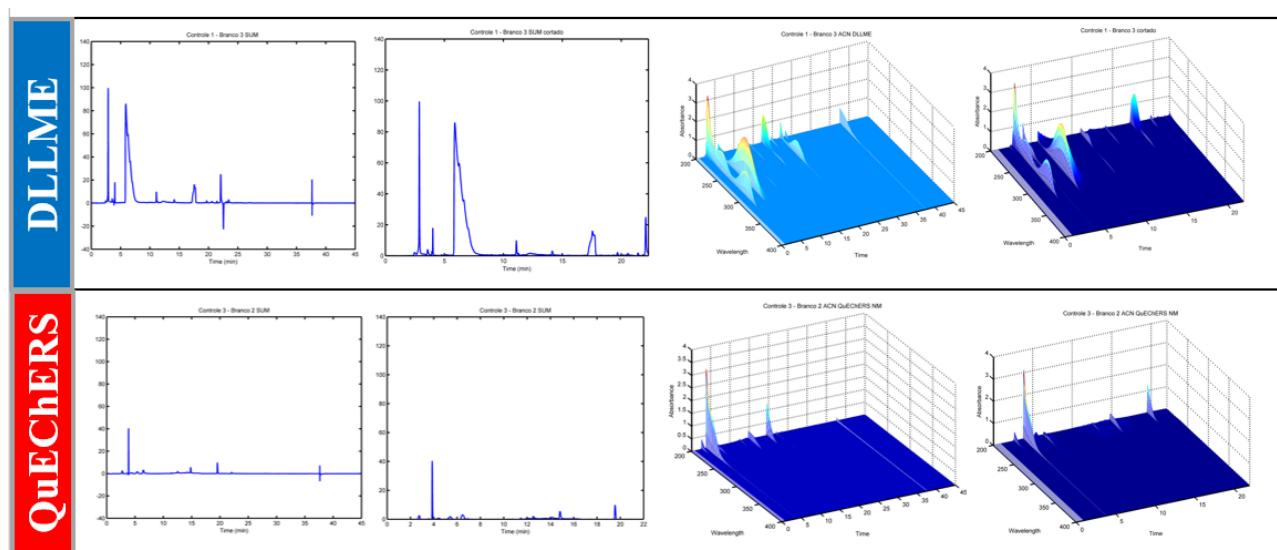
Figura 15 – Gráficos em 3D da amostra controle em dois tipos de gradientes de concentração da fase orgânica (ACN) na versão Q1 do método QuEChERS.



Fonte: A autora (2022)

Além do QuEChERS outro tipo de preparo de amostra foi testado, o DLLME. De acordo com os resultados apresentados na figura 16, o QuEChERS apresentou poucos picos, o que indica que este método não é adequado o suficiente para extrair informações bioquímicas presentes na amostra. Já o DLLME mostrou ser mais eficiente visto que, um maior número de picos foi observado, o que equivale a mais informações bioquímicas disponíveis para o estudo metabonômico. Extrair o máximo de informações (metabolitos) das amostras é crucial, já que estas serão usadas para a construção dos modelos metabonômicos.

Figura 16 – Comparaçao dos métodos DLLME vs QuEChERS na amostra controle após subtração do branco. Os cromatogramas em 2D são com os comprimentos de onda somados. Os cromatogramas em 2D e 3D se apresentam inteiros e cortados respectivamente. Fase móvel usada ACN no gradiente curvo.

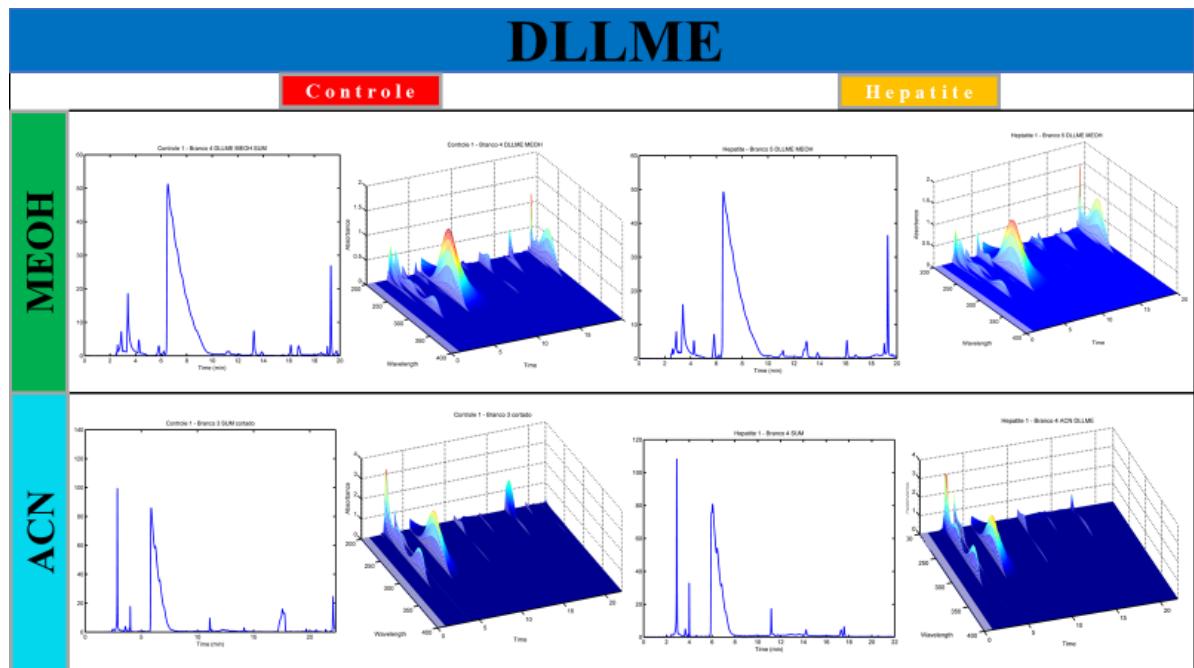


Fonte: A autora (2022)

Com respeito ao tipo de solvente orgânico utilizado na fase móvel, foram testados: ACN/água; MeOH/água; ACN/água + ácido trifluoroacético e MeOH/água + ácido trifluoroacético. No entanto os gradientes acidificados apresentaram problemas de resolução em comparação aos gradientes não acidificados. Dentre eles, o metanol apresentou melhor desempenho na resolução dos picos e na quantidade de informações extraídas, o que representa reproduzibilidade e melhor performance dos modelos metabonômicos aqui reportados. Estes resultados são apresentados na figura 17 e no ANEXO B onde estão expostos os cromatogramas em 2D e 3D das amostras controle e hepatite nos 4 gradientes avaliados.

Por conseguinte, a metodologia mais eficiente achada neste trabalho, para a análise das amostras de soro, é atingida utilizando o método DLLME para o preparo das amostras e gradiente curvo de MeOH para as análises cromatográficas. Estes parâmetros foram aplicados para a análise das 94 amostras deste estudo.

Figura 17 – Comparação dos gradientes MeOH vs ACN no método DLLME.



Fonte: A autora (2022)

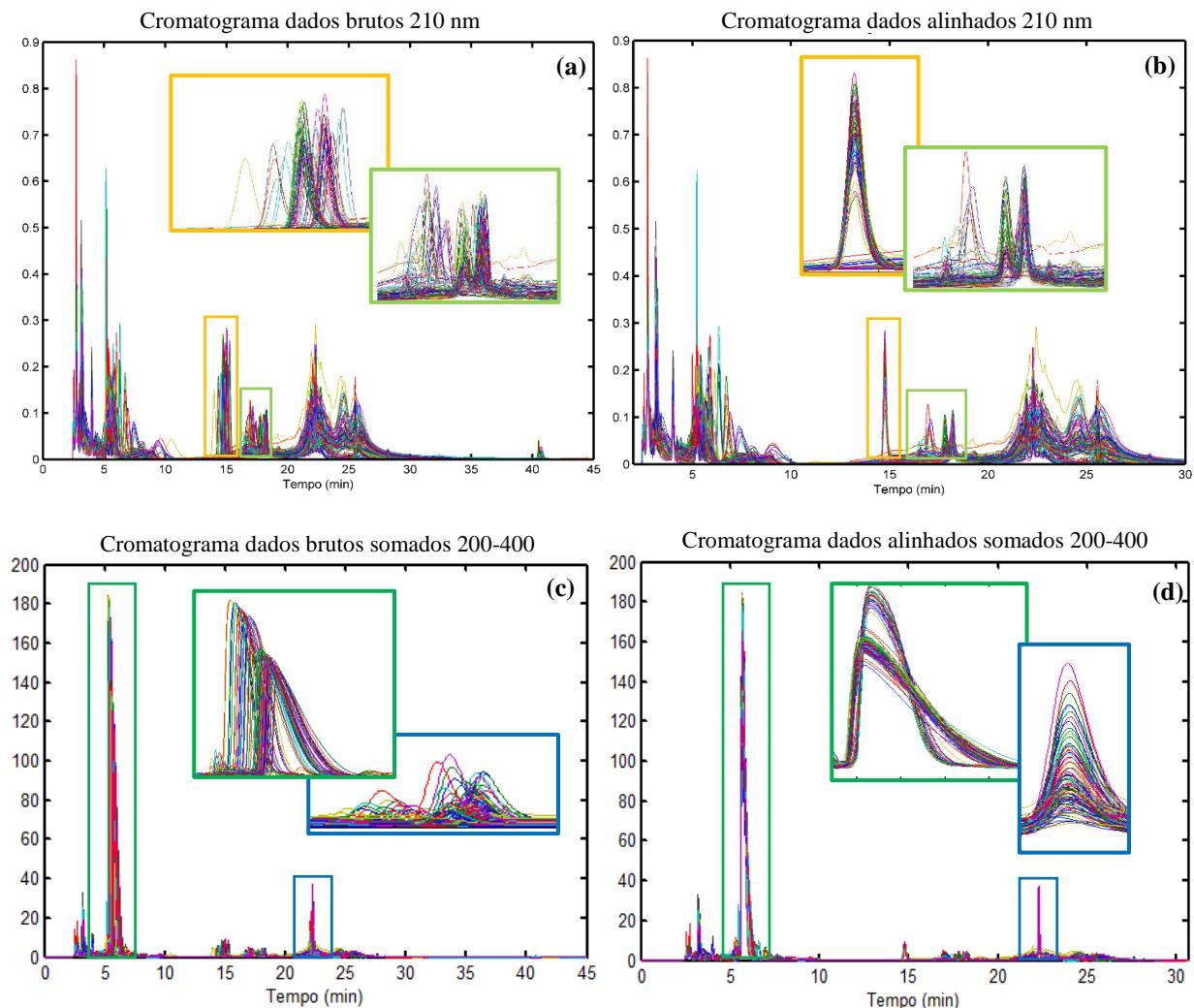
4.2 ESTUDO METABONÔMICO

4.2.1 Pré-processamento dos dados cromatográficos

As 94 amostras tiveram seu sinal subtraído pelo branco para correção de linha de base. Posteriormente foram unidas em uma matriz e ao observar os cromatogramas a 210 nm (figura 18a) e com todos os comprimentos de onda somados (figura 18c), constatou-se problemas de

deslocamentos de picos. Por tanto, foi necessário realizar um pré-processamento para alinhar os picos antes da construção dos modelos metabonômicos, caso contrário o algoritmo poderia interpretar um analito como sendo compostos diferentes. Foi utilizado o algoritmo icoshft modo “whole” para alinhar os picos por deslocamento completo do cromatograma, porém nem todos os picos foram bem alinhados, sendo necessário dividir o cromatograma em intervalos e realizar o alinhamento em cada seção do cromatograma. Deste modo, foi possível realizar os alinhamentos dos cromatogramas a 210 nm e somando os comprimentos de onda como se mostra na figura 18b e 18d respectivamente.

Figura 18 – Cromatogramas com os dados brutos (a) 210 nm e (c) somando os comprimentos de onda. E cromatogramas após alinhamento (b) a 210 nm e (d) somando os comprimentos de onda.

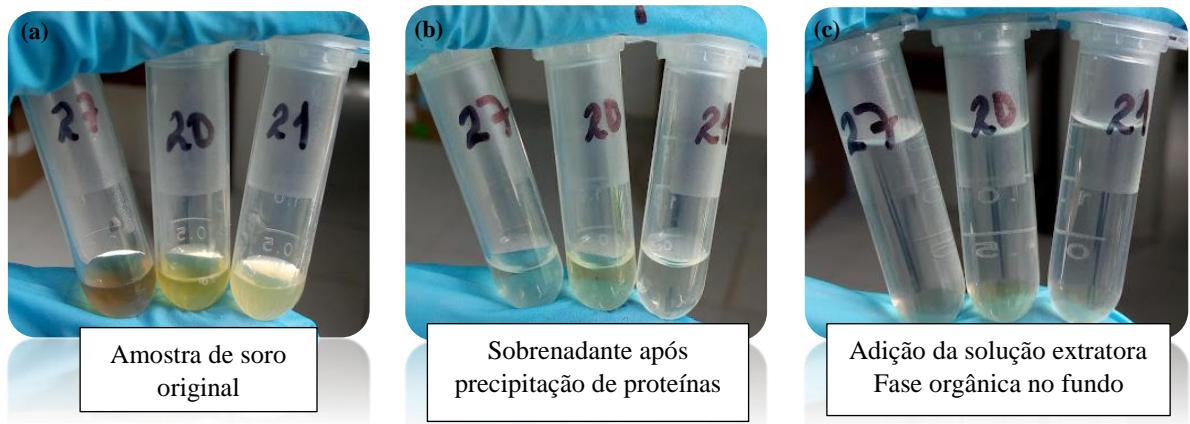


Fonte: A autora (2022)

4.2.2 Análise exploratória

Durante o preparo das 94 amostras do estudo, observou-se que algumas tinham diferenças de cor (figura 19a) e essa diferença persistia após etapa de precipitação de proteínas com acetonitrila (figura 19b) e até no final do preparo na extração da fase orgânica (figura 19c). Em vista disto, o espectro UV visível foi registrado em todas as amostras (200 a 800 nm), com o intuito de avaliar se a coloração das amostras apresentava alguma informação relacionada ao estágio da doença.

Figura 19 – Cores apresentadas nas amostras do estudo. (a) amostra original, (b) após precipitação de proteínas e (c) extração.

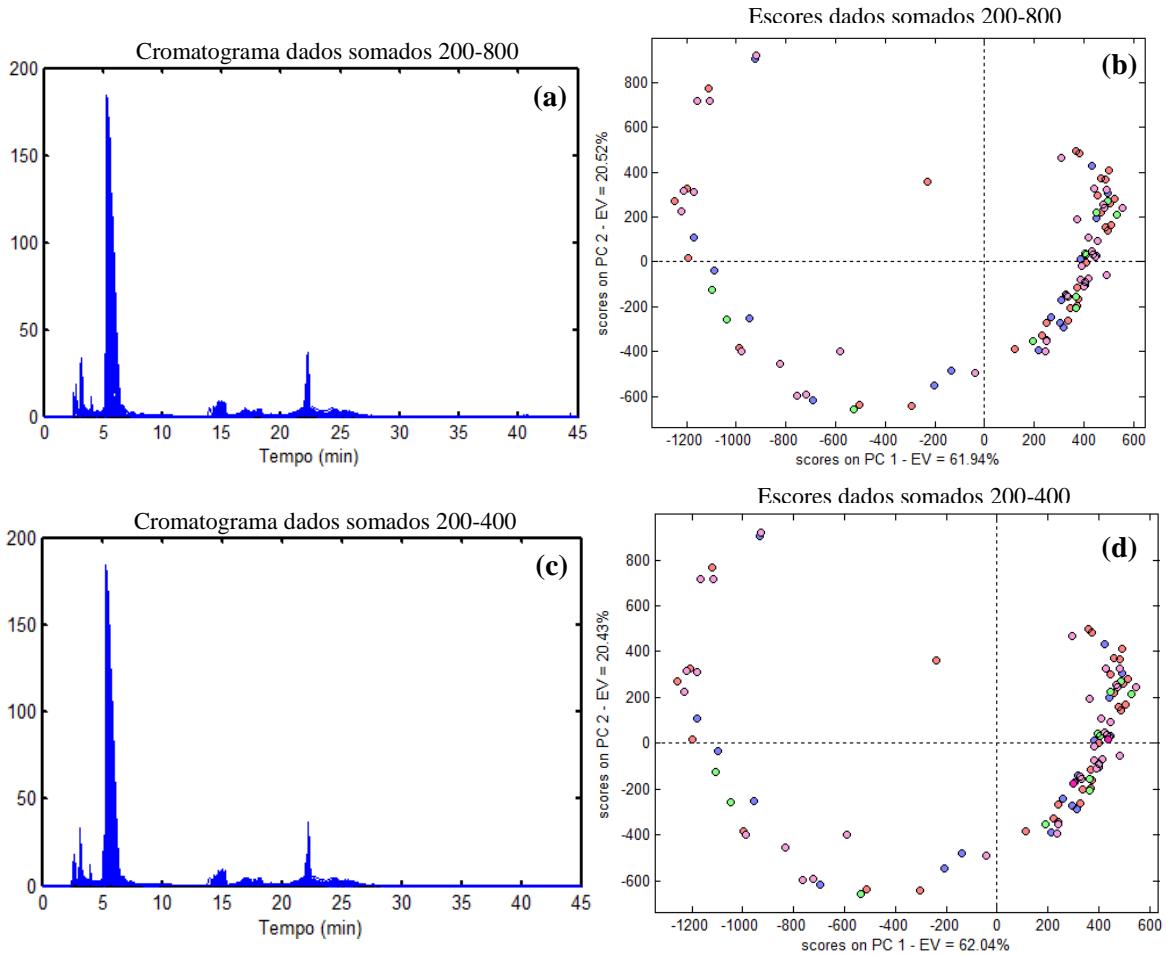


Fonte: A autora (2022)

Os dados obtidos foram avaliados somando os comprimentos de onda de 200 a 400 nm e de 200 a 800 nm. No entanto, os cromatogramas obtidos e a análise exploratória realizada foram exatamente iguais entre si, tal como é apresentado na figura 20, o que indica que a região do visível não apresentou nenhuma informação relacionada ao estadiamento, embora a aparência física de algumas amostras exibe diferenças nas cores. Dessa forma, os dados utilizados para construção dos modelos metabonômicos foram do intervalo 200 a 400 nm e os cromatogramas foram cortados em 30 minutos dado que após este tempo nenhum sinal analítico é observado.

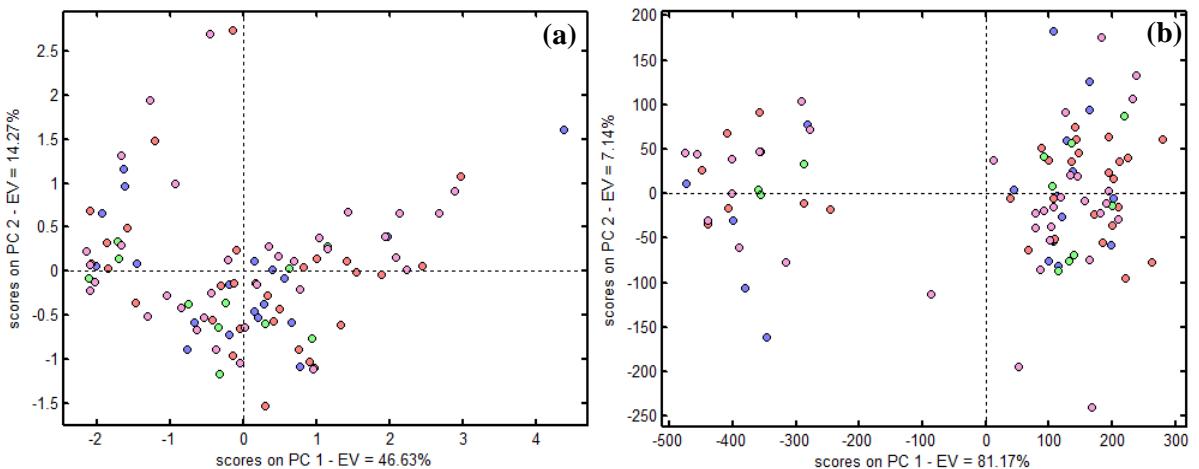
Uma vez escolhidos os dados somados de 200-400 nm para realizar as modelagens, foram pré-processados como foi reportado na seção 4.2.1 e realizado novamente análise de componentes principais utilizando também dados registrados apenas a 210 nm, para avaliar qual destas abordagens seria mais interessante empear na análise classificatória. Os escores são apresentados na figura 21. Observa-se que pela distribuição ou agrupamento natural dos dados, não é possível distinguir os 4 grupos da classificação de Niamey, todas as amostras se revelam misturadas e sem apresentar tendência de separação.

Figura 20 – Cromatogramas e gráficos de escores dos dados brutos somados entre 200-800 nm (a) e (b). E entre 200-400 nm (c) e (d) respectivamente. Foram selecionadas 5 PC's em cada caso com 93,8% de variância explicada nos dados de 200-800 nm e 94,9%, nos dados de 200 a 400nm.



Fonte: A autora (2022)

Figura 21 – Escores dos dados (a) 210 nm e (b) somando os comprimentos de onda. Foram selecionadas 5 PC's com uma variância explicada de 78% para os dados a 210 nm e 96,1% para os dados somados 200-400nm.

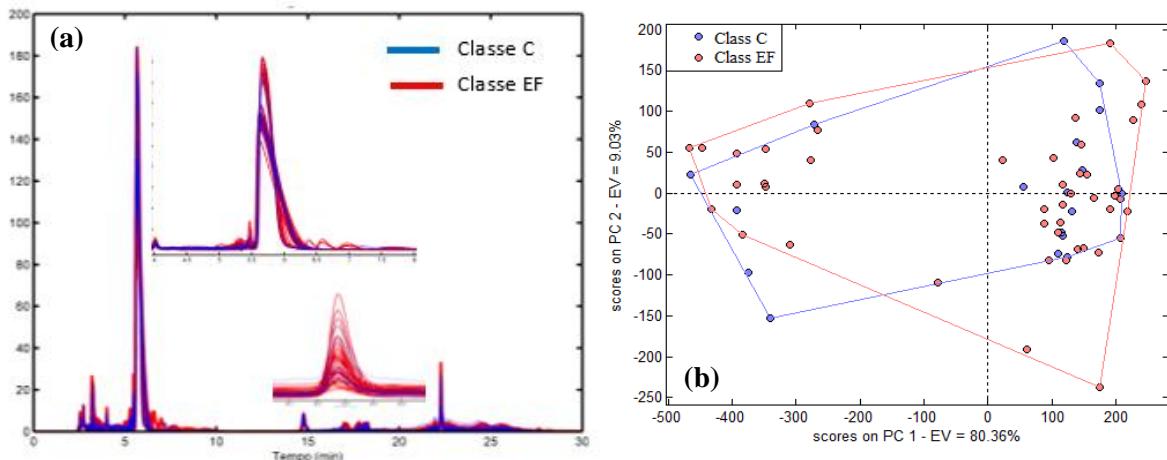


Fonte: A autora (2022)

Para usar as informações contidas em todos os comprimentos de onda registrado na região do ultravioleta, optou-se por trabalhar usando os comprimentos de onda somados de 200-400 nm uma vez que o cromatograma apresenta mais picos. As análises foram realizadas buscando a discriminação entre os estágios leve (Classe C) e avançado (Classe EF), este último foi formado pelas amostras da classe E e da classe F inicialmente. Em vista da impossibilidade de identificar as 4 classes da classificação de Niamey na análise exploratória foi decidido modelar, em primeira instância, os estágios de extremo da FPP, classes C e EF, removendo assim as amostras do estágio intermediário (classe D).

Na figura 22a é apresentado o cromatograma das amostras das classes C e EF que foram usadas na construção dos modelos metabonômicos e reportados na seção 4.2.3 *análise classificatória*. A figura 22b apresenta o gráfico dos escores do mesmo conjunto de dados, onde é observada sobreposição das classes.

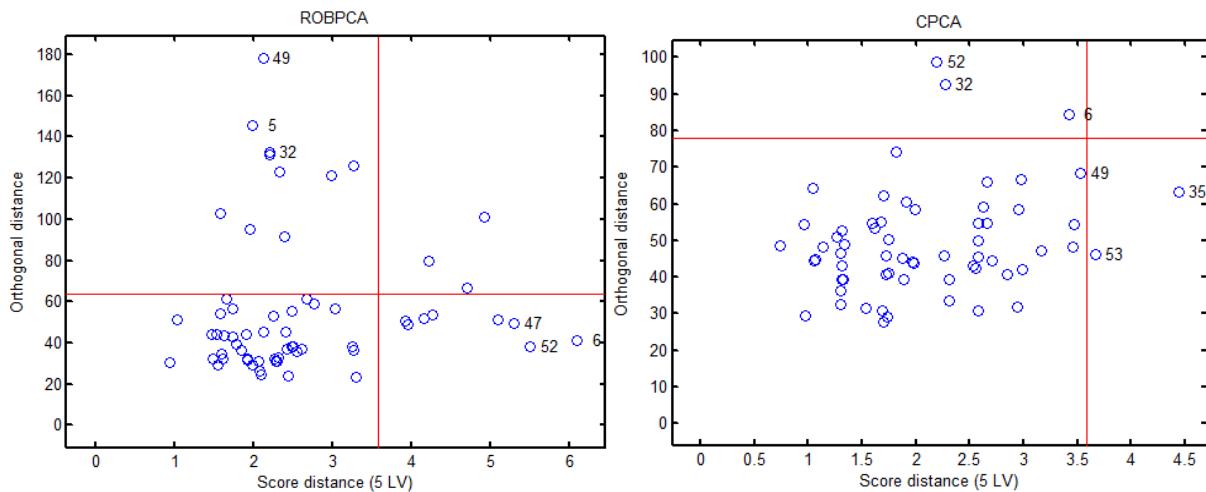
Figura 22 – Dados contendo amostras da classe C e classe EF, com os comprimentos de onda somados.
(a) Cromatograma. (b) gráfico de escores.



Fonte: A autora (2022)

Um aspecto importante a ser avaliado na análise exploratória é a presença de amostras anômalas ou *outliers* que podem influenciar na rotação dos modelos. Para detectá-las foi utilizado a PCA robusta (ROBPCA) que é mais sensível que a PCA clássica (CPCA) a este tipo de amostras. Os resultados se apresentam na figura 23. Utilizando 5 componentes principais com 95,84% de variância explicada, é possível observar que no conjunto de dados não apresenta amostras anômalas, pois não há amostras com distância ortogonal e distância de escore elevadas.

Figura 23 – Analise de componentes principais robusto e clássico.



Fonte: A autora (2022)

4.2.3 Análise classificatória

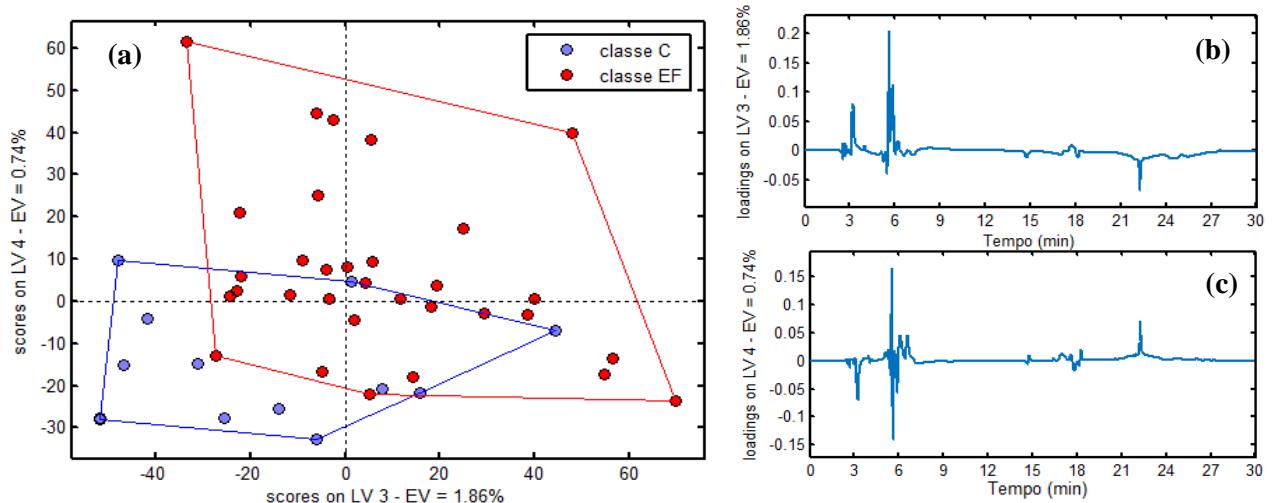
O PLS-DA foi o primeiro modelo classificatório avaliado, o número ideal de variáveis latentes (VL) necessárias para construção do modelo com o menor erro associado foi de 4, com 92% de variância explicada. Os gráficos de escores e pesos são apresentados na figura 24. Dado que é impossível vislumbrar os dados com as 4 VL simultaneamente, foi selecionado o par de variáveis latentes que, visualmente, apresentassem a melhor separação das duas classes, neste sentido as variáveis 3 e 4 foram selecionadas.

No gráfico de escores (figura 24a) se observa uma clara tendência de separação entre as duas classes, no entanto essa separação não é completa pois há uma zona de confusão ou interposição das classes. Ao observar os escores separadamente e em paralelo com os pesos (figura 24b e 24c), observa-se que a VL 4 apresenta maior distinção entre as classes do que a VL 3. No eixo positivo dos escores na VL 4 ficam a maior parte das amostras da classe EF enquanto que no eixo negativo no mesmo gráfico, ficam praticamente todas as amostras da classe C (exceto 2 posicionadas no eixo positivo). Os pesos da VL 4 associados a essa separação estão em torno de 3 e 22 minutos. Por outro lado, é importante ressaltar que, mesmo a VL 4 apresentando maior discriminação entre as classes, a melhor separação é alcançada no sentido da diagonal, por exemplo, ao traçar uma linha diagonal entre as classes no gráfico de escores da figura 24. Desta forma é atingida uma melhor separação entre as duas classes, o que indica que a VL 3 e a VL 4 em conjunto fornecem o melhor resultado para distinção da classe C e da classe EF obtido pelo modelo.

Para avaliar a eficiência do modelo é necessário examinar a tabela de contingência e as métricas de desempenho como sensibilidade, especificidade e exatidão, as quais são

apresentadas na tabela 4. A diagonal verde da tabela 4, representa os acertos do modelo e por sua vez a diagonal laranja representa os erros obtidos no modelo. A exatidão obtida no conjunto de treinamento e teste são muito diferentes, 80% e 53% respectivamente. O que indica que o modelo pode ter problemas de sobre ajuste, porém neste caso essa diferença ocorreu devido ao baixo número de amostras usado para testar o modelo e devido a classes desbalanceadas. Porém, vale a pena destacar que o número de erros é elevado para utilizar este modelo no estadiamento.

Figura 24 – Gráfico de escores (a) e pesos (b) e (c) no algoritmo PLS-DA



Fonte: A autora (2022)

Tabela 4 – Matriz de contingência e métricas de desempenho obtido do algoritmo PLS-DA

Treinamento						
Real/Predito	Classe C	Classe EF	Sensibilidade	Especificidade	ER	Exatidão
Classe C	10	3	0,77	0,81	9	80%
Classe EF	6	26	0,81	0,77		
Test						
Classe C	2	4	0,33	0,69	8	58%
Classe EF	4	9	0,69	0,33		

Fonte: A autora (2022)

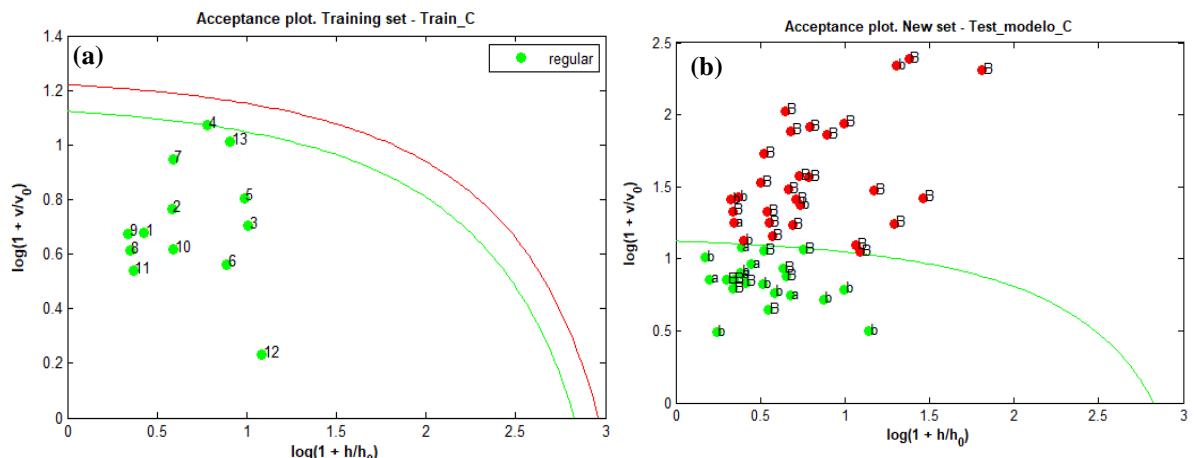
A diferença da análise exploratória (figura 22b, página 50), onde é observado que não é possível distinguir entre as duas classes da doença, o modelo PLS-DA forneceu resultados positivos (figura 24a), pois apresentou tendência de separação ou discriminação entre os pacientes com estagio leve (classe C) daqueles pacientes com estágio avançado (classe EF). No entanto essa tendência de separação ainda deixa muito a desejar, especialmente na exatidão do

grupo de teste (58%), por tanto o modelo não se apresenta robusto o suficiente para ser implementado no diagnóstico clínico.

O modelo DD-SIMCA foi usado por ser possível modelar as informações com base apenas no sinal instrumental e por fornecer mais informações que nos modelos discriminantes. Além disso, no DD-SIMCA os dados guiam a classificação, assim amostras mal rotuladas não afetam a qualidade do modelo.

DD-SIMCA da classe C

Figura 25 – Gráficos do modelo DD-SIMCA para o grupo de treinamento da classe C (a) e o grupo teste (b). Em verde as amostras da classe de treinamento, em vermelho as amostras anômalas preditas pelo modelo. A linha verde indica o intervalo de confiança (99%), já a linha vermelha indica o limite a partir do qual encontram-se as amostras não pertencentes a classe modelada. As amostras marcadas como “a” correspondem a amostras do grupo teste da classe C; as marcadas como “b” correspondem ao grupo teste da classe EF e as marcadas como “B”, ao grupo treinamento da classe EF. Foi utilizado ROBPCA com 5 PC’s.



Fonte: A autora (2022)

Tabela 5 – Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelagem da classe C

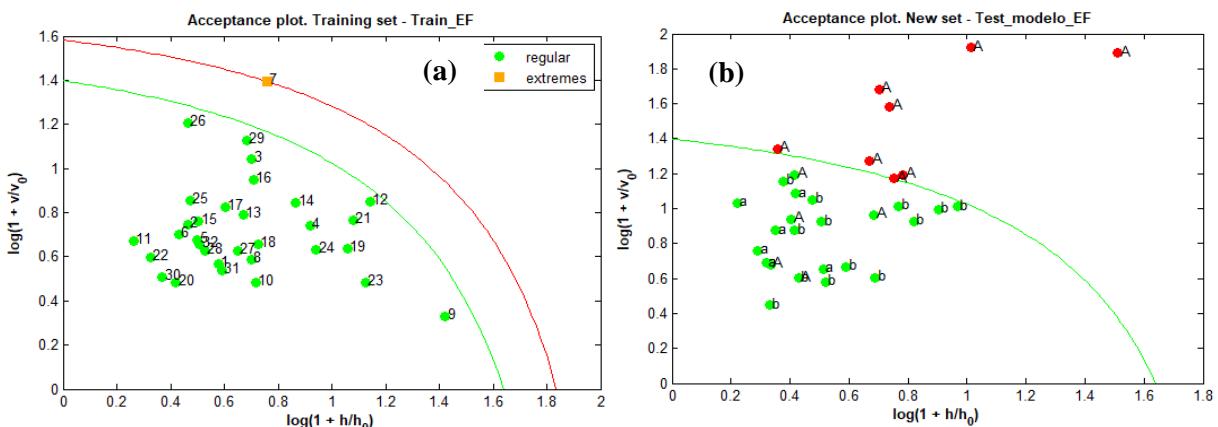
Real/Predito	Test						Exatidão
	Classe C	Classe EF	Sensibilidade	Especificidade	ER		
Classe C	5	1	0,83	0,62	18	64%	
Classe EF	17	28	0,62	0,83			

Fonte: A autora (2022)

Neste caso o modelo construído baseia-se na modelagem somente da classe C utilizando 13 amostras no conjunto de treinamento (figura 25a). Posteriormente, para testar o modelo, foi utilizado um conjunto de amostras de teste conformado por amostras de teste da classe C e amostras de teste e treinamento da classe EF, 51 amostras em total. As amostras posicionadas dentro da região delimitada pela linha verde são categorizadas como pertencendo a classe C e se apresentam de cor verde. Já as amostras que se encontram por fora da região modelada são categorizadas como não pertencendo a classe C e se apresentam em cor vermelha (figura 25b). do total de 6 amostras de teste da classe C, apenas 1 foi categorizada como não pertencendo a classe e do total de 45 amostras do conjunto EF, 17 foram classificadas como pertencendo a classe C, a exatidão obtida do modelo foi de 64% tal e como é apresentado na tabela 5.

DD-SIMCA da classe EF

Figura 26 – Gráficos do modelo DD-SIMCA para o grupo de treinamento da classe EF (a) e o grupo teste (b). em verde as amostras da classe de treinamento, em laranja as amostras classificadas como extremos, em vermelho as amostras anômalas preditas pelo modelo. A linha verde indica o intervalo de confiança (99%), já a linha vermelha indica o limite a partir do qual encontram-se as amostras não pertencentes a classe modelada. As amostras marcadas como “a” correspondem a amostras do grupo teste da classe C; as marcadas como “b” correspondem ao grupo teste da classe EF e as marcadas como “A”, ao grupo treinamento da classe C. Foi utilizado ROBPCA com 9 PC's



Fonte: A autora (2022)

No caso da modelagem da classe EF, foram utilizadas 32 amostras na construção do modelo (figura 26a) das quais uma foi categorizada como sendo de extremo. O conjunto de amostras utilizadas para testar o modelo foi conformado por amostras de teste da classe EF e amostras de teste e treinamento da classe C, um total de 32. As amostras posicionadas dentro da região delimitada pela linha verde são categorizadas como pertencendo a classe EF e se apresentam de cor verde. Já as amostras que se encontram por fora da região modelada são

categorizadas como não pertencendo a classe EF e se apresentam em cor vermelha (figura 26b). Todas as amostras de teste da classe EF foram corretamente classificadas, atingindo o valor máximo de sensibilidade (1) e do total de 19 amostras do conjunto C, 8 foram classificadas como não pertencendo a classe EF, a exatidão obtida do modelo foi de 65% tal e como é apresentado na tabela 6. A amostra 7 é uma amostra de extremo e pode indicar que nessa amostras há metabólitos que indicam que esta esteja no limite da fronteira a não pertencer a classe EF.

Tabela 6 – Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe EF

Test						
Real/Predito	Classe EF	Classe C	Sensibilidade	Especificidade	ER	Exatidão
Classe EF	13	0	1	0,42	11	65%
Classe C	11	8	0,42	1		

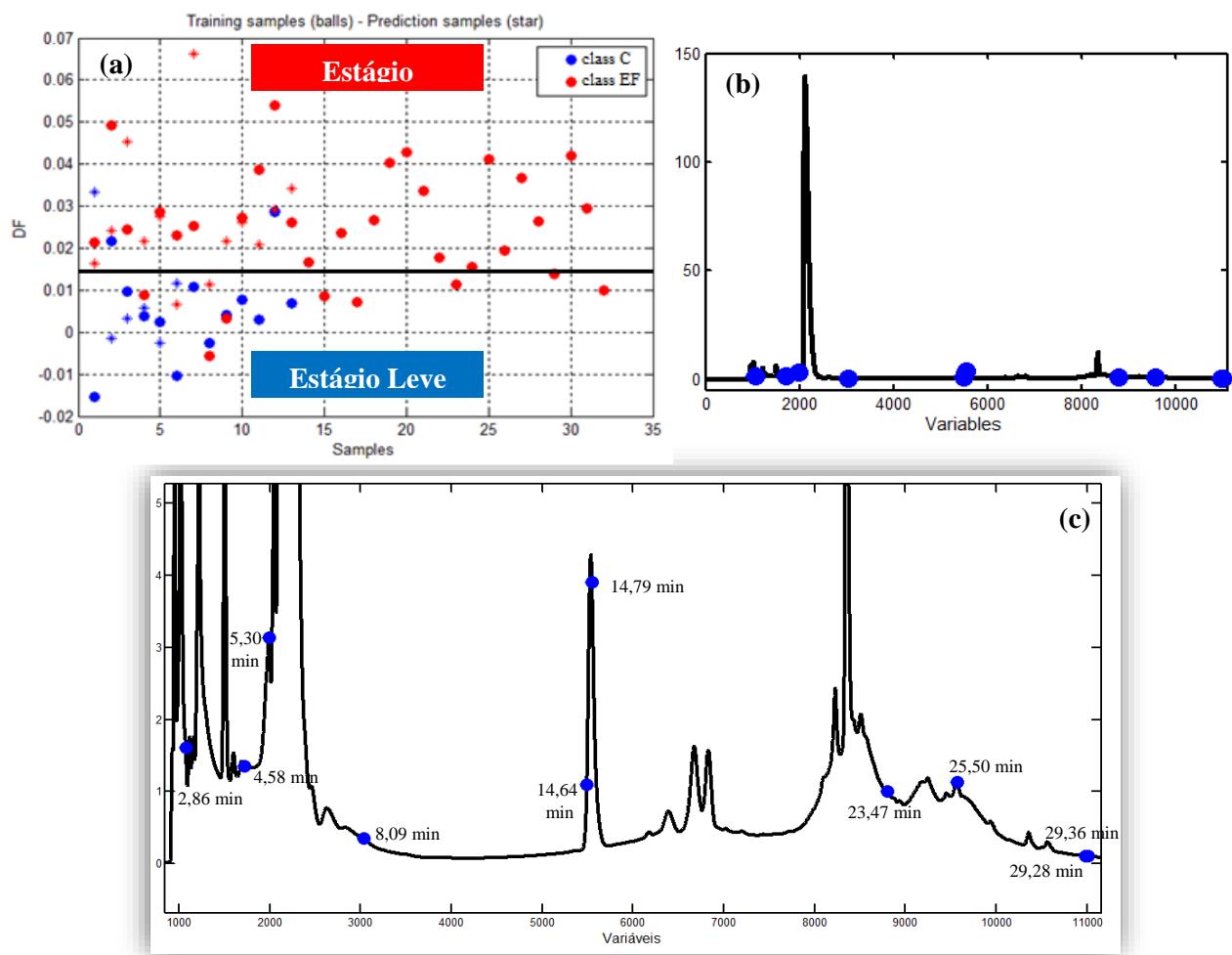
Fonte: A autora (2022)

Algoritmos como o DD-SIMCA que modelam uma classe, podem ser uma ferramenta útil para o diagnóstico diferencial da FFP, visto o modelo poderá indicar se uma amostra desconhecida ou suspeita pertence ou não a um estágio específico da doença. Isto se torna mais interessante em modelos com valores altos de sensibilidade como os aqui reportados, tanto no modelo da classe C quanto no modelo da classe EF foram obtidos valores de sensibilidade de 83% e 100% respectivamente. E dado que a sensibilidade é um tipo de acurácia para as amostras da classe alvo, os modelos da classe C e EF terão a capacidade de classificar corretamente amostras futuras como pertencendo a classe modelada com 83% e 100% de exatidão respectivamente. Por outro lado, a especificidade também é um tipo de acurácia, mas para as amostras que não pertencem a classe modelada, isto é a fração de amostras negativas classificadas corretamente. Os valores de especificidade obtidos na modelagem de ambas das classes (C e EF) não foram tão altos quanto aos da sensibilidade, desta forma os modelos ao serem mais sensíveis que específicos, poderão apresentar mais casos de falsos positivos que de falso negativo. Nenhum tipo de erro é aceitável, mas se for o caso seria melhor ter um FP e não um FN no diagnóstico clínico. Por exemplo em um caso de diagnóstico FN, um paciente com o estágio avançado da FPP será diagnosticado como não estando nesse nível da doença. Provavelmente esta pessoa não terá os cuidados e o tratamento que essa condição final da doença exige e o pior pode vir a suceder. Caso oposto de um FP onde a pessoa tomaria as previdências necessárias e mais exames seriam feitos para confirmar essa condição.

De forma global, ambas classes modeladas apresentaram exatidão praticamente igual, 64% na classe C e 65% da classe EF, os resultados obtidos foram melhores do que os fornecidos pelo modelo PLS-DA (58%), no entanto ainda continua sendo um valor baixo de exatidão.

Até agora tem sido usado algoritmos que utilizam o cromatograma inteiro para a construção dos modelos quimiométricos com uma exatidão máxima atingida de 65% para a predição. À visto disso algoritmos como GA-LDA pode ser uma boa opção para construção de modelos utilizando picos que realmente influenciem a separação das classes e assim obter melhores resultados de exatidão. Por outro lado, algoritmos com seleção de variáveis como o GA-LDA permitem a redução dos dados utilizados na construção de modelos e, portanto, a diminuição do esforço computacional. A figura 27 apresenta os gráficos dos escores e variáveis selecionadas pelo algoritmo GA-LDA.

Figura 27 – Gráficos do modelo GA-LDA. (a) escores. (b) Cromatograma com as variáveis selecionadas pelo modelo. (c) Zoom do cromatograma indicando os tempos de retenção selecionados.



Fonte: A autora (2022)

A seleção de variáveis por algoritmo genético foi feita ao selecionar o melhor resultado em termos de exatidão e parcimônia após 1000 cálculos. Nesse sentido os resultados da melhor opção encontrada são apresentados na figura 27 e na tabela 7. Foram selecionadas 10 variáveis (figura 27b) para realizar a classificação em uma função discriminante das duas classes (figura 27a). Os tempos de retenção correspondentes as variáveis selecionadas são apresentadas na figura 27c. Vale a pena ressaltar que nenhuma das variáveis selecionadas pelo algoritmo genético corresponde a região do sinal intenso do cromatograma (entre 5 e 6 minutos), por tanto esse sinal mesmo sendo de alta intensidade, não é útil para a discriminação das duas classes. É comum em análises metabonômicas alguns constituintes minoritários serem os que melhor classificam uma determinada doença, o que torna a abordagem do GA-LDA interessante. Na figura 27a observa-se a discriminação dos dois estágios da doença, com a classe EF na região superior e a classe C na região inferior. Apesar de apresentar algumas amostras confundidas, por exemplo amostras da classe EF na região da classe C e vice-versa, o modelo GA-LDA apresentou excelentes resultados, para o grupo de treinamento e grupo teste foi obtido uma exatidão de 79% e 83% respectivamente, tal e como se apresenta na tabela 7.

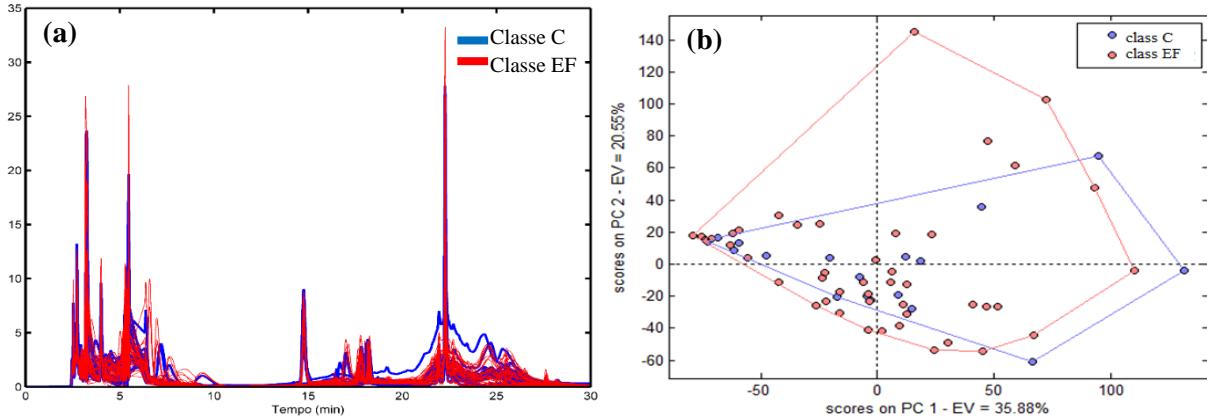
Tabela 7 – Matriz de contingência e métricas de desempenho obtido do algoritmo GA-LDA

Treinamento							
Real/Predito	Classe C	Classe EF	Sensibilidade	Especificidade	ER	Exatidão	
Classe C	11	2	0,84	0,75	10	79%	
Classe EF	8	24	0,75	0,84			
Test							
Classe C	5	1	0,83	0,84	3	83%	
Classe EF	2	11	0,84	0,83			

Fonte: A autora (2022)

Em vista que os melhores resultados foram obtidos no modelo GA-LDA utilizando tempos de retenção diferentes do sinal mais intenso do cromatograma que está localizado exatamente entre 5,53 e 6,34 minutos, testou-se remover este sinal do cromatograma e aplicar os mesmos algoritmos de classificação, já reportados, neste novo conjunto de dados reduzidos. Isso foi feito, pois em algoritmos que usam a informação dos dados inteiro como DD-SIMCA e PLS os resultados da modelagem podem ser afetados pela presença de constituintes com intensidade de sinal elevada que não são correlacionados ao objetivo da modelagem. O cromatograma e os escores da análise de componentes principais é apresentado na figura 28.

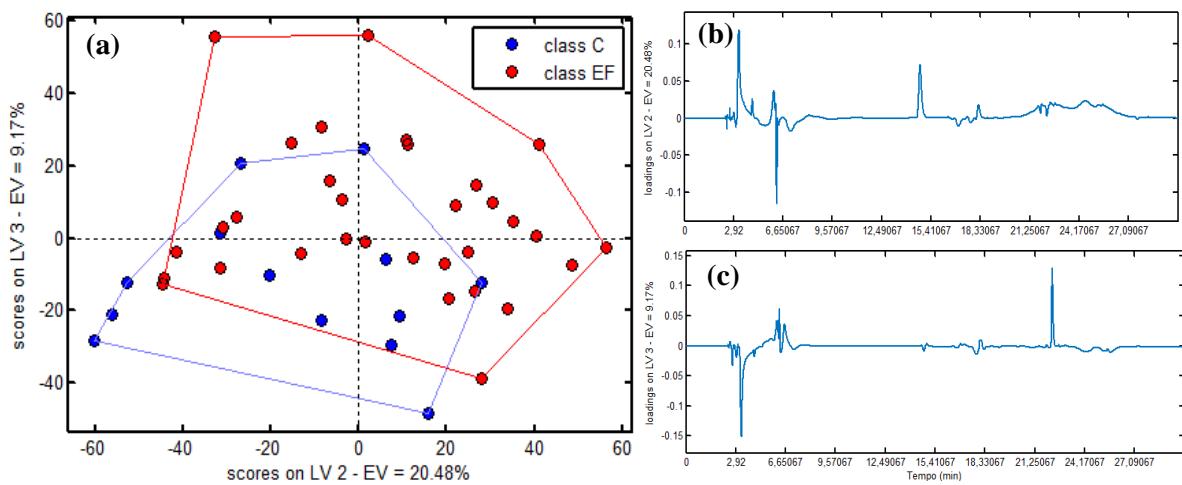
Figura 28 – Dados contendo amostras da classe C e classe EF, com o sinal entre 5,53 – 6,34 min, removido. (a) Cromatograma. (b) gráfico de escores.



Fonte: A autora (2022)

Observa-se nos escores da análise exploratória (figura 28b) deste conjunto de dados reduzidos que apesar de apresentar sobreposição das duas classes, nota-se uma melhor tendência de separação do que os escores com os dados completos (figura 22b, página 50). Desta forma infere-se que a remoção do sinal intenso do cromatograma pode favorecer a separação entre as classes e por tanto os algoritmos de análise classificatória foram aplicados a este conjunto de dados.

Figura 29 – Gráfico de escores (a) e pesos (b) e (c) do modelo PLS-DA nos dados reduzidos.



Fonte: A autora (2022)

Tabela 8 – Matriz de contingência e métricas de desempenho do modelo PLS-DA obtidos nos dados com sinal removido.

Treinamento						
Real/Predito	Classe C	Classe EF	Sensibilidade	Especificidade	ER	Exatidão
Classe C	11	2	0,85	0,75	10	78%
Classe EF	8	24	0,75	0,85		
Test						
Classe C	3	3	0,50	0,85	5	74%
Classe EF	2	11	0,85	0,50		

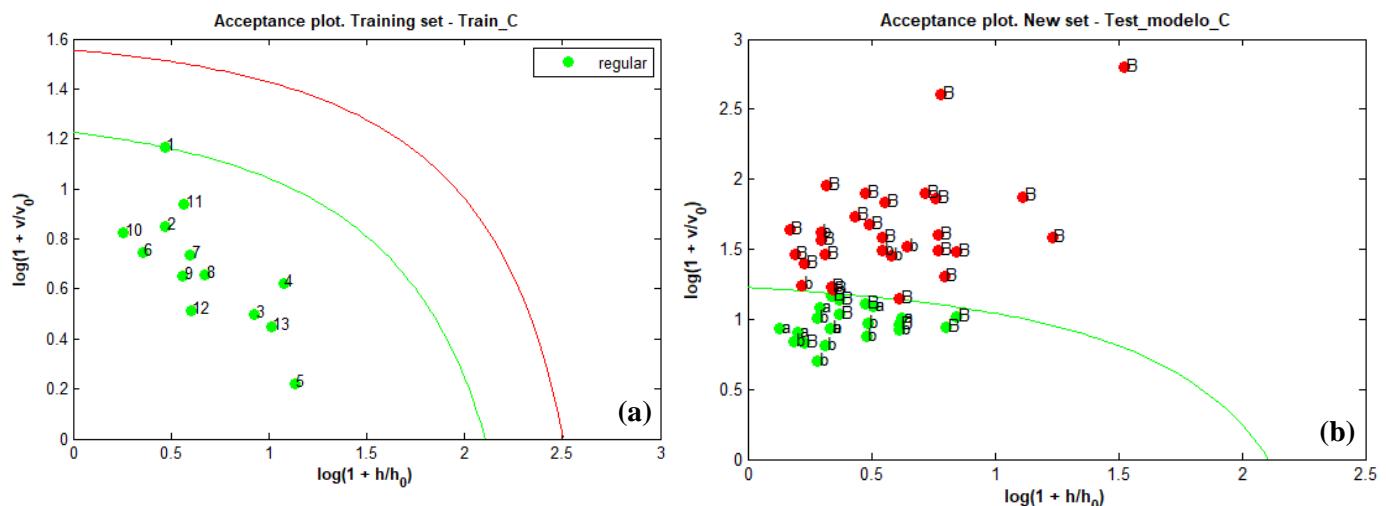
Fonte: A autora (2022)

Os resultados do modelo PLS-DA nos dados reduzidos, se apresentam na figura 29 e na tabela 8. O número ideal de variáveis latentes para construção do modelo também foi 4, com 71% de variância explicada. Foram selecionadas as variáveis latentes 2 e 3 para visualizar os resultados. Os escores de variável latente deste modelo com os dados reduzidos, apresentaram uma melhor tendência de separação das classes quando comparado aos dados completos, isso é demonstrado na matriz de contingencia da tabela 8. Ao avaliar a eficiência do modelo com as métricas de desempenho, apresentadas na mesma tabela, percebe-se que os resultados obtidos foram melhores, com valores de sensibilidade, especificidade e exatidão maiores e menor número de erros no grupo teste. Isso indica que este sinal não está correlacionado com o objetivo da modelagem e que esta informação atrapalhava o algoritmo na predição do estadiamento da doença. A exatidão obtida no modelo PLS-DA com os dados completos foi de 58% (tabela 4), já para os dados com o sinal intenso removido, a exatidão obtida foi de 74%, uma melhoria significativa. Vale a pena ressaltar também que a diferença entre a exatidão do grupo de treinamento e teste no modelo PLS-DA com os dados reduzidos, é pequena (78% e 74% respectivamente) o que indica que o modelo não terá problemas de sobre ajuste, caso oposto ao modelo com os dados completos onde a diferença da exatidão foi grande (80% treinamento – 58% teste).

Visto que foram obtidos resultados melhores com o conjunto de dados reduzidos, procedeu-se a realizar também a modelagem de cada classe com o algoritmo DD-SIMCA, os resultados são reportados a seguir.

DD-SIMCA da classe C nos dados reduzidos

Figura 30 – Gráficos do modelo DD-SIMCA para o grupo de treinamento da classe C (a) e o grupo teste (b). em verde as amostras da classe de treinamento, em vermelho as amostras anômalas preditas pelo modelo. A linha verde indica o intervalo de confiança (96%), já a linha vermelha indica o limite a partir do qual encontram-se as amostras não pertencentes a classe modelada. As amostras marcadas como “a” correspondem a amostras do grupo teste da classe C; as marcadas como “b” corresponde ao grupo teste da classe EF e as marcadas como “B”, ao grupo treinamento da classe EF. Foi utilizado ROBPCA com 5 PC’s.



Fonte: A autora (2022)

Tabela 9 – Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe C

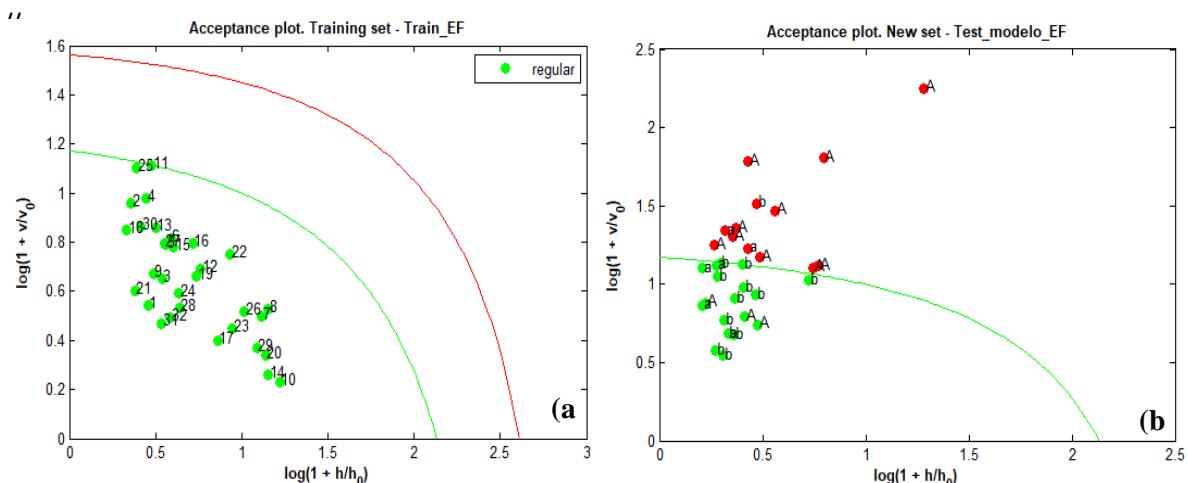
Real/Predito	Test					
	Classe C	Classe EF	Sensibilidade	Especificidade	ER	Exatidão
Classe C	6	0	1	0,64	16	68%
Classe EF	16	29	0,64	1		

Fonte: A autora (2022)

O modelo da classe C se apresenta na figura 30a e foi construído usando 13 amostras no conjunto de treinamento. Para o conjunto de teste foram utilizadas amostras de teste da classe C e amostras de teste e treinamento da classe EF, 51 amostras em total. Todas as amostras de teste C foram corretamente classificadas, ou seja, foram localizadas dentro da região delimitada pela linha verde. Do total de 45 amostras do conjunto EF, 16 foram classificadas como pertencendo a classe C (figura 30b). A exatidão obtida do modelo foi de 68% tal e como é apresentado na tabela 9. De forma geral a modelagem da classe C com os dados reduzidos forneceu resultados melhores uma vez que os valores da sensibilidade, especificidade e exatidão foram maiores do que os obtidos com a modelagem dos dados completos.

DD-SIMCA da classe EF nos dados reduzidos

Figura 31 - Modelo DD-SIMCA para o grupo de treinamento da classe EF (a) e o grupo teste (b). em verde as amostras da classe de treinamento, em laranja as amostras classificadas como extremos, em vermelho as amostras anômalas preditas pelo modelo. A linha verde indica o intervalo de confiança (95%), já a linha vermelha indica o limite a partir do qual encontram-se as amostras não pertencentes a classe modelada. As amostras marcadas como “a” correspondem a amostras do grupo teste da classe C; as marcadas como “b” corresponde ao grupo teste da classe EF e as marcadas como “A”, ao grupo treinamento da classe C. Foi utilizado ROBPCA com 12 PC’s.



Fonte: A autora (2022)

Tabela 10 – Matriz de contingência e métricas de desempenho obtido do algoritmo DD-SIMCA na modelando a classe EF

Real/Predito	Test						Exatidão
	Classe EF	Classe C	Sensibilidade	Especificidade	ER		
Classe EF	12	1	0,92	0,63	8	75%	
Classe C	7	12	0,63	0,92			

Fonte: A autora (2022)

O modelo da classe EF se apresenta na figura 31a e foi construído usando 32 amostras no conjunto de treinamento. Para o conjunto de teste foram utilizadas amostras de teste da classe EF e amostras de teste e treinamento da classe C, 32 amostras no total. Das 13 amostras de teste EF, apenas uma foi classificada como não pertencendo a classe. Do total de 19 amostras do conjunto C, 7 foram classificadas como pertencendo a classe EF (figura 31b). A exatidão obtida do modelo foi de 75% tal e como é apresentado na tabela 10. A modelagem da classe EF com os dados reduzidos também forneceu resultados melhores do que os obtidos com a modelagem dos dados completos. Especialmente no valor da exatidão que aumentou de 65% para 75%.

Dado que os modelos construídos com o PLS-DA e o DD-SIMCA forneceram melhores resultados utilizando os dados reduzidos, o algoritmo GA-LDA também foi testado neste conjunto de dados. No entanto, resultados significativamente melhores do que os previamente reportados para o modelo GA-LDA, não foram obtidos. Isto indica que o desempenho do algoritmo de seleção de variáveis é independente da presença do sinal intenso no cromatograma, caso oposto dos algoritmos que modelam os dados completos como o PLS-DA e o DD-SIMCA que apresentaram resultados diferentes quando utilizado o cromatograma reduzido.

Dos modelos quimiométricos construídos e avaliados até agora para as classes C e EF, o GA-LDA apresentou o melhor desempenho, com 83% de exatidão empregando 10 variáveis. Para efeitos comparativos e dado que o numero das amostras nas classes é desbalanceado, a tabela 11 apresenta os valores da sensibilidade, especificidade, número de erros e número de variáveis latentes, componentes principais ou variáveis, segundo cada caso, do grupo teste dos modelos construídos.

Como pode ser observado da tabela 11, em termos de sensibilidade e especificidade, os dados reduzidos apresentaram métricas com valores mais altos, quando comparado com as métricas dos dados completos. Ao comparar as métricas entre os modelos PLS-DA e DD-SIMCA com os dados reduzidos nas duas classes, o DD-SIMCA apresenta melhor desempenho e, por sua vez, os modelos construídos para as classes C e EF são mais sensíveis do que específicos, ao fornecer valores mais altos na sensibilidade (100% classe C; 92% classe EF) do que na especificidade (64% classe C; 63 classe EF).

Já o modelo GA-LDA forneceu valores altos e parelhos de sensibilidade e especificidade, tornando-o tanto sensível quanto específico, e ao mesmo tempo se converte no modelo mais simples, visto que para sua construção foi necessárias apenas 10 variáveis que em conjunto é muito menor do que a informação contida em uma variável latente ou uma componente principal. Desta forma a seleção de picos cromatográficos é a melhor opção para modelagem deste tipo de dados.

Tabela 11 – Comparaçao de resultados dos modelos construídos para as classes C e EF.

	PLS-DA		DD-SIMCA				GA-LDA		
	Dados completos		Dados reduzidos		Dados completos				
	Classe C	Classe EF	Classe C	Classe EF	Classe C Modelada	Classe EF Modelada	Classe C	Classe EF	
VL/PC/Var	4		4		5	5	9	12	10
# Erros	8		5		18	16	11	8	3
Sensibilidade	33%	69%	50%	85%	83%	100%	100%	92%	83%
Especificidade	69%	33%	85%	50%	62%	64%	42%	63%	84%

Fonte: A autora (2022)

Os dados cromatográficos modelados forneceram informações úteis para a distinção de pacientes com fibrose periportal leve de aqueles que apresentam fibrose periportal avançada. Isso indica que no soro do sangue há metabolitos que evidenciam o status da doença, o que poderia facilitar e acelerar o diagnóstico especialmente em zonas rurais onde a US não está disponível em todos os centros médicos. Além disso, o soro pode ser coletado com maior frequência do que uma biopsia, e facilmente transportado o que do ponto de vista prático são vantagens enormes quando se compara as análises feitas com a quimiometria com as análises tradicionais. Desta forma tanto o estadiamento quanto o monitoramento da doença torna-se mais acessível e possível de ser acompanhado com maior frequência.

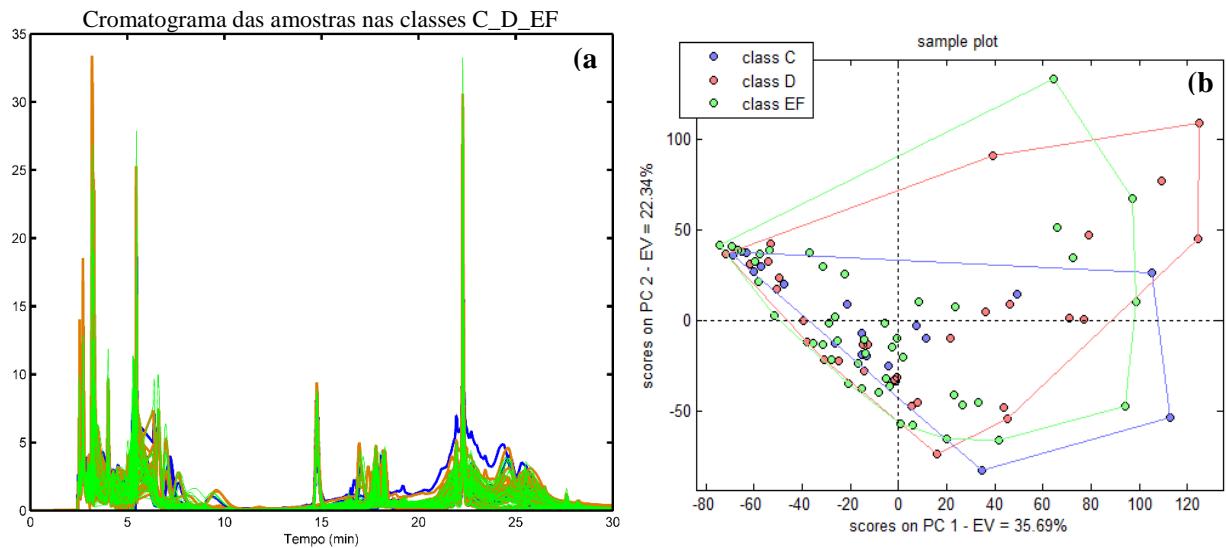
Dos 4 grupos da classificação de Niamey, 3 foram usados na construção dos modelos previamente reportados, a classe C para o estágio leve da doença e as classes E e F para o estágio avançado, 64 amostras em total. As amostras da classe D não foram utilizadas, visto que desde a análise exploratória das 94 amostras em conjunto, não apresentaram tendência de separação, como reportado na seção **4.2.2 Análise exploratória**.

Uma vez que a modelagem dos dados reduzidos apresentou o melhor desempenho, as amostras da classe D foram preditas por este modelo. Sendo que previamente foi removido o sinal intenso do cromatograma (figura 32a) e análise de componentes principais foi realizada, é possível observar os escores na figura 32b.

Para a análise exploratória foram utilizadas 6 PC's com uma variância explicada de 86,9%. No gráfico dos escores (figura 32b) apresenta-se a sobreposição das 3 classes, no entanto, é observado uma certa continuidade em relação da evolução da doença, pois a classe D fica localizada no meio das classes de extremo da FPP, leve (classe C) e avançada (classe EF).

Os algoritmos de análise classificatória foram usados neste conjunto de dados, porém não foram obtidos resultados aceitáveis, pois a incorporação da classe D acaba gerando confundimento das amostras entre as classes, resultando em métricas de desempenho com baixa eficiência quando comparado aos modelos feitos com as classes C e EF. Por tanto não foi possível construir modelos com 3 classes que forneceram um bom desempenho. Isto era esperado dado o índice de classe variável das amostras do grupo D que foram catalogadas pelo médico como sendo de vários tipos dentro da mesma classe D, como foi reportado na seção 3.1.

Figura 32 – Dados contendo amostras da classe C, classe D e classe EF, com o sinal entre 5,53 – 6,34 min, removido. (a) Cromatograma. (b) gráfico de escores.



Fonte: A autora (2022)

Uma vez que a classe D é um estágio intermédio da doença que apresenta variabilidade dentro do índice de classe e que foram desenvolvidos modelos capazes de discriminar os estágios extremos da doença, testou-se utilizar as amostras do grupo D para serem preditas pelos melhores modelos construídos, desta forma avaliar a que classe (C ou EF) as amostras do grupo D são mais próximas. O resultado da predição utilizando o modelo GA-LDA é apresentado na tabela 12.

Como é observado na tabela 12, das 30 amostras do grupo D, 12 foram classificadas como pertencendo a classe EF e 18 como pertencendo a classe C. Note-se que amostras dentro de uma mesma categoria assignada pelo médico foram classificadas em classes diferentes, por exemplo, das 11 amostras do tipo “DC” 6 foram classificadas como sendo da classe C e 5 como pertencendo a classe EF. No caso das 7 amostras do tipo “D+C” 3 foram classificadas na classe C e 4 na classe EF. Evidenciando-se assim a dificuldade na classificação deste estagio intermediário da doença.

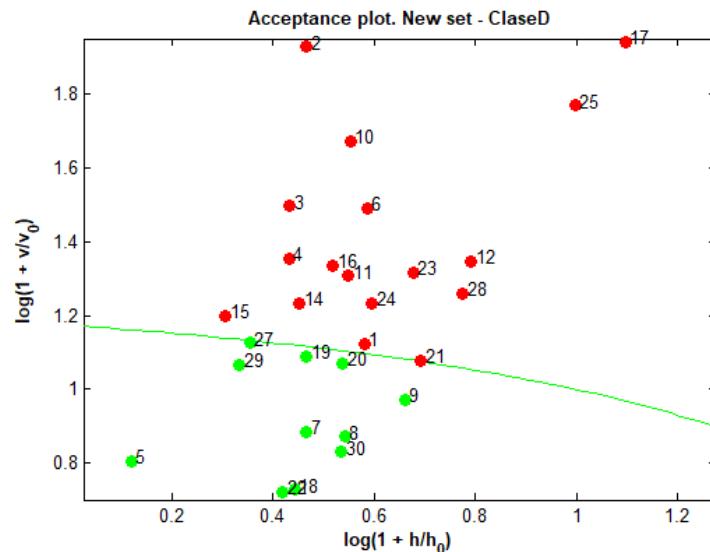
Tabela 12 – predição das amostras da classe D no modelo GA-LDA construído para as classes C e EF. Na coluna de predição 1 corresponde a classe C e 2 a classe EF.

#Amostra Matriz	US/Padrão Fibrose	Predição
1	DC+	1
2	DC+	2
3	D+C	1
4	DC+	2
5	D+C	1
6	D+C	2
7	DC	2
8	DC+	1
9	D+	1
10	DC+	1
11	DC	2
12	DC+/E-	1
13	DC+	1
14	DC	1
15	DC+	1
16	DC+	1
17	DC	2
18	DC	2
19	DC	1
20	D+C	2
21	D	2
22	D+C	2
23	D+C	1
24	DC	1
25	DC	1
26	D+C	2
27	D	1
28	DC	2
29	DC	1
30	DC	1

Fonte: A autora (2022)

O modelo GA-LDA classificou as amostras do grupo D como sendo da classe C ou EF, porém se uma amostra da classe D tiver uma característica que seja diferente da classe C e EF o modelo não irá identificar pois obrigatoriamente tem que classificar as amostras como pertencendo a uma das duas classes. Já no DD-SIMCA essas diferenças podem ser identificadas, visto que o modelo pode classificar as amostras como pertencendo ou não a uma determinada classe alvo. As amostras da classe D foram preditas utilizando o melhor modelo obtido no DD-SIMCA da classe EF. Os resultados são apresentados na figura 33 e na tabela 13.

Figura 33 – Predição das amostras da classe D no modelo DD-SIMCA da classe EF



Fonte: A autora (2022)

Tabela 13 – Predição das amostras da classe D no modelo DD-SIMCA da classe EF. Na coluna de predição os critérios pertencem e não pertencem corresponde a classe modelada (EF).

# Amostra Matriz	US/Padrão Fibrose	Predição
1	DC+	Não pertence
2	DC+	Não pertence
3	D+C	Não pertence
4	DC+	Não pertence
5	D+C	Pertence
6	D+C	Não pertence
7	DC	Pertence
8	DC+	Pertence
9	D+	Pertence
10	DC+	Não pertence
11	DC	Não pertence
12	DC+/E-	Não pertence
13	DC+	Não pertence
14	DC	Não pertence
15	DC+	Não pertence
16	DC+	Não pertence
17	DC	Não pertence
18	DC	Pertence
19	DC	Pertence
20	D+C	Pertence
21	D	Não pertence
22	D+C	Pertence
23	D+C	Não pertence
24	DC	Não pertence
25	DC	Não pertence
26	D+C	Não pertence
27	D	Pertence
28	DC	Não pertence
29	DC	Pertence
30	DC	Pertence

Fonte: A autora (2022)

No gráfico de predição das amostras da classe D da figura 33, é observado que 11 das 30 amostras foram classificadas como pertencendo a classe modelada (EF), e corresponde as amostras de cor verde que se encontram dentro da região delimitada pela linha verde. Já as 19 amostras que se encontram por fora da região modelada são categorizadas como não pertencendo a classe EF e se apresentam em cor vermelha. Esses resultados estão organizados na tabela 13, relacionando a classificação dada pela US e nota-se que das 11 amostras categorizadas como “DC” 5 foram classificadas como pertencendo a classe EF. Das 7 amostras do tipo “D+C” 3 foram classificadas como pertencendo a classe alvo. Esta mesma situação foi evidenciada na predição das amostras com o modelo GA-LDA, as amostras dentro de um mesmo tipo ou subclasse da classe D, por exemplo, DC, D+C, DC+; não são classificadas todas em uma determinada classe e sim distribuídas praticamente de forma equivalente nas classes modeladas, seja classe C, EF, pertencendo a classe EF ou não pertencendo.

Desta forma é evidenciada a complexidade para modelagem deste estágio intermediário da doença (classe D). Seria necessário utilizar outras abordagens de modelagem para desenvolver modelos que sejam mais sensíveis e permitam modelar a classe D. Isto será abordado nas perspectivas futuras.

5 CONCLUSÕES E PERSPECTIVAS FUTURAS

No presente trabalho foram desenvolvidos modelos metabonômicos capazes de discriminar, entre os pacientes esquistossomóticos, os casos extremos de fibrose periportal, isto é, os que apresentam FPP leve daqueles com FPP avançada, a partir de dados cromatográficos de amostras de soro. O DLLME foi o método de preparo de amostra mais eficiente na extração de informações da matriz.

O melhor modelo que classificou os dados cromatográficos foi o GA-LDA com 84% de sensibilidade, 83% de especificidade e 83% de exatidão. Seguido de DD-SIMCA com 92% de sensibilidade, 63% de especificidade e 75% de exatidão. A integração de técnicas analíticas e modelos metabonômicos na prática clínica pode fornecer uma ferramenta para o diagnóstico diferencial de fibrose periportal em pacientes com esquistossomose mansônica, especialmente em zonas rurais onde a US não está disponível em todos os centros médicos. Além da possibilidade de levar o controle da doença uma vez que o soro pode ser coletado repetidamente.

Apesar dos bons resultados aqui reportados, não foi possível a modelagem dos casos intermediários da doença, isto é a classe D. Para modelar esta classe é proposto como perspectiva futura avaliar uma fusão de dados utilizando informações dos sinais registrados pela técnica RMN obtido no trabalho de (LIMA RODRIGUES, 2020) em conjunto das informações dos picos cromatográficos apresentados neste trabalho, numa única modelagem. Desta forma os valores de sensibilidade, especificidade e exatidão para os modelos poderiam ser melhorados, uma vez que mais informações estão disponíveis para a modelagem das classes.

REFERÊNCIAS

ABDEL-MISIH, S. R. Z.; BLOOMSTON, M. Liver Anatomy. **Surgical Clinics of North America**, 2010. v. 90, n. 4, p. 643–653. Disponível em: <<http://dx.doi.org/10.1016/j.suc.2010.04.017>>.

ÅBERG, K. M.; ALM, E.; TORGRIP, R. J. O. The correspondence problem for metabonomics datasets. **Analytical and Bioanalytical Chemistry**, 2009. v. 394, n. 1, p. 151–162.

ADANIYA HIGA, B. Multivariate analysis applied to the supply of liquefied petroleum gas (LPG) in Peru. **Revista Industrial Data**, 2019. v. 22, p. 47–64.

ADENAN, M. N. H. *et al.* Screening Malaysian edible bird's nests for structural adulterants and geographical origin using Mid-Infrared – Attenuated Total Reflectance (MIR-ATR) spectroscopy combined with chemometric analysis by Data-Driven – Soft Independent Modelling of Class Ana. **Forensic Chemistry**, 2020. v. 17, p. 100197. Disponível em: <<https://doi.org/10.1016/j.forc.2019.100197>>.

ANASTASSIADES, M. *et al.* Fast and easy multiresidue method employing acetonitrile extraction/partitioning and. **Journal of AOAC International**, 2003. v. 86, p. 412–431. Disponível em: <<https://www.semanticscholar.org/paper/Fast-and-easy-multiresidue-method-employing-and-for-Anastassiades-Lehotay/f55b9b419a4f84392501cd90866c03ec01cd16cd>>.

ARAÚJO, Mário César Ugulino *et al.* The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, 2001. v. 57, n. 2, p. 65–73.

BARRETO, A. V. M. S. Letter to editor about “new index for the diagnosis of liver fibrosis in schistosomiasis mansoni”. **Arquivos de Gastroenterologia**, 2017. v. 56, n. 1, p. 108–109.

Estudo de biomarcadores para morbidade e acompanhamento pós terapêutico em pacientes com esquistossomose mansoni. **Stroke**, 2018. v. 32, n. 1, p. 139–146.

BARROS NETO, B. De; SCARMINIO, I. S.; BRUNS, R. E. 25 anos de quimiometria no Brasil.

Química Nova, 2006. v. 29, n. 6, p. 1401–1406.

BATISTA, A. *et al.* World Journal of Psychiatry. **World Journal of Psychiatry**, 2018. v. 6, n. 1, p. 43–53.

BÍBLIA. BÍBLIA. Nova Almei ed. [S.l.]: Sociedade Bíblica do Brasil, [s.d.].

BOLBOT, Ol. Imagens de liver fibrosis. 2021. Disponível em: <[https://www.shutterstock.com/pt/g/Olga+Bolbot?searchterm=liver fibrosis](https://www.shutterstock.com/pt/g/Olga+Bolbot?searchterm=liver%20fibrosis)>. Acesso em: 12 nov. 2021.

BRASIL, M. Da S. MANSONI Diretrizes Técnicas Diretrizes técnicas. [S.l.]: [s.n.], 2014.

BRITO, M. I. B. Da S.; SILVA, M. B. A.; QUININO, L. R. De M. Situação epidemiológica e controle da esquistossomose em Pernambuco : estudo descritivo , 2010-2016 *. 2020. v. 29, n. 2, p. 2010–2016.

CALDAS, S. S. *et al.* Principais técnicas de preparo de amostra para a determinação de resíduos de agrotóxicos em água por cromatografia líquida com detecção por arranjo de diodos e por espectrometria de massas. **Quimica Nova**, 2011. v. 34, n. 9, p. 1604–1617.

CAMPÊLO, J. De M. *et al.* Optimization of QuEChERS extraction for detection and quantification of 20 antidepressants in postmortem blood samples by LC-MS/MS. **Forensic Science International**, 2021. v. 319.

CANUTO, G. A. B. *et al.* **METABOLÔMICA: DEFINIÇÕES, ESTADO-DA-ARTE E APLICAÇÕES REPRESENTATIVAS.** **Quimica Nova**. Sociedade Brasileira de Quimica.

CAVIGLIA, G. P. *et al.* Liver fibrosis: The 2017 state of art. **Panminerva Medica**, 2017. v. 59, n. 4, p. 320–331.

COLLEY, D. G. *et al.* Human schistosomiasis. **The Lancet**, 2014. v. 383, n. 9936, p. 2253–2264. Disponível em: <[http://dx.doi.org/10.1016/S0140-6736\(13\)61949-2](http://dx.doi.org/10.1016/S0140-6736(13)61949-2)>.

COLLINS, C. H.; BRAGA, G. L.; BONATO, P. S. **Fundamentos de CROMATOGRAFIA**. Campinas - SP- Brasil: Editora da Unicamp, 2006.

COSTA FILHO, P. A. DA; POPPI, R. J. Genetic algorithm in chemistry. **Quimica Nova**, 1999. v. 22, n. 3, p. 405–411.

COSTA, T. DA. **METABONÔMICA APLICADA AO DIAGNÓSTICO E ESTADIAMENTO DE DOENÇAS HEPÁTICAS**. [S.I.]: UNIVERSIDADE FEDERAL DE PERNAMBUCO, 2016. Disponível em: <<https://repositorio.ufpe.br/handle/123456789/20263>>.

DAHAB, A. A.; SMITH, N. W. Drug-induced hepatotoxicity: Application of mass spectrometry based metabonomics. **Analytical Methods**, 2012. v. 4, n. 7, p. 1887–1902.

FERREIRA, Márcia Miguel Castro. ANÁLISE EXPLORATÓRIA DOS DADOS. In: LIMA, R. et al. (Org.). **QUIMIOMETRIA Conceitos, Métodos e Aplicações**. Campinas - SP- Brasil: Editora da Unicamp, 2015, p. 495.

FIDELIS, M. et al. Authentication of juices from antioxidant and chemical perspectives: A feasibility quality control study using chemometrics. **Food Control**, 2017. v. 73, p. 796–805. Disponível em: <<http://dx.doi.org/10.1016/j.foodcont.2016.09.043>>.

FRIEDMAN, S. L. Mechanisms of Hepatic Fibrogenesis. **Gastroenterology**, 2008. v. 134, n. 6, p. 1655–1669.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta**, 2005. v. 67, n. 4, p. 736–740.

GIKA, H. G. et al. Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. **Journal of Pharmaceutical and Biomedical Analysis**, 2014. v. 87, p. 12–25. Disponível em: <<http://dx.doi.org/10.1016/j.jpba.2013.06.032>>.

GODOY, G. D. M. M. **EMPREGO DA METABONÔMICA NO DIAGNÓSTICO DE DOENÇAS HEPÁTICAS EM AMOSTRAS DE URINA**. [S.I.]: UNIVERSIDADE

FEDERAL DE PERNAMBUCO, 2015. ISBN 9781450349185. Disponível em: <<https://repositorio.ufpe.br/handle/123456789/16247>>.

GOUVEIA, L. R. *et al.* Diagnosis of coinfection by schistosomiasis and viral hepatitis B or C using ¹H NMR-based metabonomics. **PLoS ONE**, 2017. v. 12, n. 8, p. 1–11.

Metabonômica Aplicada ao Diagnóstico Diferencial de Doenças Hepáticas. [S.l.]: UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO, 2017. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/29084/4/DISSERTAÇÃO_Liana_Ribeiro_Gouveia.pdf>.

GUERRA, S. *et al.* Liver diseases: what is known so far about the therapy with human amniotic membrane? **Cell and Tissue Banking**, 2016. v. 17, n. 4, p. 653–663.

HAMZENEJAD, A.; GHOUSHCHI, S. J.; BARADARAN, V. Clustering of Brain Tumor Based on Analysis of MRI Images Using Robust Principal Component Analysis (ROBPCA) Algorithm. **BioMed Research International**, 2021. v. 2021, n. 2008.

HEYENS, L. J. M. *et al.* Liver Fibrosis in Non-alcoholic Fatty Liver Disease : From Liver Biopsy to Non-invasive Biomarkers in Diagnosis and Treatment. 2021. v. 8, n. April, p. 1–20.

HUANG, Y. *et al.* **UHPLC-MS-Based Metabolomics Analysis Reveals the Process of Schistosomiasis in Mice.** **Frontiers in Microbiology**.

HUBERT, M.; ROUSSEEUW, P. J.; BRANDEN, K. VANDEN. ROBPCA: A new approach to robust principal component analysis. **Technometrics**, 2005. v. 47, n. 1, p. 64–79.

KOSMINSKY, E. Entendendo de uma vez por todas: Fisiopatologia da esquistossomose. **Eu Médico Residente Ensino**, 2020. Disponível em: <<https://www.eumedicoresidente.com.br/post/entendendo-de-uma-vez-por-todas-fisiopatologia-da-esquitossomose>>.

LAGO, E. M. *et al.* Antischistosomal agents: State of art and perspectives. **Future Medicinal Chemistry**, 2018. v. 10, n. 1, p. 89–120.

LANÇAS, F. M. A cromatografia líquida moderna e a espectrometria de massas: Finalmente “compatíveis”? **Scientia Chromatographica**, 2009. v. 5, n. 1, p. 27–46. Disponível em: <<http://doi.editoracubo.com.br/10.4322/sc.2013.005>>.

LI, Y. *et al.* 3E-LDA: Three Enhancements to Linear Discriminant Analysis. **ACM Transactions on Knowledge Discovery from Data**, 2021. v. 15, n. 4.

LIMA RODRIGUES, M. **DISCRIMINAÇÃO DE PACIENTES COM FIBROSE PERIPORTAL ESQUISTOSSOMÓTICA POR METABONÔMICA**. [S.l.]: Universidade Federal de Pernambuco, 2020.

LIU, Y. *et al.* A metabolic profiling analysis of symptomatic gout in human serum and urine using high performance liquid chromatography-diode array detector technique. **Clinica Chimica Acta**, 2011. v. 412, n. 23–24, p. 2132–2140.

LOPES, R. E. C. **Discriminação de madeiras similares por NIRS e PLS-DA considerando variações de temperatura e umidade**. [S.l.]: Universidade de Brasília, 2015. Disponível em: <http://repositorio.unb.br/bitstream/10482/19807/3/2015_RosylaneElaineCostaLopes.pdf>.

LOVERGNE, L. *et al.* Investigating optimum sample preparation for infrared spectroscopic serum diagnostics. **Analytical Methods**, 2015. v. 7, n. 17, p. 7140–7149.

LOYO, R. M. *et al.* Gas chromatography-mass spectrometry (GC/MS) reveals urine metabolites associated to light and heavy infections by *Schistosoma mansoni* in mice. **Parasitology International**, 2021. v. 80, n. August 2020.

LYRA, W. Da S. *et al.* Classificação periódica: um exemplo didático para ensinar análise de componentes principais. **Química Nova**, 2010. v. 33, n. 7, p. 1594–1597.

MADSEN, R.; LUNDSTEDT, T.; TRYGG, J. Chemometrics in metabolomics-A review in human disease diagnosis. **Analytica Chimica Acta**, 2010. v. 659, n. 1–2, p. 23–33.

MARIA SOARES TOJAL DE BARROS LIMA, L. *et al.* Valor preditivo de marcadores séricos

de fibrose hepática em pacientes portadores de hepatite crônica viral C Predictive value of serum markers of hepatic fibrosis in patients with chronic hepatitis C. 2008. p. 185–191. Disponível em: <<http://www.scielo.br/pdf/jbpm/v44n3/05.pdf>>.

MAZIVILA, S. J. *et al.* Detection of melamine and sucrose as adulterants in milk powder using near-infrared spectroscopy with DD-SIMCA as one-class classifier and MCR-ALS as a means to provide pure profiles of milk and of both adulterants with forensic evidence: A short communic. **Talanta**, 2020. v. 216, n. February, p. 120937. Disponível em: <<https://doi.org/10.1016/j.talanta.2020.120937>>.

MEDEIROS, T. B. *et al.* Correlation Between Platelet Count and Both Liver Fibrosis and Spleen Diameter in Patients With Schistosomiasis Mansoni. **Arquivos de Gastroenterologia**, 2014. v. 51, n. 1, p. 34–38.

MERCIECA, G. *et al.* Rapid and simple procedure for the determination of cathinones, amphetamine-like stimulants and other new psychoactive substances in blood and urine by GC–MS. **Journal of Pharmaceutical and Biomedical Analysis**, 2018. v. 149, p. 494–501. Disponível em: <<http://dx.doi.org/10.1016/j.jpba.2017.11.024>>.

MONZÓN, C. M. *et al.* Chemometric optimization and validation of a novel dispersive liquid–liquid microextraction-HPLC method for gliclazide, glibenclamide and glimepiride quantitation in serum samples. **Microchemical Journal**, 2016. v. 127, p. 113–119.

NEVES, M. D. G.; POPPI, R. J. Authentication and identification of adulterants in virgin coconut oil using ATR/FTIR in tandem with DD-SIMCA one class modeling. **Talanta**, 2020. v. 219, n. May.

NIAMEY WORKING GROUP. ULTRASOUND IN SCHISTOSOMIASIS: A Practical Guide to the Standardized Use of Ultrasonography for the Assessment of Schistosomiasis-related Morbidity. **World Health Organization technical report series**, 2000. p. 1–55.

PRESTES, Osmar D. *et al.* QuEChERS - Um método moderno de preparo de amostra para determinação multirresíduo de pesticidas em alimentos por métodos cromatográficos acoplados à espectrometria de massas. **Quimica Nova**, 2009. v. 32, n. 6, p. 1620–1634.

RAYCHAUDHURI, S. INTRODUCTION TO MONTE CARLO SIMULATION. (IEEE, Org.). Miami, FL, EUA: [s.n.], 2008. p. 91–100.

REZAEI, M. *et al.* Determination of organic compounds in water using dispersive liquid-liquid microextraction. **Journal of Chromatography A**, 2006. v. 1116, n. 1–2, p. 1–9.

RODIONOVA, O. Ye *et al.* Quantitative risk assessment in classification of drugs with identical API content. **Journal of Pharmaceutical and Biomedical Analysis**, 2014. v. 98, p. 186–192. Disponível em: <<http://dx.doi.org/10.1016/j.jpba.2014.05.033>>.

ROUSSEEUW, P. J.; HUBERT, M. Anomaly detection by robust statistics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 2018. v. 8, n. 2, p. 1–14.

SANTANA, D. F. B. *et al.* EXPERIMENTO DIDÁTICO DE QUIMIOMETRIA PARA CLASSIFICAÇÃO DE ÓLEOS VEGETAIS COMESTÍVEIS POR ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO COMBINADO COM ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS: UM TUTORIAL, PARTE V. **Quim. Nova**, 2020. v. 43, n. 3, p. 371–381.

SANTOS, I. G. De A. *et al.* Aspectos relacionados com a positividade para a esquistossomose : estudo transversal em área de baixa prevalência em Alagoas , 2020 *. 2021. v. 30, n. 2, p. 1–12.

SANTOS, J. C. a Utilização Da Metabonômica No Diagnóstico Da Infecção Crônica Pelos Vírus Das Hepatites B E C, E Na Avaliação Da Atividade Das Enzimas Hepáticas. [S.l.]: UNIVERSIDADE FEDERAL DE PERNAMBUCO, 2018. Disponível em: <<https://repositorio.ufpe.br/handle/123456789/29729>>.

SANTOS, L. O. DOS *et al.* The use of ANOVA-PCA and DD-SIMCA in the development of corn flour laboratory reference materials. **Food Chemistry**, 2022. v. 367, n. August 2021.

SCHEICH, T. EL *et al.* The WHO ultrasonography protocol for assessing hepatic morbidity due to *Schistosoma mansoni*. Acceptance and evolution over 12 years. **Parasitology Research**,

2014. v. 113, n. 11, p. 3915–3925.

SCHUPPAN, D.; AFDHAL H, N. Comparison of temperament traits among intercollegiate athletes and physical education majors. **Research Quarterly of the American Association for Health, Physical Education and Recreation**, 2008. v. 38, n. 4, p. 615–622.

SILVA, C. P. DA *et al.* Simple extraction of toxicologically relevant psychotropic compounds and metabolites from whole blood using mini-QuEChERS followed by UPLC–MS/MS analysis. **Biomedical Chromatography**, 2021. v. 35, n. 9.

SOARES DA SILVA BURATO, J. *et al.* Recent advances and trends in miniaturized sample preparation techniques. **Journal of Separation Science**, 2020. v. 43, n. 1, p. 202–225.

SOARES, S. F. C. *et al.* The successive projections algorithm. **TrAC - Trends in Analytical Chemistry**, 2013. v. 42, p. 84–98.

SOUZA, A. M. DE; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: UM tutorial, parte I. **Quimica Nova**, 2012. v. 35, n. 1, p. 223–229.

SOUZA ALMEIDA, L. **INVESTIGAÇÃO DE SIBILÂNCIA RECORRENTE E NÃO RECORRENTE EM LACTENTES USANDO METABONÔMICA BASEADA EM RMN DE 1H**. [S.l.]: [s.n.], 2021.

SOUZA, W. *et al.* Esquistossomose mansônica: aspectos gerais, imunologia, patogênese e história natural*. **Sociedade Brasileira de Clínica Médica**, 2011.

TADOKORO, T. Diagnosis and Therapeutic Management of Liver Fibrosis by MicroRNA. 2021.

TAHARA, Y.; SHIBATA, S. Circadian rhythms of liver physiology and disease: Experimental and clinical evidence. **Nature Reviews Gastroenterology and Hepatology**, 2016. v. 13, n. 4, p. 217–226. Disponível em: <<http://dx.doi.org/10.1038/nrgastro.2016.8>>.

THARWAT, A. Classification assessment methods. **Applied Computing and Informatics**,

2018. v. 17, n. 1, p. 168–192. Disponível em: <<https://doi.org/10.1016/j.aci.2018.08.003>>.

TSUKADA, S.; PARSONS, C. J.; RIPPE, R. A. Mechanisms of liver fibrosis. **Clinica Chimica Acta**, 2006. v. 364, n. 1–2, p. 33–60.

VALDERRAMA, L. *et al.* PROPOSTA EXPERIMENTAL DIDÁTICA PARA O ENSINO DE ANÁLISE DE COMPONENTES PRINCIPAIS. **Quimica Nova**, 2016. v. 39, n. 2, p. 245–249.

VUCKOVIC, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. **Analytical and Bioanalytical Chemistry**, 2012. v. 403, n. 6, p. 1523–1548.

WOLD, S. Chemometrics; what do we mean with it, and what do we want from it? **Chemometrics and Intelligent Laboratory Systems**, 1995. v. 30, n. 1, p. 109–115.

WORLEY, B.; POWERS, R. Multivariate Analysis in Metabolomics. **Current Metabolomics**, 2013. v. 1, n. 1, p. 92–107.

ZANELLA, R.; PRESTES, Osmar D.; *et al.* QuEChERS. In: NACIONAL, G.-G. E. (Org.). **Preparo de amostras para análise de compostos orgânicos**. Rio de Janeiro: [s.n.], 2015, p. 231–244.

ADAIME, M. B.; *et al.* Microextração líquido-líquido dispersiva. **Preparo de amostras para análise de compostos orgânicos**. Rio de Janeiro: [s.n.], 2015, p. 125–133.

ZONTOV, Y. V. *et al.* DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach. **Chemometrics and Intelligent Laboratory Systems**, 2017. v. 167, p. 23–28. Disponível em: <<http://dx.doi.org/10.1016/j.chemolab.2017.05.010>>.

ANEXO A – PARECER CONSUBSTANCIADO DO CEP

UFPE - HOSPITAL DAS
CLÍNICAS DA UNIVERSIDADE
FEDERAL DE PERNAMBUCO -
HC/UFPE



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS METABONÔMICOS PARA ESTADIAMENTO DE FIBROSE PERIPORTAL EM PACIENTES COM ESQUISTOSSOMOSE MANSONI USANDO CROMATOGRAFIA À LÍQUIDO E QUIMIOMETRÍA

Pesquisador: Julieth patricia Gonzalez Herrera

Área Temática:

Versão: 1

CAAE: 47667021.0.0000.8807

Instituição Proponente: Centro de Ciências Exatas e da Natureza

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.844.474

Apresentação do Projeto:

Projeto de mestrado da candidata supramencionada sob orientação do Prof. Dr. José Licarion Pinto Segundo Neto do departamento de Química fundamental que visa avaliar através de Modelos Metabonômicos o grau de fibrose periportal de 100 pacientes esquistossomóticos através de amostras de sangues de um banco de amostras de soro previamente aprovado em pesquisas anteriores do Serviço de Gastroenterologia do HC-UFPE, comparando os resultados com os achados do ultrassom (US) abdominal.

Objetivo da Pesquisa:

Geral:

Desenvolver e avaliar modelos metabonômicos, a partir de dados de cromatografia líquida de alta eficiência com detector por arranjo de Diodos e espectrometria de massas de amostras de soro, para estadiamento de fibrose periportal em pacientes com esquistossomose, tendo a US como padrão de referência.

Específicos:

- Realizar o preparo e as análises cromatográficas das amostras de soro.
- Desenvolver e avaliar modelos metabonômicos, a partir dos dados obtidos das análises cromatográficas, que permitam classificar os padrões de fibrose periportal.

Endereço: Av. Professor Moraes Rego, 1235, 3º andar do prédio principal

Bairro: Cidade Universitária **CEP:** 50.670-901

UF: PE **Município:** RECIFE

Telefone: (81)2126-3743

E-mail: cep.hcpe@ebserh.gov.br

**UFPE - HOSPITAL DAS
CLÍNICAS DA UNIVERSIDADE
FEDERAL DE PERNAMBUCO -
HC/UFPE**



Continuação do Parecer: 4.844.474

- Descrever os metabólitos associados à discriminação dos grupos.
- Comparar os resultados da US com os modelos de classe adquiridos pela metabonômica.

Avaliação dos Riscos e Benefícios:

Segundo os autores, os riscos possíveis são a manipulação do soro dos pacientes já coletados. Para amenizar os riscos a manipulação será realizada com os devidos equipamentos de proteção individual seguindo o protocolo de boas práticas de laboratório desde a manipulação do soro até o descarte dos materiais. Quanto aos benefícios, pretende-se desenvolver uma metodologia para construir modelos metabonômicos que possam ser utilizados como método de triagem e diagnóstico de pacientes com difícil acesso a equipamento de US em grandes centros.

Com a ascensão da pesquisa sobre o tema no futuro, se espera que a metabonômica esteja empregada nos diversos laboratórios e centros de saúde para que a população possa ter acesso ao diagnóstico de maneira ágil, eficaz e segura, principalmente em áreas de difícil acesso.

Comentários e Considerações sobre a Pesquisa:

Acreditamos que o desenvolvimento desses modelos metabonômicos possam no futuro, juntamente com os resultados do US abdominal nos pacientes esquistossomóticos com suspeita de fibrose hepática, contribuir para um diagnóstico mais precoce e, consequentemente, um melhor manejo terapêutico dos pacientes acometidos.

Considerações sobre os Termos de apresentação obrigatória:

vide conclusões ou pendências e lista de inadequações

Conclusões ou Pendências e Lista de Inadequações:

Aprovado

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJECTO_1721114.pdf	02/06/2021 12:56:07		Aceito
Projeto Detalhado / Brochura Investigador	Projeto_detalhado_JGH.doc	02/06/2021 12:54:48	Julieth patricia Gonzalez Herrera	Aceito

Endereço: Av. Professor Moraes Rego, 1235, 3º andar do prédio principal	
Bairro: Cidade Universitária	CEP: 50.670-901
UF: PE	Município: RECIFE
Telefone: (81)2126-3743	E-mail: cep.hcpe@ebserh.gov.br

**UFPE - HOSPITAL DAS
CLÍNICAS DA UNIVERSIDADE
FEDERAL DE PERNAMBUCO -
HC/UFPE**



Continuação do Parecer: 4.844.474

Outros	Carta_de_anuencia_carimbada_Andrea.pdf	02/06/2021 12:54:00	Julieth patricia Gonzalez Herrera	Aceito
Outros	SEI_SEDE_13642979_Carta_anuencia_HCEBSERH_UFPE.pdf	18/05/2021 12:25:50	Julieth patricia Gonzalez Herrera	Aceito
Outros	Termo_compromisso_e_confidencialida_de_Edmundo.pdf	18/05/2021 12:22:37	Julieth patricia Gonzalez Herrera	Aceito
Outros	CARTA_DE_ANUNCIA_DQF_Carimbada.pdf	18/05/2021 12:20:50	Julieth patricia Gonzalez Herrera	Aceito
Outros	declaracao_de_vinculo_mestrado.pdf	28/04/2021 13:43:14	Julieth patricia Gonzalez Herrera	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	Solicitacao_de_Dispenса_de_TCLE.docx	28/04/2021 13:35:09	Julieth patricia Gonzalez Herrera	Aceito
Folha de Rosto	folhaDeRosto.pdf	28/04/2021 13:33:21	Julieth patricia Gonzalez Herrera	Aceito
Outros	CV_Julieth.pdf	20/04/2021 17:27:22	Julieth patricia Gonzalez Herrera	Aceito
Outros	EDITAL_FACEPE_14_2019_INOVACAO_IAM_Resultado.pdf	20/04/2021 17:18:55	Julieth patricia Gonzalez Herrera	Aceito
Outros	TERMO_DE_CONFIDENCIALIDADE.docx	20/04/2021 17:17:13	Julieth patricia Gonzalez Herrera	Aceito
Outros	TERMO_DE_COMPROMISSO_DO_PESQUISADOR.docx	20/04/2021 17:17:03	Julieth patricia Gonzalez Herrera	Aceito
Outros	Formulario_EBSERH_sobre_o_projeto.docx	20/04/2021 17:15:57	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVmarcelino.pdf	20/04/2021 17:08:56	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVmilena.pdf	20/04/2021 17:08:34	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVanaluaciacoutinho.pdf	20/04/2021 17:08:04	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVedmundo.pdf	20/04/2021 17:07:36	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVgeorge.pdf	20/04/2021 17:07:07	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVricardo.pdf	20/04/2021 17:06:31	Julieth patricia Gonzalez Herrera	Aceito
Outros	CVlicarion.pdf	20/04/2021 17:05:53	Julieth patricia Gonzalez Herrera	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Endereço: Av. Professor Moraes Rego, 1235, 3º andar do prédio principal	CEP: 50.670-901
Bairro: Cidade Universitária	
UF: PE	Município: RECIFE
Telefone: (81)2126-3743	E-mail: cep.hope@ebsrh.gov.br

UFPE - HOSPITAL DAS
CLÍNICAS DA UNIVERSIDADE
FEDERAL DE PERNAMBUCO -
HC/UFPE



Continuação do Parecer: 4.844.474

Não

RECIFE, 13 de Julho de 2021

Assinado por:

Ana Caetano
(Coordenador(a))

Endereço: Av. Professor Moraes Rego, 1235, 3º andar do prédio principal
Bairro: Cidade Universitária CEP: 50.670-901
UF: PE Município: RECIFE
Telefone: (81)2126-3743 E-mail: cep.hcpe@ebserh.gov.br

ANEXO B – CROMATOGRAMAS 2D E 3D DAS AMOSTRAS CONTROLE E HEPATITE NOS 4 GRADIENTES ANALISADOS - DLLME

