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

HÉLDER VINÍCIUS CARNEIRO DA SILVA

ANÁLISE METABONÔMICA PARA DISCRIMINAÇÃO DE PACIENTES COM VARICOCELE QUANTO SUA FERTILIDADE USANDO DADOS CLÍNICOS E CROMATOGRAFIA A LÍQUIDO

Recife

2021

HÉLDER VINÍCIUS CARNEIRO DA SILVA

ANÁLISE METABONÔMICA PARA DISCRIMINAÇÃO DE PACIENTES COM VARICOCELE QUANTO SUA FERTILIDADE USANDO DADOS CLÍNICOS E CROMATOGRAFIA A LÍQUIDO

Este trabalho foi apresentado à Pós-Graduação em Química do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Química.

Área de Concentração: Química Analítica

Orientador: Prof. Dr. José Licarion Pinto Segundo Neto

Recife

2021

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB15-861

S586a Silva, Hélder Vinícius Carneiro da
Análise metabonômica para discriminação de pacientes com varicocele quanto sua fertilidade usando dados clínicos e cromatografia a líquido / Hélder Vinícius Carneiro da Silva. – 2021.
91 f.: il., fig., tab.

Orientador: José Licarion Pinto Segundo Neto.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Química, Recife, 2021.
Inclui referências e apêndices.

1. Química analítica. 2. Aprendizado de máquina. 3. Dados faltantes. 4. Fertilidade. I. Segundo Neto, José Licarion Pinto (orientador). II. Título.

543 CDD (23. ed.) UFPE- CCEN 2021 - 145

HÉLDER VINÍCIUS CARNEIRO DA SILVA

ANÁLISE METABONÔMICA PARA DISCRIMINAÇÃO DE PACIENTES COM VARICOCELE QUANTO SUA FERTILIDADE USANDO DADOS CLÍNICOS E CROMATOGRAFIA A LÍQUIDO

Dissertação apresentada ao Programa de Pós Graduação em Química da Universidade Federal de Pernambuco, Centro Acadêmico CCEN, como requisito para a obtenção do título de Mestre em Química. Área de concentração: Química Analítica.

Aprovado em: 13/08/2021.

BANCA EXAMINADORA:

Prof. Dr. José Licarion Pinto Segundo Neto
Departamento de Química Fundamental - UFPE
Orientador

Profa. Dra. Claudete Fernandes Pereira
Departamento de Química Fundamental - UFPE
Examinador Interno

Prof. Dr. Wellington Pinheiro Dos Santos
Departamento de Engenharia Biomédica - UFPE
Examinador Externo

AGRADECIMENTOS

Ao meu professor e orientador Licarion Pinto, por todo o conhecimento que me foi passado, sempre atento e com paciência para me tirar dúvidas, que contribuiu diretamente na minha formação acadêmica.

Ao Prof. Salvador Vilar Correia Lima, ao Ronmilson Marques e ao Filipe Tenório Lira Neto, por terem cedido as amotras.

À técnica Abene Ribeiro pela grande ajuda na operação dos equipamentos. Ao Prof Severino Junior e seu aluno Jefferson, pela disponibilidade para usar equipamentos do BSTR

Aos meus pais, Marli Batista e Joaquim Silva e à minha irmã, Tatiana Carneiro, pela paciência, compreensão, carinho e investimento.

À Adriane Seguins, Danielle Santana, Cecília Primavera, Elizângela Barbosa, Leonora Santana e Ramon Vinícius, pelas diversas alegrias e tristezas compartilhadas.

Aos meus amigos da Igreja Mangue, em especial aos meus amigos do grupo -Amém na mesa do bar-, pelas risadas, conversas, devocionais e pelo ombro na hora no choro.

Aos meus colegas e amigos do LabMeQ, em especial Julieth González, por ter sido minha companhia de labotarório nessa fase tão -quase- solitária de nossas vidas.

Aos meus amigos do DQF: Clara, Eric, Sarah, Lucas, Luisa, Matheus e Monica, por ter tornado essa etapa da minha vida bem menos difícil.

Ao PPG em Química do Departamento de Química Fundamental da UFPE, em especial aos professores que sempre estiveram muito presentes na minha formação. A Patrícia Rosa, sempre disponível e atenciosa na secretaria do curso.

Á Central Analítica do dQF, pela estrutura e equipamentos disponibilizados.

À CAPES pela bolsa concedida e a UFPE pelo suporte institucional.

Por fim, a todos que de alguma forma contribuíram para a realização desse trabalho.

“The opposite of war isn’t peace, it’s creation.” (LARSON, 1996).

RESUMO

A varicocele é uma doença caracterizada pela dilatação abnormal e tortuosidade na veia espermática, sendo a principal causa de infertilidade entre homens. A relação entre a varicocele e a infertilidade ainda não é bem definida. Além disso, o diagnóstico de infertilidade também é muito demorado, podendo chegar a 1 ano, e impreciso. Nesse âmbito, um estudo metabonômico pode oferecer mais pistas sobre a relação entre a varicocele, infertilidade e parâmetros seminais. A primeira etapa de um estudo metabonômico envolve o preparo do material biológico para as análises químicas, onde são necessárias técnicas que extraiam o máximo de informação mesmo em pequenas quantidades, como o QuEChERS miniaturizado e o DLLME. Assim, como a diversidade de metabólitos alcançadas pode ser muito vasta, o uso de técnicas cromatográficas de separação e análise tornam-se cruciais. Quando se pensa em um estudo metabonômico com objetivo de identificar, classificar e diagnosticar indivíduos, é importante que a instrumentação seja acessível para hospitais e clínicas de médio e pequeno porte, como o HPLC-DAD, que é de fácil operação, robusto, reprodutível e de baixo custo. As amostras de sêmen foram coletadas e divididas em três classes: controle (C), varicocele fértil (VF) e varicocele infértil (VI). O preparo de amostra deu-se por Microextração Líquido-líquido dispersiva (DLLME) e QuEChERS. A otimização do estudo metabonômico foi realizado em um HPLC-DAD utilizando acetonitrila e metanol. Definidas as melhores condições, as amostras foram submetidas a análise multivariada exploratória e classificatórias de dados através do software MATLAB, utilizando o somatório dos comprimentos de onda (de 200 a 400 nm) e a absorção apenas em 210 nm. Os dados clínicos foram colhidos de acordo com os parâmetros da OMS. A inserção dos dados faltantes foi efetuada com os algoritmos de SVD, KNN, BPCA e substituição pela média dos valores observados. A qualidade da inserção foi feita através do teste de Kolmogorov-Smirnov. As mesmas análises quimiométricas foram realizadas com os dados clínicos. A otimização revelou que o melhor preparo de amostra foi o DLLME e a melhor fase móvel foi o MeOH. A análise por PCA robusta não indicou a presença de amostras anômalas, mas também não foi capaz de promover uma separação visual entre as classes. Já a análise por PLS-DA conseguiu promover uma separação melhor entre as classes, com exatidão de 95% e 75 % para o somatório e em 210 nm, respectivamente. Já a análise por LDA com seleção de variáveis apresentou exatidão de 80% para ambas as classes. Na tentativa de melhorar os resultados, foram construídos modelos utilizando apenas as classes VF e VI, onde a LDA atingiu 100% de classificação correta em todos os dados avaliados. Em relação aos dados clínicos, o modelo de KNN foi o que apresentou os melhores resultados segundo o teste de Kolmogorov-Smirnov. Também não foi verificado aqui a presença de amostras anômalas. A análise exploratória indicou que os parâmetros seminais e hormonais são explicam as principais diferenças entre as classes VF e VI. Os modelos classificatórios obtiveram acurácia de 77,5% e 87,5% para o PLS e LDA, respectivamente. Assim, os estudos apresentados aqui, apesar de preliminares, apresentam elevado potencial para funcionarem como triagem para o diagnóstico de infertilidade em homens com varicocele.

Palavras-chaves: aprendizado de máquina; dados faltantes; fertilidade; metabolômica; quimiometria; varicocele.

ABSTRACT

Varicocele is a disease characterized by abnormal dilation and tortuosity in the spermatic vein, being the principal cause of infertility among men. The relationship between varicocele and infertility is not well defined. In addition, the diagnosis of infertility is also very time-consuming, reaching up to 1 year, and imprecise. In this context, a metabonomic study can offer more shreds of evidence about the relationship between varicocele, infertility, and seminal parameters. The first stage of a metabonomic lead-up involves the biological material preparation for chemical analyses, where techniques that extract the maximum amount of information even in small quantities are needed, such as the miniaturized QuEChERS and DLLME. Once the metabolite diversity achieved by these methods is wide, the use of chromatographic techniques for discrimination analysis becomes fundamental. A metabonomic study to identify, classify, and diagnosing individuals, the instrumentation must be accessible for medium and small hospitals and clinics, such as the HPLC-DAD, which is easy to operate, robust, reproducible, and Low cost. Semen samples were divided into three classes: control (C), fertile varicocele (VF), and infertile varicocele (VI). Sample preparation was made through DLLME and QuEChERS. The optimization of the metabonomic study was performed on an HPLC-DAD using ACN and MeOH. Once defined the best conditions, the samples were submitted to exploratory multivariate analysis and data classification using the MATLAB software, using the sum of wavelengths and absorption only at 210 nm. Clinical data were collected according to WHO parameters. The insertion of missing data was performed with the algorithms of SVD, KNN, BPCA, and replacement by the mean. Insertion quality was evaluated using the Kolmogorov-Smirnov test. The same chemometric analyzes were applied to clinical data. The optimization revealed that the best sample preparation was DLLME and, the best mobile phase was MeOH. Robust PCA analysis did not indicate anomalous samples presence, but there wasn't a good visual separation between the classes. PLS analysis promoted better discrimination between the classes, with an accuracy of 95% and 75% for the sum and at 210 nm, respectively. The analysis by LDA with variable selection showed an accuracy of 80% for both classes. Models were built using VF and VI classes to improve the results, where LDA reached 100% of correct classification in all evaluated data. In clinical data, the KNN model showed the best results according to the Kolmogorov-Smirnov test. ROBPCA didn't identify anomalous samples here either. Exploratory analysis indicated that seminal and hormonal parameters explain the main differences between classes VF and VI. The classification models obtained an accuracy of 77.5% and 87.5% for the PLS and LDA, respectively. The studies presented here, although preliminary, have a high potential to work as a screening method to diagnose infertility in men with varicocele.

Keywords: chemometrics; fertility; machine learning; metabolomics; missing Data; varicocele.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AG	Algoritmo Genético
BPCA	Análise de Componentes Principais Baeyseana
C	Controle
CG	Cromatografia Gasosa
CPE	Extração em Ponto de Nuvem
DA	Análise Discriminante
DAD	Detector de Arranjo de Diodos
DLLME	Microextração Líquido-Líquido Dispersiva
DR	Distância Robusta
ERO	Espécie Reativa do Oxigênio
FN	Falso Negativo
FP	Falso Positivo
FSH	Hormônio Folículo-estimulante
HLLE	Extração Líquido-Líquido Homogênea
HPLC	Cromatografia Líquida de Alta Eficiência
KNN	K Vizinhos mais próximos
KS	Kolmogorov-Smirnov
LDA	Análise Discriminante Linear
LH	Hormônio Luteinizante
MAR	Dados Faltantes Aleatórios
MCAR	Dados Faltantes Completamente Aleatórios
MCD	Determinante de Covariância Mínimo
MNAR	Dados Faltantes não Aleatórios
OMS	Organização Mundial de Saúde
PC	Componente Principal
PCA	Análise de Componentes Principais
PLS	Mínimos Quadrados Parciais
PP	Busca de Projeções

QuEChERS	Quick, Easy, Cheap, Effective, Rugged, Safe
ROBPCA	PCA Robusta
SHBG	Globulina Ligadora de Hormônios Sexuais
SVD	Decomposição de Valores Singulares
TSH	Hormônio Estimulante da Tireoide
VF	Varicocele Fértil
VI	Varicocele Infértil
VL	Variáveis Latentes
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

A infertilidade pode ser definida como a incapacidade de gerar filhos após um ano regular de sexo desprotegido. Segundo a OMS (Organização Mundial de Saúde), este problema atinge cerca de 48 milhões de casais e 186 milhões de indivíduos no mundo inteiro [2]. As doenças que ocorrem no trato genital feminino são responsáveis por cerca de 50% a 60% dos diagnósticos de infertilidade, enquanto as doenças do trato masculino são responsáveis por 40% a 50% [3].

Dentre as doenças do trato masculino, a varicocele é uma das principais causas de infertilidade entre homens, sendo responsável por cerca de 19% dos casos de infertilidade primária e de 45% nos casos de infertilidade secundária[1]. Na população de homens saudáveis, de 10% a 15% possuem algum grau de varicocele [4].

Apesar dos mais diversos estudos sobre as causas da varicocele e como ela afeta a fertilidade masculina, ainda não se há um diagnóstico preciso sobre os mecanismos que sofrem interferência da doença. Além disso, não se sabe também o porque de alguns homens possuírem varicocele e serem férteis, enquanto outros são inférteis [5].

Dentro deste cenário, é preciso se buscar técnicas capazes de inferir sobre o status de fertilidade ou ao menos realizar uma triagem quanto ao quadro do paciente sem que seja necessário aguardar um período de 1 ano para isso [6].

Em 2020 o nosso grupo de pesquisa realizou o estudo metabonômico para o dignóstico de infertilidade em homens com varicocole utilizando espectros de RMN de amostras de soro seminal [7]. Nesse contexto e com o objetivo de continuar os estudos metabonômicos da varicocele, propõe-se aqui um estudo metabonômico utilizando dados de cromatografia líquida de alta eficiência.

1.1 OBJETIVO GERAL

Desenvolver um modelo quimiométrico capaz de diferenciar entre pacientes com varicocele fértil e varicocele infértil utilizando dados clínicos e de cromatografia líquida de alta eficiência assistidas por técnicas quimiométricas.

[1] Infertilidade primária: quando não há gestação anterior; Infertilidade secundária: há gestação anterior.

1.1.1 **Objetivos específicos**

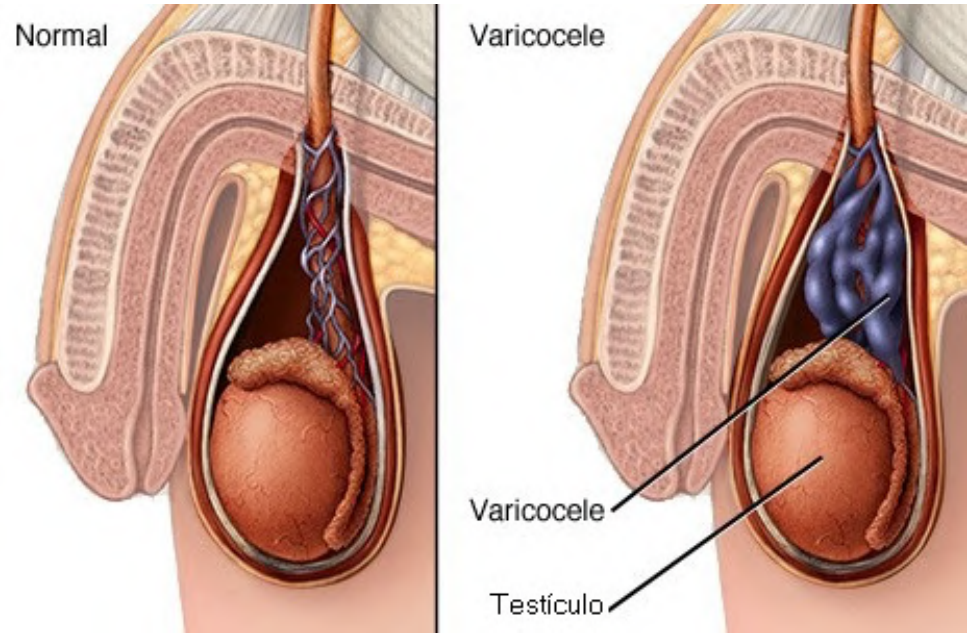
- Desenvolver um método analítico para a análise de soro de sêmen em HPLC-DAD;
- Realizar a predição e validação de dados faltantes no conjunto de dados clínicos;
- Realizar análise exploratória utilizando a PCA Robusta para a detecção de amostras anômalas;
- Construir modelos quimiométricos de classificação para a triagem de pacientes com varicocele quanto sua fertilidade usando dados clínicos e cromatográficos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 VARICOCELE

A varicocele é uma doença caracterizada pela dilatação abnormal e tortuosidade na veia espermática [8]. Está associada com a dor, atrofia testicular e redução das taxas de fertilidade. Na Figura 1 está a representação de um testículo saudável e um testículo com varicocele.

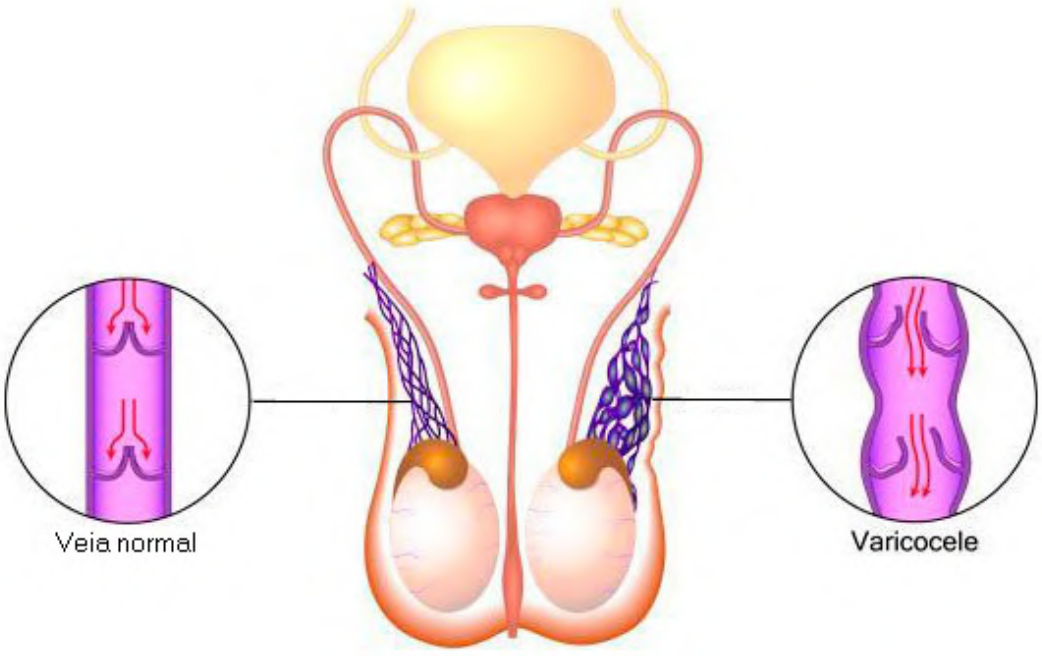
Figura 1 - Representação de um testículo saudável comparado ao um testículo com varicocele.



Fonte: Mayo Foundation for Medical Research and Education (2015). Modificado pelo autor.

Essa dilatação é causada pelo mal funcionamento do refluxo sanguíneo nas veias. Em um indivíduo saudável, o sistema de válvulas presente nas veias impede o retorno do sangue, sendo o mecanismo falho em um indivíduo com varicocele, ocasionando o acúmulo de sangue nos testículos e o consequente dilatamento das veias [8]. Na Figura 2 está representado o esquema de como acontece o refluxo.

Figura 2 - Representação de como ocorre o refluxo nas veias escrotais.



Fonte: Shutterstock (2018). Modificado pelo autor.

A varicocele é um problema comum na medicina reprodutiva, estando presente em 15% dos homens em todo o mundo. Sabe-se também que essa doença é uma das principais causadoras da infertilidade masculina, sendo responsável por 19% dos casos de infertilidade primária. A epidemiologia ainda não é bem entendida, sendo ainda necessários estudos em larga escala para determinar os fatores que contribuem para o seu desenvolvimento [9].

Um fator interessante sobre a varicocele é o seu agravamento com o passar dos anos. Foi comprovado que a incidência da doença aumenta cerca de 10% para cada década de vida e que 75% dos homens com mais de 90 anos possuem varicocele. Como está associada com a diminuição da produção de testosterona, a varicocele também pode ter correlação com o envelhecimento acelerado entre os homens [10].

Embora existam muitas possíveis causas, nenhuma única é capaz de descrever todos os casos, podendo muitas delas estar presentes em um único indivíduo. Além disso, identificar a causa específica de cada paciente pode não ser economicamente viável e pouco provavelmente traria algum benefício, considerando os tratamentos hoje já disponíveis. Contudo, é interessante conceituar as origens da varicocele, uma vez que isto pode guiar para o desenvolvimento de novas intervenções cirúrgicas que sejam mais direcionadas

[5, 11].

O diagnóstico da varicocele é feito em um exame de rotina, em um consultório quente, sendo o paciente avaliado de pé ou deitado. O uso da manobra de Valsalva pode ajudar na visualização das veias dilatadas, caso elas não cheguem a ser palpáveis. A varicocele é dividida em três graus, de acordo com o exame físico proposto por Dubin and Amelar [12, 13].

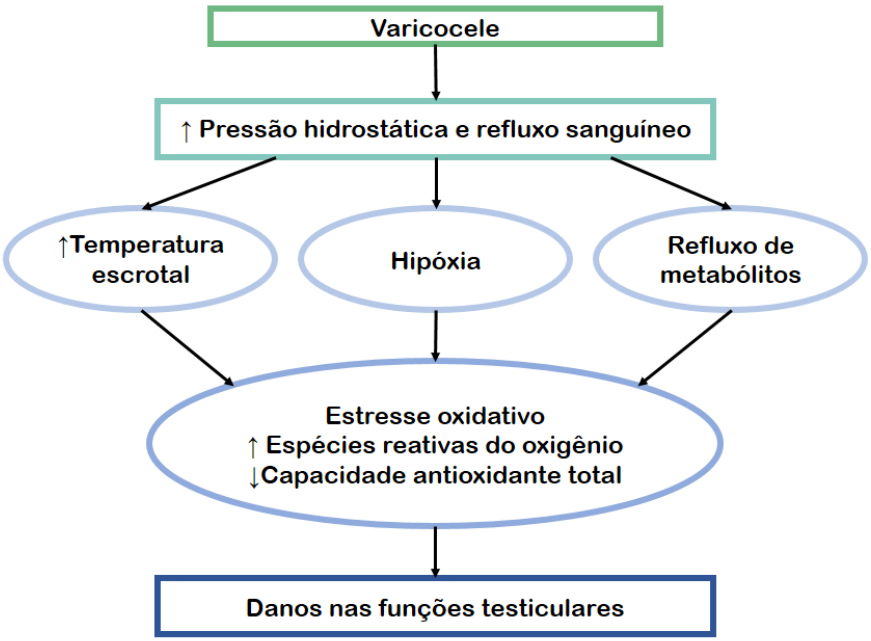
- **Grau 1** - Palpável quando o paciente está de pé e fazendo a manobra de Valsalva.
- **Grau 2** - Palpável quando o paciente está de pé, sem realizar a manobra de Valsalva.
- **Grau 3** - Palpável e visível pelo saco escrotal quando o paciente está de pé.

2.1.1 Varicocele e Infertilidade

A relação entre a varicocele e a infertilidade ainda não é bem definida. Acredita-se que o agravamento da doença ao ponto de causar efeitos negativos nos testículos e nos parâmetros seminais pode ser ocasionado por diversos mecanismos que ocorrem de forma simultânea. Dentro dessa cadeia de complexos mecanismos, o estresse oxidativo pode ser a principal causa da disfunção [14].

O estresse oxidativo pode ser definido como a produção elevada de espécies reativas do oxigênio (ERO). Como exemplos, podem ser citados o peróxido de hidrogênio e radicais livres que contém um elétron desemparelhado. Em concentrações normais, a presença dos EROs é essencial para produção de espermatozoides saudáveis. Entretanto, o excesso dessas espécies já é uma conhecida causa da infertilidade masculina. O fenômeno bioquímico pode ser resultado de vários mecanismos compensatórios que trabalham para assegurar a produção dos espematozóides, mas que, a nível molecular, levam a formação de muitos radicais livres [15]. Como apresentado na Figura 3, o estresse oxidativo pode ser causado por três condições principais: A hipertermia escrotal, a hipóxia e o refluxo de metabólitos.

Figura 3 - Principais fatores que levam ao estresse oxidativo em pacientes com varicocele.



Fonte: Jensen et al.(2017) [14]. Modificado pelo autor.

Alguns estudos já correlacionaram a presença da varicocele com o aumento da temperatura escrotal, denominada também como hipertermia [16–18]. Com o refluxo sanguíneo, a produção de espermatozoides é afetada pois ela costuma ocorrer idealmente a cerca de 2ºC abaixo da temperatura corporal [19]. Os mecanismos moleculares que levam os compromentimento dos parâmetros seminais estão relacionados principalmente com a diminuição da síntese de proteínas e de enzimas específicas [20].

Um segundo aspecto a ser avaliado é a pressão sanguínea nos testículos, que é muito menor quando comparada com outras regiões do corpo. Portanto, é de se esperar que qualquer mínima alteração da pressão seria significativa no microambiente testicular. O aumento da pressão venosa leva uma reação contrária na pressão arterial, que diminui em um efeito compensatório, a fim de manter a homeostase em relação à pressão intratesticular. Tal efeito leva a diminuição no fornecimento de oxigênio e nutrientes, conhecido como hipóxia. Um dos principais problemas que podem ser gerados nesse cenário é a apoptase, a morte programada de células [21,22]. A resposta à hipóxia induzida por varicocele promove a apoptose das células germinativas, contribuindo assim para a infertilidade masculina [23].

Outro fator contribuinte para o estresse oxidativo é o refluxo de metabólitos tóxicos

vindos dos rins e das glândulas suprarrenais, como a ureia, prostaglandinas E, prostaglandina F2-alfa e epinefrina. Esses metabólitos são responsáveis por contribuir no estresse oxidativo em outras culturas celulares em testes realizados *in vitro*. Entretanto, apesar de estarem presentes em elevados níveis em pacientes com varicocele, não há estudos que comprovem a relação com a infertilidade, sendo os poucos realizados em animais muito contraditórios [24].

Camoglio e colaboradores estudaram os efeitos do refluxo de metabólitos renais e adrenais na função testicular utilizando ratos. Os dados obtidos indicaram comprometimento nos níveis hormonais e nos parâmetros seminais, apoiando a hipótese de que os metabólitos citados aumentam o dano testicular induzido por varicocele [25]. Entretanto, outros estudos indicam que não há diferença significativa entre os níveis desses metabólitos no sangue venoso e arterial, o que indica que não ocorre o refluxo. Por isso, esse fator precisa de estudos mais aprofundados [26, 27].

Além de problemática em definir as causas da infertilidade, o diagnóstico também é muito demorado e impreciso. Hoje, os testes de infertilidade são recomendados quando a gravidez não ocorre após doze meses de sexo sem o uso de métodos contraceptivos. São analisados então os parâmetros seminais e físicos, sendo o diagnóstico dado por um urologista ou profissional na área de reprodução humana. Entretanto, a análise individualizada dos parâmetros seminais pode não ser uma boa preditora da infertilidade, o que torna o diagnóstico bem mais desafiador [6, 28].

Como foi exposto, a relação entre a varicocele, infertilidade e parâmetros seminais não é totalmente compreendida. A resposta pode estar em uma análise holística dos diversos pontos que contribuem para o desenvolvimento desta disfunção. Nesse âmbito, um estudo metabonômico pode oferecer mais pistas para a elucidação dessa relação.

2.2 METABONÔMICA

As ciências ômicas podem ser definidas como o estudo, identificação e caracterização de moléculas biológicas envolvidas em um processo bioquímico a fim de inferir sobre a dinâmica e estrutura desse processo em uma célula, tecido ou organismo. Exemplos de ciências ômicas são a genômica, transcriptômica, proteômica e a metabolômica, que estudam o genoma, transcriptoma, proteoma e o metaboloma, respectivamente [29].

Como já citado, a metabonômica, a ciência ômica utilizada neste trabalho, é o

estudo do metaboloma, que representa o conjunto de metabólitos em um ambiente bioquímico específico, que por sua vez são os produtos moleculares finais de processos celulares. O principal objetivo nos estudos metabolômicos é identificar metabólitos ou marcadores químicos que diferenciem processos ou condições características [30]. Pode ser categorizada em duas abordagens: a metabolômica *targeted*, onde os metabólitos já são conhecidos e o objetivo é a quantificação, e a metabolômica *untargeted* ou metabonômica, onde o objetivo é determinar o maior número de metabólitos possíveis, envolvendo principalmente a identificação e reconhecimento de padrão [31].

O fluxograma de trabalho de um estudo metabonômico deve seguir uma padronização definida pela Sociedade de Metabolômica (do inglês, *Metabolomics Society*), onde inicialmente devem ser determinados o objetivo do estudo, o número de amostras e a abordagem a ser utilizada. Após definido o planejamento experimental, segue-se então com a coleta e preparo das amostras, análise instrumental, processamento e análise estatística dos dados, identificação dos metabólitos e interpretação biológica final [32].

2.2.1 **Preparo de amostras e Técnicas miniaturizadas de extração**

A primeira etapa de um estudo metabonômico envolve o preparo do material biológico para as análises químicas. Logo, a escolha do tipo de preparo é fundamental, uma vez que a composição dos metabolitos a serem observados pode ser afetada pelo processo e interferir na interpretação biológica final. Para ser considerado um método ideal, o preparo deve: ser não seletivo, a fim de garantir que o maior número possível de metabólitos sejam alcançados; ser simples e rápido, evitando a perda de material; ser reprodutível e garantir que os metabólitos sejam os mesmos no momento da extração e no momento da análise [33].

O preparo de amostra pode ir desde uma diluição usando tampões e solventes deuterados a extrações com solventes orgânicos. Em análises cromatográficas é mais comum a realização de uma extração com solvente orgânico para adequar a amostra à análise e reduzir a presença de interferentes que possam danificar o equipamento, por exemplo macromoléculas [34].

Geralmente, o volume de amostras biológicas disponível é bastante reduzido, sendo necessárias técnicas que extraiam o máximo de informação mesmo em pequenas quantidades. Neste cenário, o uso de técnicas miniaturizadas de extração é uma alternativa viável

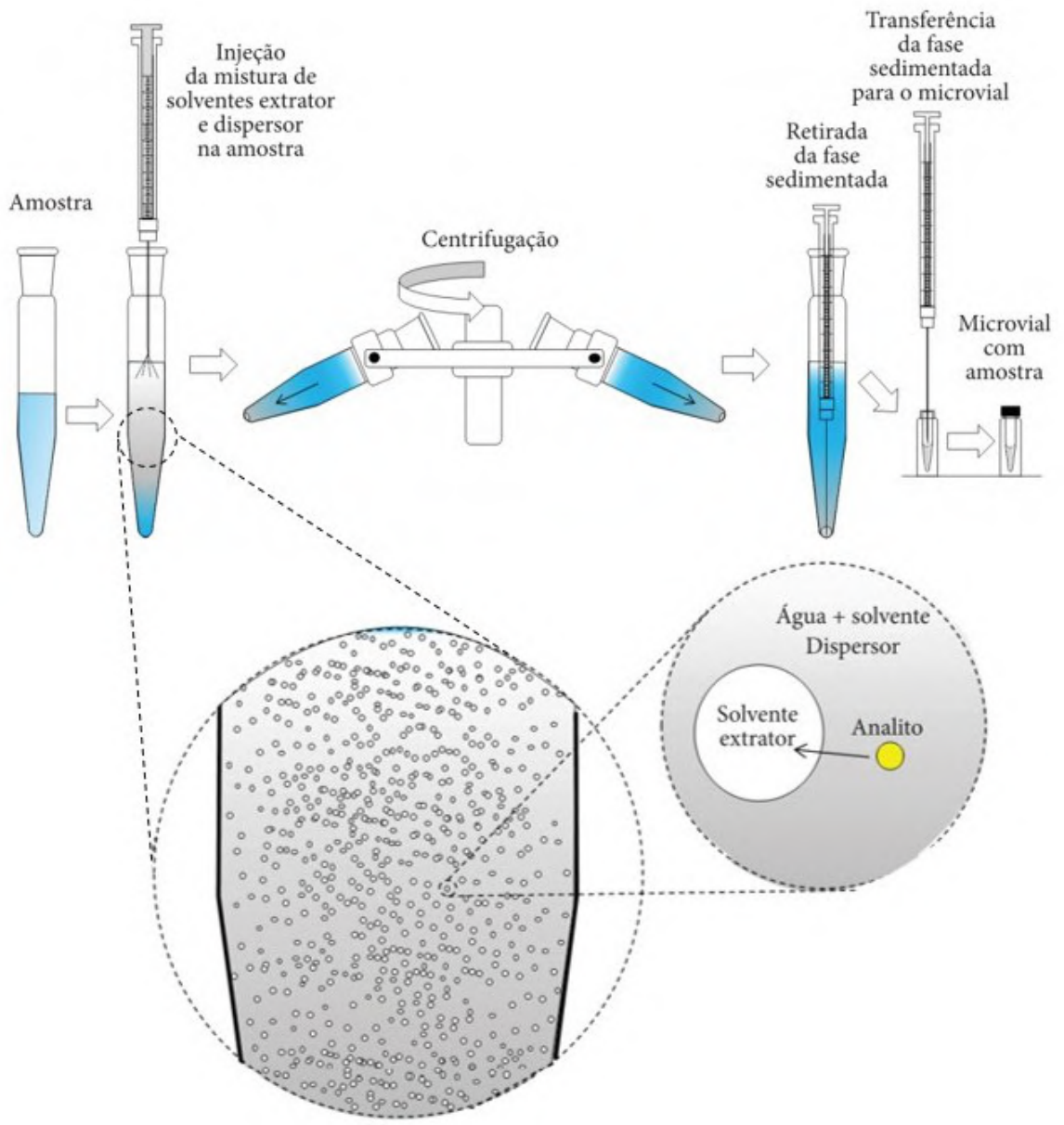
para contornar esse tipo de problema. As técnicas miniaturizadas, além de proporcionarem o manejo mais adequado para amostras biológicas, ainda permitem uma abordagem mais sustentável, onde a demanda de solventes e materiais também é reduzida. Ademais, essas técnicas permitem análises mais simples e diretas, aumentando a eficiência da extração [35].

Dois métodos foram avaliados, um que se baseia na total miscibilidade do solvente extrator e amostra com posterior partição com fase sólida (QuEChERS) e um que se baseia na imiscibilidade entre solvente extrator e amostra (DLLME). O primeiro método a ser citado é o QuEChERS, tipo de preparo de amostra recebe o nome através da sigla do inglês para *Quick, Easy, Cheap, Effective, Rugged, Safe - QuEChERS*, que significa rápido, fácil, barato, eficaz, robusto e seguro, respectivamente . O método foi desenvolvido por Anastassiades *et al.* [36] em 2003 para a extração de pesticidas em produtos agrícolas. Desde então, o modelo de extração vem sofrendo modificações, inclusive na sua miniaturização [37].

Alguns estudos metabonômicos já utilizaram o QuEChERS ou versões modificadas para extrair metabólitos de amostras biológicas, revelando que a técnica tem potencial para acessar um variado número de informações por sua extração eficiente e com possibilidade de ser não seletiva [38–40]. Recentemente, Casado *et al.* desenvolveram uma versão miniaturizada da técnica para a quantificação de fenóis em produtos alimentícios para bebês. A abordagem proposta conseguiu extrair as informações necessárias e utilizou bem menos materiais [41].

Outra técnica miniaturizada de extração é a Microextração Líquido-Líquido Dispersiva (DLLME, do inglês *Dispersive Liquid-Liquid Microextraction*), proposta inicialmente por Rezaee e colaboradores [42]. A DLLME é caracterizada pela utilização de dois tipos de solventes. O primeiro é o solvente extrator (fase orgânica), que irá concentrar os analitos desejados e estará em menor quantidade. O segundo, o solvente dispersor, que estará presente em maior quantidade e deve ser miscível tanto no solvente extrator como na amostra (fase aquosa). O sistema de extração é definido baseado no sistema ternário de solventes como a Extração Líquido-Líquido Homogênea (HLLE, do inglês *Homogeneus Liquid-Liquid Extraction*) e a Extração em Ponto de Nuvem (CPE, do inglês *Cloud Point Extraction*) [43]. Na Figura 4 está um esquema que apresenta as etapas do processo.

Figura 4 – Esquema do método DLLME.



Fonte: Martins et al.(2012) Modificado pelo autor [43]

Conforme ilustrado na Figura 4, com o auxílio de uma seringa, a mistura de solventes é colocada em contato com a amostra, onde ocorre a mistura das fases. Assim, é gerada uma perturbação vigorosa no sistema, tornando-se a mistura turva. A turbidez observada é resultado da presença de microbolhas de solvente extrator dispersas por todo a amostra. Como resultado, a área superficial disponível para a transferência de analitos é infinitamente melhorada quando comparada a uma extração líquido-líquido tradicional, sendo o estado de equilíbrio atingido rapidamente. Após a sedimentação acelerada por

centrifugaçãodo solvente extrator, este pode ser coletado e analisado [43].

O uso da DLLME pode ser encontrado em alguns estudos metabonômicos disponíveis na literatura. Por exemplo, Zhao *et al.* determinaram a presença de neurotransmissores e metabólitos presentes em amostras de cérebro de ratos que sofriam de Parkinson [44]. Este tipo de preparo de amostra também foi utilizado em um estudo para a determinação de biomarcadores de diabetes na urina [45]. No campo dos estudos urológicos, Huang *et al.* conseguiram extrair e determinar metabólitos na urina que indicam a presença de câncer de próstata [46].

A diversidade de metabólitos alcançadas por eficientes métodos de extração pode ser muito vasta. Por exemplo, no plasma humano podem ocorrer cerca de 4000 metabólitos em nove diferentes ordem de grandeza. Nesse cenário, o uso de técnicas cromatográficas de separação e análise tornam-se cruciais para o estudo metabonômico [47].

2.2.2 Técnicas Cromatográficas - aquisição de dados para estudos metabonômicos

A evolução dos estudos metabonômicos sempre esteve atrelada ao desenvolvimento de técnicas cromatográficas cada vez mais precisas. O avanço tecnológico no empacotamento das colunas, processamento de dados e detecção, permitiram o aumento do acesso a vários tipos de metabólitos com maior resolução cromatográfica. Hoje, a combinação de tecnologias de separação e detecção utilizadas na metabonômica é muito ampla, sendo as mais comuns o HPLC-MS, CE-MS e CG-MS [47].

O uso do detector de massas é extensamente aplicado em estudos metabonômicos por oferecer a possibilidade de determinação das rotas metabólitas. Conduto, é importante frisar que este tipo de técnica, quando associada ao CG, exige que as moléculas a serem analisadas sejam voláteis e termicamente estáveis, o que demanda um preparo de amostra bem mais delicado, como a necessidade de derivatização. Já quando utilizado com a cromatografia líquida, o sistema apresenta pouca reprodutibilidade e diversas dificuldades de operação e manutenção [48]. Assim, quando se pensa em um estudo metabonômico com objetivo de identificar, classificar e diagnosticar indivíduos, é importante que a instrumentação seja acessível para hospitais e clínicas de médio e pequeno porte. Logo, o HPLC-DAD apresenta-se como uma boa alternativa para um situação como esta. O HPLC-DAD é de fácil operação, robusto, reprodutível e de baixo custo, chegando a cus-

tar cerca de 1/10 do valor de um LC/MS. Além destas vantagens, a técnica é bem menos sensível à interferência da matriz, como a presença de carboidratos, ácidos graxos e colesterol [48, 49]. Muitos estudos metabonômicos classificatórios já foram feitos utilizando o HPLC-DAD como método de separação e detecção de metabólitos [50, 51] , principalmente em alimentos. Como exemplo, podem ser citados os estudos pra a classificação de óleo de oliva [52], vinho [53] e café [54].

Considerando-se os aspectos citados, pode-se afirmar que o estudo metabonômico é um importante passo para a investigação de um fenômeno bioquímico. Porém, com o avanço de tecnologias citadas capazes de monitorar milhares de metabólitos em uma única análise, extrair informações relevantes desses complexos e extensos conjuntos de dados torna-se uma tarefa crucial. A quimiometria permite a utilização de variados métodos para melhor acessar e compreender o comportamentos de metabólitos em diferentes ambientes químicos [55].

2.3 QUIMIOMETRIA - ANÁLISE MULTIVARIADA DE DADOS

Apesar de serem comuns na rotina de análises químicas, os métodos univariados passam a ser limitados na presença de inúmeras análises. No contexto metabonômico, um conjunto de dados $\mathbf{X}_{i\times j}$ pode facilmente ter dimensões $j \gg i$, sendo necessário então o uso de abordagens multivariadas capazes de explorar e detectar informações importantes. Os métodos multivariados de reconhecimento de padrão podem ser divididos em dois grupos: os não supervisionados, que buscam a redução da dimensionalidade e análise exploratória, e os supervisionados, dedicados à classificação e predição.

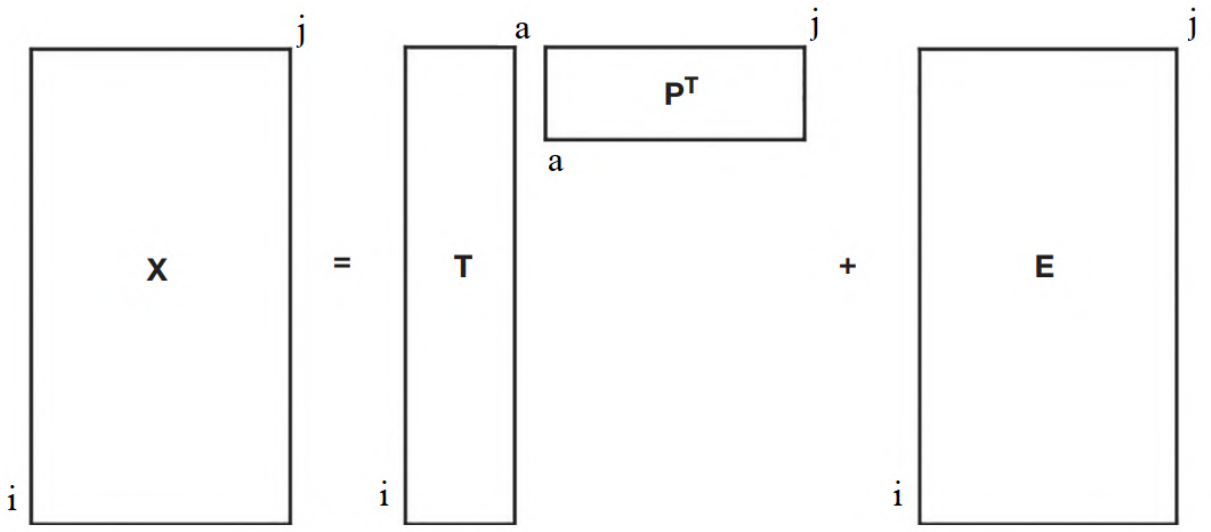
2.3.1 **Análise Exploratória de Dados**

A análise por componentes principais (PCA, do inglês - *Principal Component Analysis*) é o principal método de análise exploratória de dados multivariados que pode ser utilizado para a redução da dimensionalidade, detecção de amostras anômalas, agrupamento gráfico e entre outras aplicabilidades. Esse tipo de análise exploratória é pré-requisito para outras análises mais específicas, como a calibração multivariada e análise classificatória, sendo assim a primeira etapa na análise multivariada de dados [56]. O cálculo da PCA está indicado na Equação 1. Na Figura 5 está representada de forma

gráfica as matrizes envolvidas nos cálculos.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

Figura 5 - Representação matricial da equação 1.



Fonte: (Geladi *et al.*, 2020.) [57].Modificado pelo autor.

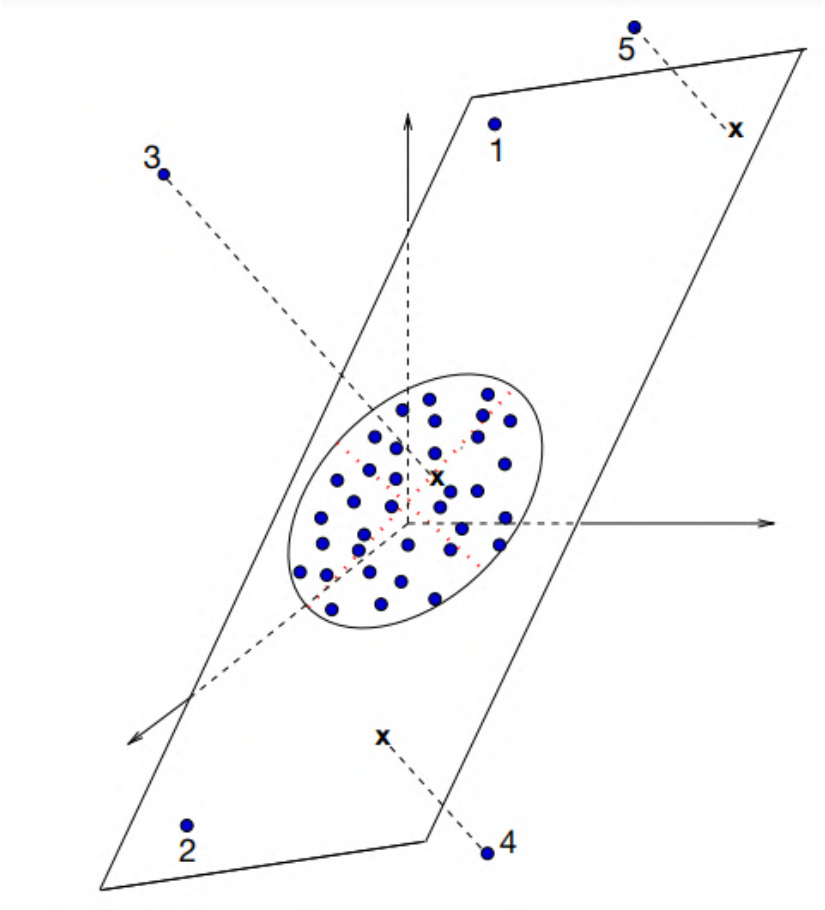
Como indicado na Figura 5, a matriz $\mathbf{X}_{i\times j}$ representa os dados originais, pode ser decomposta em uma matriz $\mathbf{T}_{i\times a}$ de escores, onde a correponde ao número de componentes calculadas). A matriz $\mathbf{X}_{i\times j}$ é multiplicada pela transposta da matriz $\mathbf{P}_{a\times j}$, que contêm os pesos de cada variável. Os resíduos não explicados pela transformação estão contidos na matriz $\mathbf{E}_{i\times j}$ com as mesmas dimensões da matriz original.

O cálculo da primeira componente principal (PC1) se dá pela projeção de uma linha no espaço multidimensional, onde a soma dos quadrados dos objetos projetados naquela direção são maximizadas, ou seja, a PC1 corresponde a direção onde há a maior variação dos dados. Os escores são a representação das amostras no espaço das novas variáveis denominadas de PC, enquanto os pesos contém a informação das variáveis usadas para avaliar similaridades e diferenças entre as amostras em cada respectiva PC. Assim, quanto maior o valor do peso de uma variável, maior a sua contribuição para a variabilidade entre os dados. O mesmo ocorre para a segunda componente principal (PC2), que extrai mais informações da matriz X não explicadas pela PC1. Da mesma forma se calcula a terceira e quarta PC, de forma que todas as PCs são ortogonais entre si. O número ideal de PCs depende da complexidade das amostras analisadas, sendo que um modelo com

mais de 10 PCs muito complexo e pode indicar que a análise por PCA não é a mais adequada para o caso, sendo necessárias abordagens mais específicas, como os algoritmos de classificação. [57]

Um dos problemas apresentados por essa versão clássica da PCA é a sua sensibilidade a presença de amostras anômalas, que diferem da maioria das observações encontradas no espaço multidimensional. Como pode ser observado na Figura 6, existem diferentes tipos de amostras anômalas que podem influenciar na rotação do modelo. A PCA está representada no plano n-dimensional [58].

Figura 6 - Tipos de amostras anômalas encontrados em um conjunto de dados



Fonte: Hubert, M., Rousseeuw, P. J., Vanden Branden, K. (2020). [59]

Nota-se que maioria das amostras se encontram no centro das origens das PCs. A observação 1 e 2, apesar de afastadas do centro (resíduo elevado), não influenciam na rotação do modelo pois se encontram no mesmo plano. A amostra 3 possui elevada distância ortogonal do plano da PC, que não pode ser percebido no gráfico dos resíduos devido a sua projeção do plano. As observações 4 e 5 possuem grande distância do centro

da origem e ortogonal, o que torna o modelo mais propício de sofrer rotações na tentativa que explicar essas amostras. Para melhor identificar a presença e minimizar o efeito de amostras anômalas na rotação do modelo PCA, foi proposta a abordagem da PCA robusta (ROBPCA). [59].

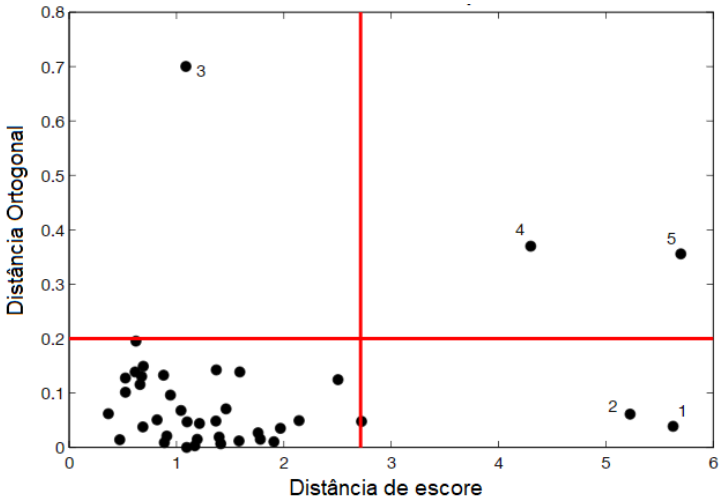
Dentre as estratégias utilizadas para o cálculo da ROBPCA, o mais comum é o determinante de covariância mínimo, o MCD (do inglês, *Minimum Covariance Determinant*). O MCD trabalha de forma a encontrar um subconjunto de amostras que minimizem o determinante da matriz de covariância. Determinado o subconjunto, é calculada a distância robusta (DR) de cada amostra (x_i) de acordo com a equação 2, em que $\hat{\mu_0}$ é a média das distâncias das amostras.

$$DR(\boldsymbol{x_i}) = \sqrt{(x_i - \hat{\mu_0})^t \mathbf{X^T X} (x_i - \hat{\mu_0})} \tag{2}$$

Considerando a propriedade das matrizes, o conjunto de dados $\mathbf{X}$ deve ter mais linhas do que colunas, evitando assim que a matriz de covariância tenha o determinante nulo. Deste modo, o uso do MCD torna-se limitado, sendo necessário então uma redução de dimensionalidade para conjunto de dados maiores. Inicialmente, esta redução era feita por decomposição de valores singulares, no entanto, algoritmos mais avançados utilizam o *project pursuit* (PP, busca de projeções). Este algoritmo busca inicialmente uma projeção onde a distribuição de dados seja a menos gaussiana possível. Uma vez que a distribuição de ruídos apresenta um caráter normal, projeções que se afastam ao máximo desse comportamento são as que mais explicam a variação entre as amostras.

A ROBPCA é então calculada utilizando apenas os pontos onde a distância ortogonal não é muito grande definidos pelo PP e MCD, resultando em um novo espaço k-dimensional. Como ferramenta de diagnóstico, pode ser utilizado o mapa de *outliers*, que consiste em verificar a distância ortogonal pela distância de escores. Um exemplo está representado na Figura 7, onde é possível observar as amostras em 4 quadrantes distintos. Os valores de corte para a distância ortogonal e de escores estão indicados nas linhas vermelhas.

Figura 7 - Mapa de *outliers*.

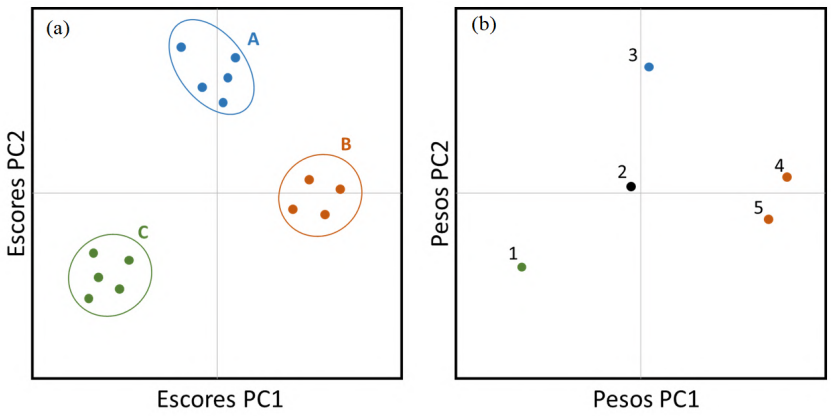


Fonte: Hubert, M., Rousseeuw, P. J., Vanden Branden, K. (2020). [58], modificado pelo autor.

Como pode ser observado, as amostras mais homogêneas se encontram no terceiro quadrante. Já no segundo e quarto quadrante, temos amostras com elevada distância ortogonal e de escore, respectivamente. Por fim, as amostras inseridas no primeiro quadrante, indicam elevada distância ortogonal e de escore ao mesmo tempo, sendo potencialmente anômalas e podem dificultar na construção do modelo [58].

Definido o conjunto de dados, pode-se então dar continuidade à análise exploratória. Para isso, é construído um gráfico que relaciona os escores e pesos em cada uma das PCs calculadas. Um exemplo de gráficos de escores e pesos está indicado na Figura 8.

Figura 8 - Exemplo de gráfico de escores(a) e pesos(b) obtidos no cálculo da PCA.



Fonte:O autor (2021)

A interpretação do gráfico pode ser feita da seguinte forma: Amostras próximas entre si indica a presença de similaridades, sendo possivelmente agrupas em classes, como

as classes A, B e C demonstradas no gráfico. Já em relação aos pesos, pode-se inferir sobre a contribuição de cada variável para a construção da PC. Por exemplo, as variáveis 4 e 5 possuem influência na separação das amostras em PC1, mas têm baixa influência em PC2. O comportamento exatamente contrário é encontrado na variável 3, com alta influência em PC2 e baixa em PC1. Já a variável 3 apresenta influência em ambas as PC, ao contrário da variável 2, que apresenta baixa influência na separação das amostras. Além disso, pode-se averiguar sobre o comportamento das variáveis dentro de cada classe. Por exemplo, pode-se esperar que as variáveis 4 e 5 possuam um maior valor numérico no grupo B, assim como 3 para o grupo A e 1 para o grupo C [60].

2.3.2 **Análise Classificatória de Dados**

Dentre os métodos supervisionados de dados, podem ser destacados o PLS-DA (do inglês, *Partial Least Square - Discriminant Analysis*) e a LDA (do inglês, *Linear Discriminant Analysis*), este último comumente acoplada a seleção de variáveis.

O PLS-DA é um algoritmo muito utilizado como ferramenta quimiométrica para otimizar a discriminação entre dois ou mais conjuntos de dados, através da análise conjunta entre duas matrizes, uma contendo os dados originais ($\mathbf{X}$) e outra contendo a indicação da classe ($\mathbf{Y}$) [61]. O PLS-DA trabalha de forma a maximizar as covariâncias entre as duas matrizes citadas através da criação de um subespaço linear. Esse subespaço permite a predição do índice de classes através de fatores reduzidos, aqui denominadas variáveis latentes (VL). As VLs descrevem o comportamento dos valores de$\mathbf{Y}$ no subespaço onde as amostras de $\mathbf{X}$ foram projetadas [62].

Já a LDA foi proposta inicialmente por R. Fisher para a discriminação de diferentes tipos de flores, sendo hoje utilizado em diversas aplicações [63]. O objetivo da LDA é realizar uma transformação linear através da projeção dos dados originais em um espaço de dimensões reduzidas e maximizando a separabilidade das classes. O critério para a redução das dimensões é maximizar a distância entre as amostras de diferentes classes e minimizar a distância das amostras dentro da mesma classe [64]. Entretanto, um fator limitante para o uso da LDA está relacionado à dimensionalidade dos dados originais. Caso a matriz inicial possua mais colunas do que linhas (como frequentemente encontrado em conjunto de dados espectrais, por exemplo), a matriz torna-se singular e a sua inversa não pode ser definida com exatidão. Além disso, variáveis correlacionadas também podem

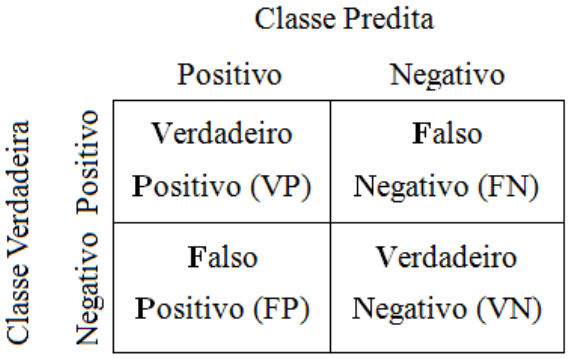
levar a singularidades no processo de inversão da matriz. Para contornar essa questão, são utilizados alguns métodos, dentre eles, a seleção de variáveis que busca reduzir o elevado número de variáveis a um subconjunto de tamanho inferior ao número de amostras e que seja o minimamente correlacionadas entre sí. [64].

A seleção de variáveis, além de permitir que o cálculo da LDA possa ser realizado, apresenta vantagens como a diminuição do custo computacional com a redução do conjunto de variáveis iniciais. A combinação das variáveis selecionadas são as que melhor explicam a variação entre as classes e há um menor risco de sobreajuste [65]. O melhor algoritmo para a seleção depende da natureza dos dados e dos objetivos da aplicação. No presente trabalho, o método avaliado foi o algoritmo genético (GA, do inglês, *Genetic Algorithm*), que é uma ferramenta de otimização usada para selecionar as variáveis mais representativas para o caso em estudo. O algoritmo genético é baseado na teoria evolutiva de Darwin, onde o processo evolutivo é simulado matematicamente. O método foi proposto por John H. Holland na década de 60, com o objetivo de otimizar sistemas complexos. Desde então, o algoritmo vem sendo modificado e melhorado para ser utilizado em diversas áreas, inclusive na química [66].

2.3.2.1 Validação dos métodos de classificação

Após a construção de um modelo de classificação, é necessário validá-lo. Há diversas formas de realizar uma validação de um modelo classificatório. Nesta dissertação optou-se por particionar o conjunto de amostras em um de treinamento e um de teste e avaliar a eficiência de classificação correta das amostras de teste baseado em erros do tipo I e II. O diagnóstico pode ser dado através de uma matriz de confusão, onde são dispostos os números de amostras designadas em cada classe [67] .

Figura 9 - Matriz de confusão.



Fonte: Filho R.(2015) . [68]

Como pode ser observada na Figura 9 em um exemplo para duas classes, a matriz de confusão gera quatro informações diferentes. Inicialmente, na diagonal da matriz encontram-se os verdadeiros positivos (VP) e os verdadeiros negativos (VN). Os VP são as amostras que foram classificadas corretamente como pertencentes à classe, enquanto os VN são as amostras classificadas corretamente como não pertencentes à classe. No canto superior direito, encontram-se os falsos negativos (FN), que são as amostras que pertencem à classe, mas que foram classificadas erroneamente como não-pertencentes, ou seja erro do tipo II. Já no canto inferior esquerdo, encontram-se os falsos positivos (FP), que são as amostras que não fazem parte da classe, mas que foram classificadas erroneamente como pertencentes, ou seja, erro do tipo I [69].

Com o conjunto de dados classificados, é possível calcular as figuras de mérito, que indicam o quão eficiente é o modelo. Dentre essas métricas pode-se citar a exatidão, uma medida das classificações corretas de VP e VN, a sensibilidade, que indica o quão eficiente o modelo foi para evitar falsos negativos, a especificidade, que indica a habilidade do modelo em evitar falsos positivos, e a precisão, que indica dentre todas as classificações de classe positivas que o modelo fez, quantas estão corretas [70].

2.4 DADOS CLÍNICOS

Dados clínicos podem ser definidos como conjuntos de observações coletadas de pacientes, os quais se encontram dentro de algum critério específico, como uma doença ou condição médica [71]. Normalmente, tais dados contém variados tipos de informações, sendo possível encontrar variáveis distintas entre si e de diferentes unidades. Assim, torna-

se imprescindível o conhecimento aprofundado das particularidades de cada observação, a fim de chegar em conclusões mais apropriadas sobre o problema a ser estudado [72]. O objetivo da análise multivariada em dados clínicos é a extração de máxima informação e reconhecimento de padrões entre as amostras, muitas vezes não acessados com uma simples análise univariada.

2.4.1 **Dados Faltantes**

Um problema frequentemente encontrado em dados clínicos são os dados faltantes. Eles ocorrem quando o valor de uma variável em uma amostra não está registrado. Isso pode ocorrer por uma série de razões, como: o paciente se recusou a responder uma determinada questão; erro ou indisponibilidade de algum equipamento para realizar um exame; tal informação simplesmente não foi solicitada para aquele paciente específico, etc [73].

A princípio, é importante entender em qual padrão os dados estão faltantes. Um sistema desenvolvido por Rubin permite identificar os mecanismos dos dados faltantes, os quais foram divididos em três categorias [74]:

1. MCAR (*Missing completely at random*): a probabilidade do dado estar faltante não está relacionada com os valores observados nem com os valores não observados, são completamente aleatórios.

2. MAR (*Missing at random*): a probabilidade do dado estar faltante está relacionada com os valores dos dados observados.

3. MNAR (*Missing not at random*): a probabilidade do dado estar faltante está relacionada com o próprio valor do dado faltante.

Uma das formas de lidar com a ausência de dados seria excluir completamente uma amostra ou uma variável até que não houvessem mais dados faltantes. Entretanto, a redução do conjunto de dados implica na redução da precisão estatística e dos parâmetros analisados. Uma abordagem válida seria a substituição de um dado faltante por um valor plausível, que não adicione nenhuma tendência nas análises estatísticas [73].

O SVD (do inglês, *singular value decomposition*) consiste em um método de imputação baseado na decomposição em valores singulares de uma matriz de dados. Inicialmente, os dados faltantes são mapeados e substituídos pela média de sua respectiva

coluna. É realizada então uma decomposição da matriz em três matrizes, U, S e V. Uma nova matriz M é construída a partir destas três matrizes ($M = U * S * V'$) e os valores obtidos para os dados faltantes são substituídos na matriz original. O processo é repetido até que os valores obtidos convirjam [75].

Um segundo método é o BPCA (do inglês, *Baeysean Principal Component Analysis*). O método está baseado na estimação dos valores faltantes através da abordagem bayesiana, acoplada ao algoritmo interative expectation maximization. A estimativa de valores é divida em basicamente três processos. A regressão dos valores existentes em componentes principais, inserção dos dados através da estimativa bayesiana e expectation–maximization (EM)-like repetitiven algorithm [76].

O último método abordado neste trabalho é baseado na imputação de valores através dos vizinhos mais próximos (KNN, do inglês, *k nearest neighbors*). A premissa do algoritmo é bastante simples. Os dados são projetados em um espaço multidimensional e os valores estimados para os dados faltantes são obtidos através da média do valores dos k vizinhos mais próximo. A distância entre as amostras é definida pelo cálculo da distância de Mahalanobis. Em alguns casos, essa média pode ser ponderada de acordo o quão distante está o vizinho do indivíduo em questão. Quanto mais distante, menor o peso atribuído, sendo o número de vizinhos utilizados deve otimizado caso a caso [77].

Essa versão clássica do KNN pode exigir um esforço computacional muito elevado, uma vez que é necessário calcular a distância entre todas as amostras. Por isso, foi desenvolvido um algoritmo chamado SKNN (do inglês, *sequential KNN, ou KNN sequencial*). Aqui, as amostras são divididas em dois grupos. As que não possuem dados faltantes e as que possuem dados faltantes. No segundo, as amostras são colocadas em ordem crescente em relação ao número de dados faltantes. A primeira amostra a ser realizada a imputação é a que possui menos dados faltantes. Após o cálculo da distância em relação às amostras completas e preenchimento dos dados, essa amostra passa a ser parte do conjunto de amostras completas, sendo também utilizada para as predições futuras. Assim, as amostras são preenchidas sequencialmente até que a última amostra, aquela que possui mais dados faltantes, seja completa. Vale salientar que para que o algoritmo funcione, deve haver um conjunto inicial de amostras completas. Nessa abordagem, o consumo computacional é bem menor e pode ser utilizado para conjuntos de dados com muitas variáveis e amostras [78].

Considerando que os modelos de imputação utilizam os valores observados para predizer os valores faltantes, resultados com tendência podem ser produzidos e conduzir a interpretações finais equivocadas. Assim, faz-se necessária a avaliação da qualidade de predição dos dados. Uma alternativa plausível é o uso do teste de Kolmogorov-Smirnov [79].

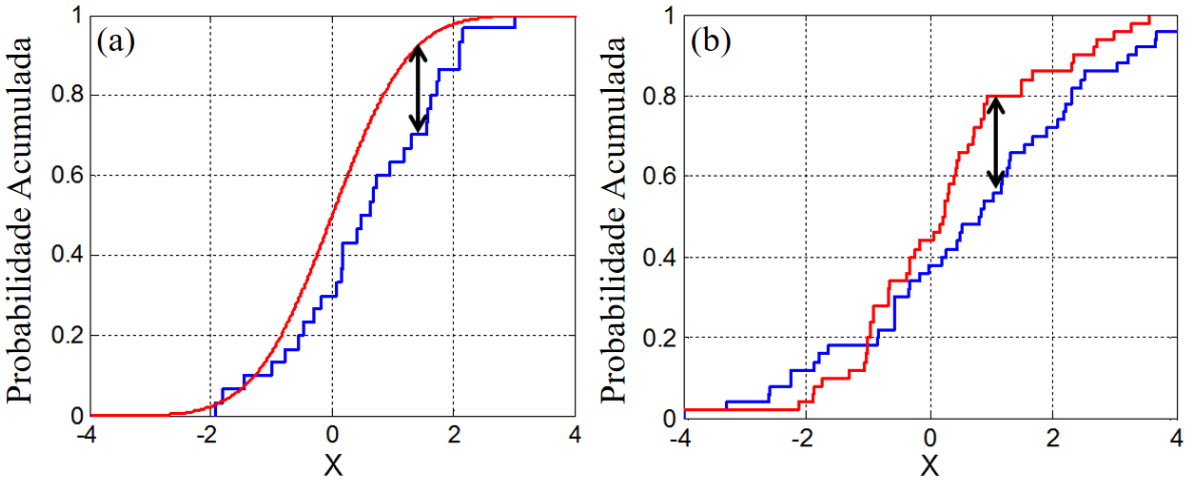
2.4.2 **Teste De Kolmogorov-Smirnov**

O teste de Kolmogorov-Smirnov (teste KS) é utilizado para medir e comparar a distribuição acumulativa com um modelo de referência ou entre duas amostras unidimensionais (2KS). O teste torna-se um bom método para verificar se duas amostras têm origem na mesma população e possuem distribuições semelhantes. o teste KS calcula a distância máxima (D) entre a função de referência ou entre outra amostra através da equação 4 [80].

$$D_{(a,b)} = sup|F_a - F_b| \tag{3}$$

onde F_a e F_b as funções de distribuição acumulativas das amostras a e b e *sup* a função de máximo [80]. Na Figura 10, D está representada graficamente.

Figura 10 - Representação gráfica de D em (a) KS teste para uma amostra e (b) KS teste para duas amostras. Em vermelho a distribuição antes da inserção e em azul a distribuição após a inserção.



Fonte: Teste Kolmogorov-Smirnov. Em: Wikipédia (2021). [81], modificado pelo autor.

Deve ser aplicada a hipótese nula de que:

H_0: as amostras são da mesma distribuição.

H_1: as amostras são de distribuições diferentes.

A hipótese nula é aceita a nível de confiança α se:

$$D_{(n,m)} < c(\alpha)\sqrt{\frac{n+m}{n \cdot m}} \tag{4}$$

sendo n e m os tamanhos das amostras, respectivamente, e $c(\alpha)$ o valor tabelado para cada nível de confiança [80].

O teste KS para duas amostras é muito sensível para diferenças locais e de forma das amostras, sendo considerado um bom teste para comparar amostras antes e após a inserção de dados que estavam faltantes. O teste já se mostrou eficiente quando o mecanismo presente do conjunto de dados é MCAR. O diagnóstico utilizando o teste KS foi realizado em dados de precipitação de chuva [82], em dados sociais [83] e em estudo clínicos epidemiológicos [84].

3 METODOLOGIA

3.1 SELEÇÃO DOS INDIVÍDUOS

O presente estudo foi aprovado pelo comitê de ética da Universidade Federal de Pernambuco e pelo Instituto de Medicina Integral Professor Fernando Figueira (número de aprovação: 2.075.028). Homens que possuíam idade entre 18 e 50 anos atendidos em uma clínica de fertilidade foram avaliados por um médico especialista em reprodução humana. A infertilidade foi diagnosticada segundo o *Practice Committee of the American Society of Reproductive Medicine* [85]. Os pacientes que apresentaram as seguintes características foram excluídos do estudo:

- Evidência de infecção urinária;
- Doenças urológicas diagnosticadas por qualquer exame hormonal;
- Defeitos genéticos;
- Histórico de criptorquidia;
- Uso de testosterona ou de qualquer outro anabolizante durante os últimos 12 meses;
- Histórico de quimioterapia ou radioterapia;
- Histórico de qualquer lesão nos testículos;
- Histórico de cirurgia escrotal.

No total, 90 pacientes foram avaliados e tiveram o sêmen coletado. Desses, dez foram descartados devido a baixa qualidade do material coletado (baixo volume ou eluição elevada do sêmen). Os 80 pacientes restantes foram divididos em três classes, apresentados na Tabela 1:

Tabela 1 - Número e características em grupos dos pacientes que tiveram o sêmen coletado.

Nome	Código	Número de amostras	Características
Grupo Controle	C	24	Homens saudáveis sem varicocele palpável, que possuíam pelo menos um filho nascido nos últimos 12 meses, sem histórico de infertilidade e tratamento, que desejavam fazer a cirurgia de vasectomia
Varicocele Fértil	VF	21	Homens férteis com varicocele palpável, que possuíam pelo menos um filho nascido nos últimos 12 meses, sem histórico de infertilidade ou tratamento.
Varicocele Infértil	VI	35	Homens inférteis com varicocele palpável, que não conseguiram ter filhos após 12 meses de sexo regular, sem proteção e sem evidências de infertilidade na parceira.

Fonte: O autor (2021).

O exame físico foi realizado em um ambiente quente e bem iluminado com o paciente de pé. O tamanho do testículo foi medido usando um orquidômetro de Prader e o grau da varicocele foi determinado segundo o critério de Dubion e Amelar [13]. As parceiras dos pacientes foram questionadas sobre a sua idade, histórico de tratamento de fertilidade, histórico de cirurgia ou doença pélvica e sobre a regularidade do ciclo menstrual.

3.2 ESTUDO METABONÔMICO

Devido ao tempo entre coleta e o estudo metabonômico, as 80 amostras de sêmen não estavam mais completamente disponíveis, uma vez que o material também foi utilizado para outros estudos e o volume disponível era reduzido. Deste modo, o estudo metabonômico foi conduzido utilizando 9 amostras do grupo de controle, 11 amostras de

varicocele fértil e 11 amostras de varicocele infértil.

3.2.1 **Preparo de Amostra**

Para o preparo de amostra com o DLLME, foram coletados 250 μL de cada amostra, sendo então adicionados 250 μL de acetonitrila para a precipitação das proteínas e macromoléculas. A mistura foi levada ao centrifugador (Hettich - Mikro 185) por dez minutos , 6000 RPM e 45° de inclinação (1814 g). Logo após, todo o sobrenadante foi coletado, ao qual foi adicionado rapidamente, com o auxílio de uma seringa, 1,25 mL de uma solução 1:4 de diclometano/acetona. As amostras foram então levadas a -40 °C para a separação total do solvente extrator (diclorometano) e para favorecer a remoção de água residual e gorduras (lipídeos). Em seguida as amostras ainda congeladas foram centrifugadas nas mesmas condições anteriores. O solvente extrator foi coletado, filtrado e armazenado em um vial com insert de 150 μL. No método QuEChERS foram adicionados 250 μL de acetonitrila e centrifugada nas mesmas condições do método prévio ao DLLME. Todo o sobrenadante foi coletado e adicionou-se uma mistura 2:1 de sulfato de magnésio ($MgSO_4$) e acetato de sódio ($ACNa$) até a saturação para promover a partição das fases orgânica e aquosa. A mistura, agitada vigorosamente para a solubilização dos sais, foi levada a centrifugação nas mesmas condições já descritas. A fase orgânica foi coletada, filtrada e armazenada em um insert de 150 μL.

3.2.2 **Otimização**

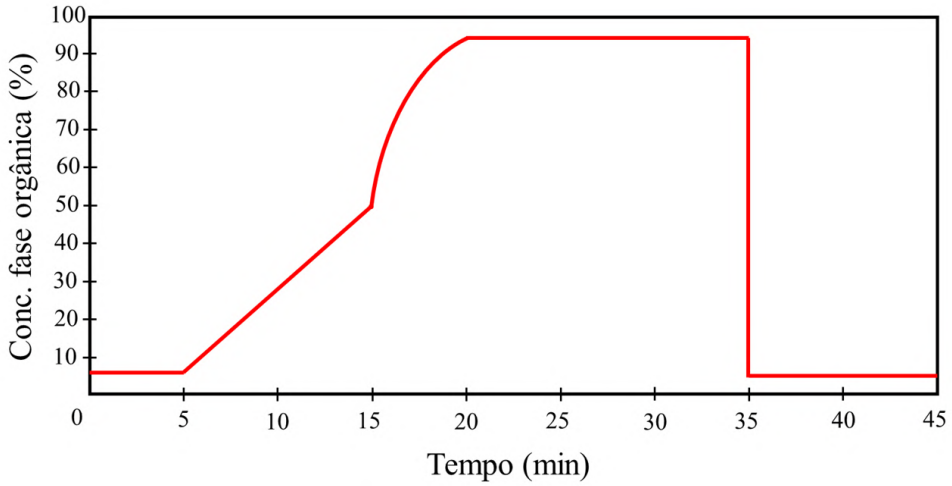
A otimização foi estruturada com o objetivo de avaliar o melhor resultado cromatográfico em quatro condições diferentes, utilizando dois tipos de preparo de amostra (DLLME e QuEChERS) e de duas fases móveis na corrida cromatográfica(ACN e MeOH). Uma amostra de cada classe foi selecionada, totalizando três amostras para cada método de preparo.

3.2.3 **Cromatografia Líquida**

A definição do método cromatográfico foi realizada com base no maior número de picos e resolução cromatográfica. As análises e obtenção dos espectros foram realizadas em um HPLC-DAD (Shimadzu UFLC - 13327). A coluna utilizada foi uma C18 (Luna,

00G-4252-E0, H20-209701, 250 x 4,6 mm, 100 Å, 5 µm).Uma bomba binária de água ultra pura (Milli-Q) e fase orgânica foi utilizada para fazer a mistura dos solventes. O fluxo definido foi de 1,5 mL/min e temperatura de forno constante de 30°C. O detector DAD (Shimadzu - SPD-M20A). O gradiente de concentração da fase orgânica seguiu os valores descritos na Figura 11.

Figura 11 - Gradiente de concentração da fase orgânica.



Fonte: O autor (2021).

A aquisição de dados foi realizada utilizando o *software* LabSolutions. Após a definição da melhor condição de trabalho, foram realizadas algumas mudanças no método cromatográfico. Cada amostra foi injetada em duplicata, sendo injetado um branco entre diferentes amostras. Para a correção do sinal cromatográfico foi realizada a subtração pelo branco, que consistiu em acetonitrila pura.

3.2.4 **Modelos quimiométricos com dados cromatográficos**

A construção dos modelos quimiométricos com os dados cromatográficos foram realizados utilizando duas abordagens: (1) Com o λ selecionado em 210 nm; (2) com o somatório de todos os comprimentos de onda (de 200 a 400 nm).

A avaliação e correção dos dados foi feita no *software* Matlab (MATLAB R2010a, MathWorks). O cálculo da ROBPCA foi realizado através da biblioteca LIBrary for Robust Analysis (LIBRA) [58], enquanto que para o métodos de Classificação foram utilizados os *ToolBox* Classification Milano [67]. Para a seleção das variáveis pelo algoritmo genético, foi utilizada uma população inicial de 200 indivíduos, 100 gerações e taxa de

mutação de 5%. Todas as figuras foram geradas em ambiente Python (versão 3.5.6) em interface Jupyter Notebook (versão 6.2.0).

3.3 DADOS CLÍNICOS

Todos os participantes selecionados passaram por análise de sêmen, ultrassonografia do escroto e o nível dos hormônios sexuais. A análise seminal foi feita de acordo com os parâmetros da Organização Mundial de Saúde (OMS), enquanto o pH foi testado utilizando uma fita de pH. A ultrassonografia do escroto foi realizada utilizando uma sonda de alta frequência (MARCA), onde foi definido o tamanho dos testículos, o diâmetro da varicocele e o refluxo venoso. Sangue venoso foi coletado de cada paciente entre as 7 e 11 horas da manhã. Os níveis de testosterona total (206 a 1200 ng/dL), estradiol (11.6 a 41.2 pg/mL), hormônio folículo-estimulante (FSH) (1.4 a 18.1 mUI/mL), hormônio luteinizante (LH) (1.5 a 9.3 mUI/mL) e a globulina ligadora de hormônios sexuais (SBHG) (10 a 57 nmol/L) foram medidos em tempo real por um imunoensaio quimioluminescente de fase sólida com o uso de um analisador automático (ADVIA Centaur XP, Siemens Healthcare Diagnostics). A albumina (3.4 a 4.8 g/dL) foi quantificada usando um ensaio colorimétrico (Abbott Diagnostics, Abbott Park, IL, USA) com um analisador automático (Architect® c16000, Abbott Diagnostics). Os níveis de testosterona livre (49.9 a 199.9 pg/mL) foram calculados usando a fórmula validada de Vermeulen et al [37]. A tabela inicial obtida com os dados coletados está indicada no Anexo A.

3.3.1 **Modelos quimiométricos com dados clínicos**

As análises exploratórias e classificatórias dos dados clínicos foram realizadas utilizando apenas os grupos VF e VI, totalizando 56 observações. O grupo controle foi removido desta análise, pois uma única variável (grau de varicocele palpável) é suficiente para distinguir pacientes com varicocele dos demais. Algumas variáveis foram removidas da tabela disponível no Anexo A. Foram elas: Idade do paciente, da parceira, da puberdade e o IMC, pois não traziam informações úteis ao modelo; e o grau de varicocele e o refluxo da ultrassonografia, por serem dados categóricos. Foram então consideradas 22 variáveis para a construção do modelo. São elas: exame clínico para avaliar o tamanho dos testículos (2 variáveis); os parâmetros seminais (7 variáveis); níveis hormonais (9 va-

riáveis); ultrassografia (4 variáveis).Todas as variáveis e o número de dados faltantes estão dispostos na Tabela 2.

Tabela 2 - Dados faltantes por variável.

n	Variável	nº de dados	nº de dados faltantes
1	Tamanho testículo direito	56	0
2	Tamanho testículo esquerdo	56	0
3	Concentração	48	8
4	Contagem Total	48	8
5	Motilidade Progressiva	48	8
6	Total de espermatozoides progressivos	48	8
7	Kruger	45	11
8	Volume	55	1
9	pH	55	1
10	Testosterona	56	0
11	Testosterona Livre	53	3
12	Estradiol	55	1
13	Hormônio luteinizante (LH)	56	0
14	Hormônio folículo-estimulante (FSH)	56	0
15	Globulina carreadora de hormônios sexuais (SHBG)	53	3
16	Albumina	55	1
17	Hormônio estimulante da tireoide (TSH)	47	9
18	T4 Livre	47	9
19	USG Testículo Direito	55	1
20	USG Testículo Esquerdo	55	1
21	USG veia Direita (Tamanho)	49	7
22	USG veia esquerda (Tamanho)	54	2

Fonte: O autor (2021).

O conjunto de dados inicial continha dimensão 56x22, com 6,65 % de dados faltantes. Todos os cálculos foram realizados em ambiente MATLAB 2010a. Foram escolhidos

quatro tipos diferentes de algoritmos para a inserção dos dados, sendo eles: SVD, KNN, BPCA e Média. Após a inserção dos dados, os conjuntos de dados gerados foram avaliados quanto a sua qualidade de predição através do teste de Kolmogorov-Smirnov.

Selecionado o melhor conjunto de dados, este foi selecionado e autoescalonado para as análises quimiométricas posteriores. Foi realizada uma PCA robusta a fim de detectar amostras anômalas. Após isso, foi realizada uma análise exploratória através dos escores e pesos gerados pela PCA robusta.

Em seguida, foram realizados cálculos de classificação através dos algoritmos PLS e LDA-GA utilizando as mesmas metodologias aplicadas aos dados cromatográficos.Para a seleção das variáveis pelo algoritmo genético, foi utilizada uma população inicial de 22 indivíduos, 50 gerações e taxa de mutação de 5% O conjunto inicial foi divido entre conjunto de treinamento (70%) e conjunto de teste (30%), sendo a divisão feita pelo algoritmo de Kennard-Stone. Os modelos foram avaliados quanto a exatidão, sensibilidade, precisão e especificidade, através de uma tabela de confusão.

4 **RESULTADOS**

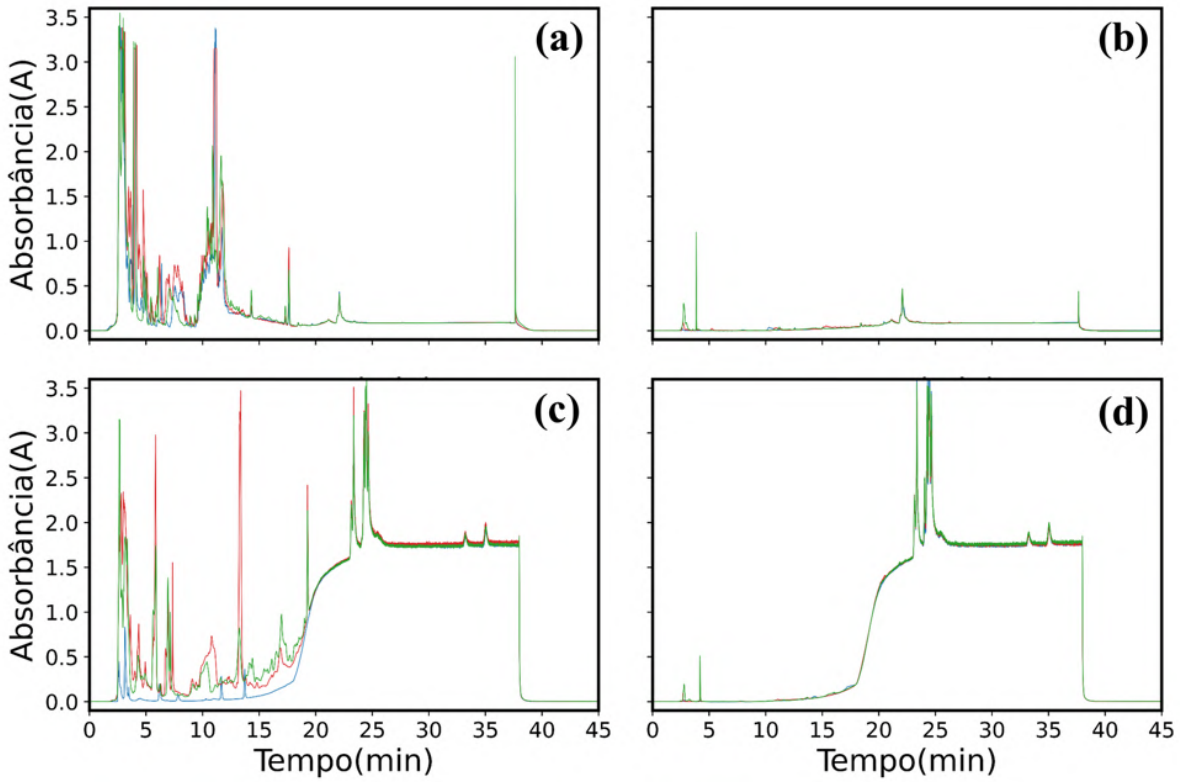
4.1 ESTUDO METABONÔMICO

4.1.1 **Otimização dos parâmetros de análise cromatográfico**

A otimização foi realizada avaliando-se o desempenho de dois tipos de preparo de amostra e dois tipos fase móvel. O cromatograma em 210 nm de cada classe estudada nos diferentes parâmetros utilizados está apresentado na Figura 12.

De acordo com os resultados apresentados na Figura 12, nas condições (b) e (d), onde foi utilizado o método QuECHERS, praticamente não há picos, indicando que a metodologia não foi capaz de extrair informações bioquímicas necessárias para o estudo metabonômico, sendo então esta abordagem descartada. Já nas condições (a) e (b), onde foi utilizado o DLLME, é possivel constatar que o método extraiu muita informação bioquímica, evidenciada pela presença numerosa de picos. É importante que o preparo de amostra escolhido apresente o máximo de informações possíveis que poderão ser usadas para melhorar a sensibilidade e seletividade dos modelos metabonômicos.

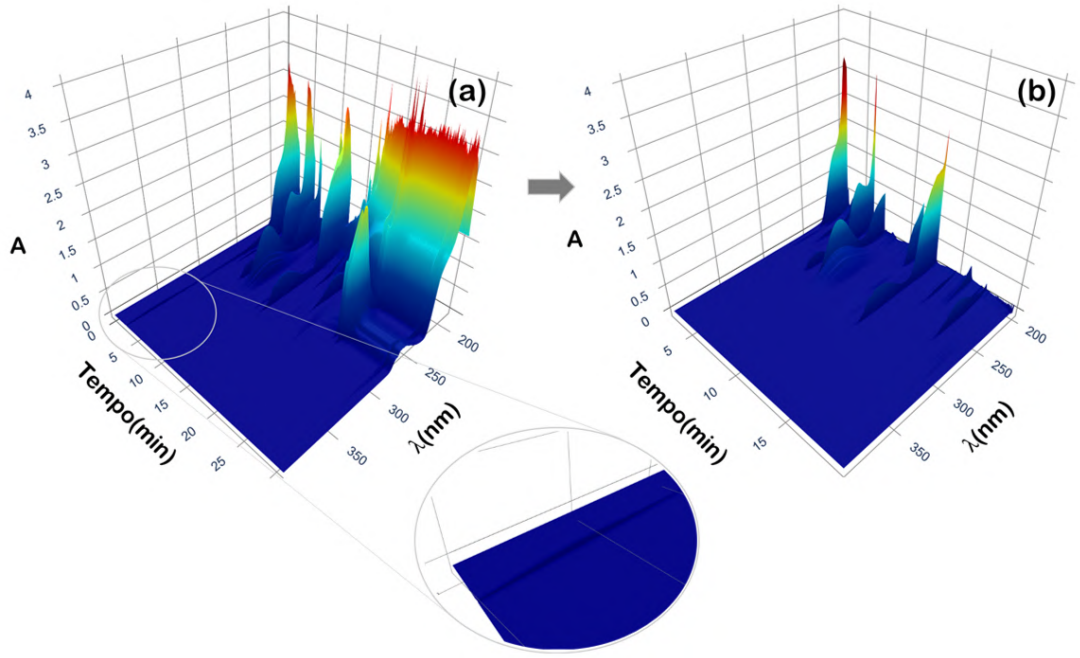
Figura 12 - Cromatograma registrado em 210 nm nos diferentes parâmetros avaliados. Preparo de amostra usando DLLME (a) e QuEChERS (b) com fase orgânica de ACN e Preparo de amostra usando DLLME (c) e QuEChERS (d) com fase orgânica de MeOH. Volume de injeção = 20 μL; Vazão = 1,0mL.min^{-1}; Temperatura do forno = 30°C; λ = 210 nm. Onde: C (—), VF (—) e VI (—).



Fonte: O autor (2021).

O metanol apresentou melhor desempenho na resolução dos picos, o que garantiu reprodutibilidade e melhor desempenho dos estudos quimiométricos apresentados nesta dissertação. Sendo assim, o DLLME foi selecionado como método de extração e o metanol como fase móvel orgânica. Todos os constituintes de interesse eluiram até 20 minutos, por este motivo o gradiente foi ajustado de forma a reduzir o tempo de análise de 45 para 36 minutos. Outro ajuste no método foi que o volume de injeção da amostra foi reduzido de 20 para 5 μL, uma vez que a absorção em 210 nm estava saturada em alguns picos o que poderia levar a desvios da lei de Lambert-Beer. A fim de visualizar os sinais analíticos em todos os comprimentos de onda, um cromatograma em 3D de uma amostra do grupo de controle está representado na Figura 13.

Figura 13 - Cromatograma em 3D de uma amostra do grupo de controle (a) e após a subtração do branco usado para corrigir a linha de base (b). Em destaque, o sinal referente ao tempo morto da coluna (deflexão do solvente). Volume de injeção = 5 μL; Vazão = 1,0mL.min^{-1}; Temperatura do forno = 30°C; λ = 200 a 400 nm.



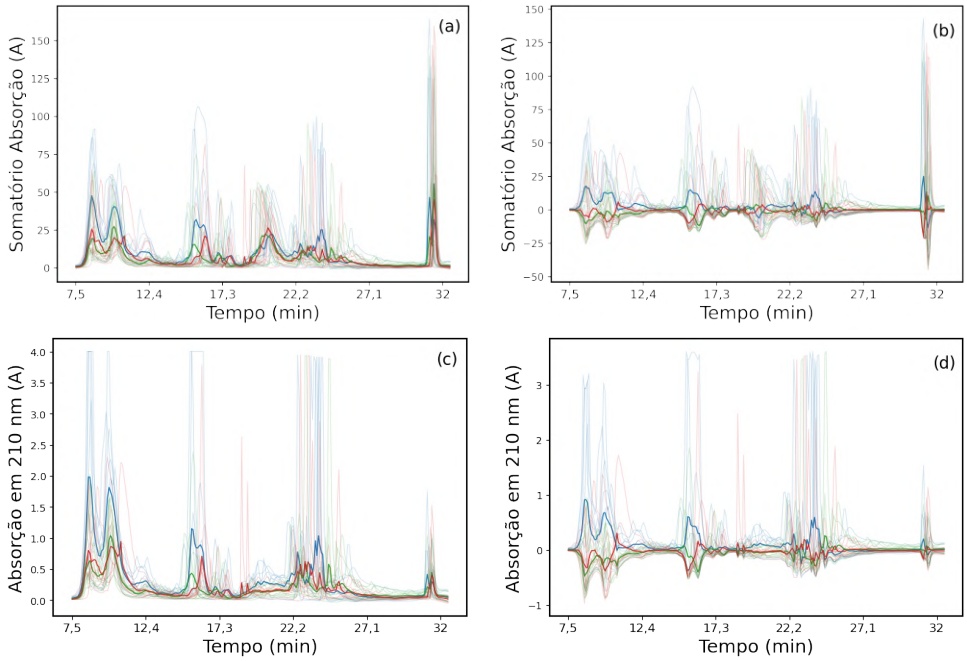
Fonte: O autor (2021).

Dois aspectos importantes podem ser visualizados na Figura 13(a). O primeiro diz a respeito do deslocamento da linha de base que começa em torno de 18 minutos. Isso ocorre devido ao ponto de corte do metanol que é próximo de 205 nm. O segundo aspecto, é o sinal em cerca de três minutos, como mostrado no destaque. Tal efeito acontece pela mudança no índice de refração causada pela passagem do solvente da amostra injetada no detector o que caracteriza o tempo morto da coluna. Devido a irreprodutibilidade do sinal próximo ao tempo morto os sinais analíticos de interesse são observados após este tempo de retenção. Após a subtração do branco nenhum sinal analítico foi observado após 20 minutos de eluição, assim os dados selecionados para os estudos quimiométricos posteriores compreendia os sinal entre os tempos de retenção de 2,9 e 20 minutos, como indicado na Figura 13(b).

As modelagens matemáticas foram realizadas utilizando-se dois tipos de dados. O primeiro, utilizando o somatório de todos os comprimentos de onda para cada tempo de retenção, enquanto o segundo selecionando-se apenas os dados no comprimento de onda de 210 nm. Ambos os dados foram posteriormente avaliados, pois apesar da maioria dos

constituintes apresentarem absorbância em 210 nm alguns apresentaram absorbância mais intensa em comprimento de onda distintos. Os cromatogramas corrigidos do somatório dos comprimentos de onda e em 210 nm estão indicados na Figura 14.

Figura 14 - Somatório do Cromatograma para cada amostras de 200 a 400 nm após a subtração e seleção entre o ponto de deflexação do branco e o final da região analítica de interesse (a) e após a centragem na média (b). Em destaque, a média de cada classe. Controle: (—), Varicocele Fértil (—) e Varicocele Infértil (—) Volume de injeção = 5 μL; Vazão = 0,6 mL.min^{-1}; Temperatura do forno = 30°C;



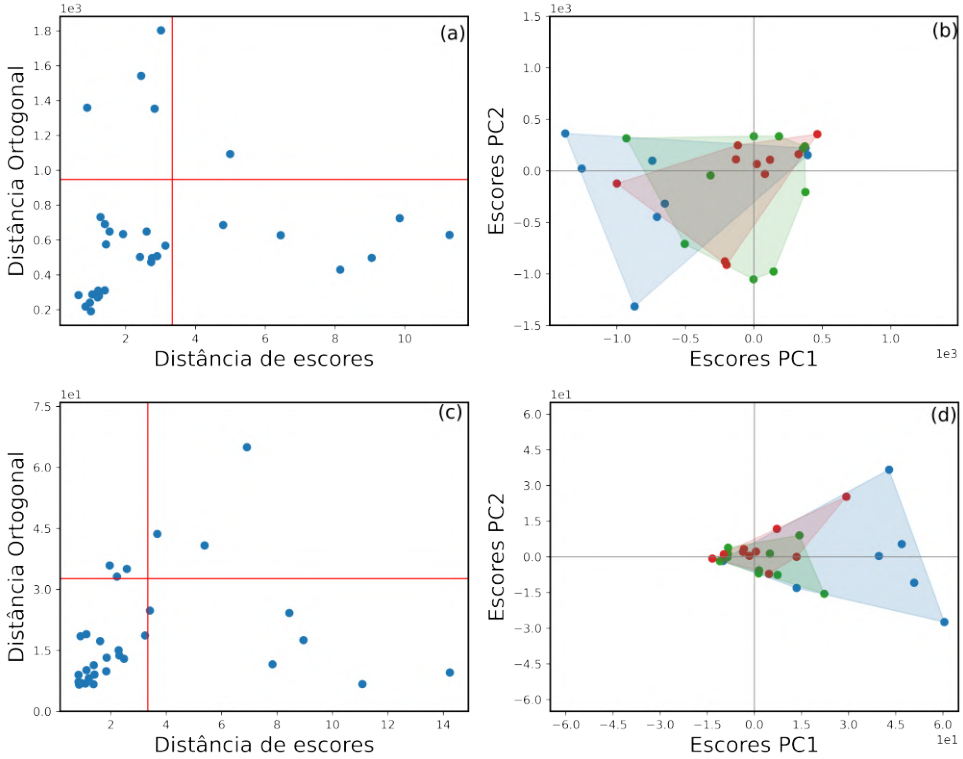
Fonte: O autor (2021).

Como indicado na Figura 14(a), na região onde havia sinal analítico de interesse há a presença de muitos picos cromatográficos que indicam diferenças entre as três classes estudadas. Nota-se também que o pico presente em todas as amostras um pouco antes de 32 minutos tem o mesmo tempo de retenção, o que indica reprodutilidade no método. Na Figura 14 está apresentada a centragem na média das amostras. Todas as modelagens matemáticas foram realizadas utilizando esse tipo de pré-processamento para os dados cromatográficos.

4.1.2 Detecção de amostras anômalas e Análise Exploratória - Dados cromatográficos

A identificação de amostras anômalas dentro de um conjunto de dados é fundamental para evitar que sejam carregados erros e tendências na modelagem. A detecção dessas amostras foi realizada utilizando o algoritmo da PCA robusta (ROBPCA). A Figura 15 apresenta o gráfico que dispõe a distância ortogonal e a distância de escores para cada amostra e os escores obtidos para cada conjunto de dados.

Figura 15 - Mapa de *outliers* para a somatório (a) e para os dados em 210 nm (c). Escores obtidos para as primeiras PCs para o somatório (b) e em 210 nm (d), onde : C (●), VF (●), VI (●).



Fonte: O autor (2021).

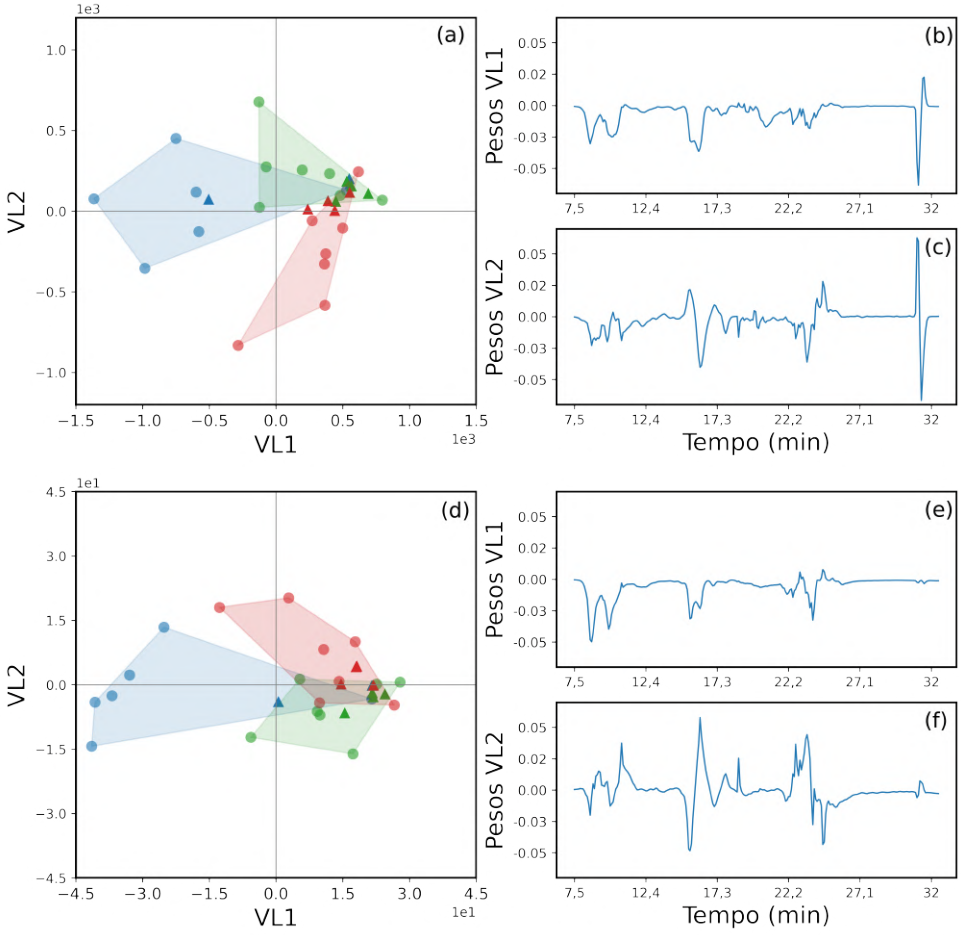
Ao observar os mapas de *outliers* na Figura 15(a) e Figura 15(b), em ambas as abordagens não foram verificadas a presença de amostras anômalas no primeiro quadrante que pudessem influenciar no modelo exploratório. Logo, todas as amostras foram consideradas para os modelos posteriores. Em relação aos gráficos de escores obtidos através do cálculo da PCA robusta (Figura 15(b) e Figura 15(d)), não foi capaz de proporcionar uma boa separabilidade entre as classes, onde pode-se notar que todas elas se encontram sobrepostas. Assim, faz-se necessária a utilização de modelos supervisionados apresentados

a seguir.

4.1.3 **Análise Classificatória dos Dados Cromatográficos**

O primeiro modelo classificatório avaliado foi o PLS-DA. O número ideal de variáveis latentes foi definido através da validação cruzada, onde se relaciona o erro associado ao número de VLs. Deste modo foi obtido um número de seis VLs para o somatório, com 70% de variância total explicada no conjunto de dados e de 54% no índice de classes. Para os dados em 210 nm, foi obtido apenas uma VL, com 44% de variância total explicada no conjunto de dados e de 28% no índice de classes. O gráfico de escores e pesos obtidos estão indicados na Figura 16.

Figura 16 - Gráfico de escores (a) e pesos (b) e (c) para o somatório e de escores (d) e pesos (e) e (f) para 210 nm no algoritmo PLS-DA. Onde: C treino (●), C teste (▲) VF treino (●), VF teste (▲), VI treino (●) e VI teste (▲).



Fonte: O autor (2021).

Começando pelos somatório dos dados cromatográficos, na Figura 16(a) se observa

que as classes apresentaram um padrão de separação melhor do que o apresentado pela análise exploratória (Figura 15(b)). Na variável latente 1 há a melhor separação entre os indivíduos saudáveis (c) e com varicocele (VF e VI). Ou seja, é possível inferir que a varicocele em si já influencia no perfil cromatográfico obtido através do soro do sêmen. Já a variável latente dois consegue explicar melhor as diferenças em relação ao status de fertilidade entre os indivíduos que possuem varicocele. Tal resultado indica que pode-se obter diferenças bioquímicas no sêmen de homens inférteis com varicocele quando comparados com homens férteis com varicocele. Assim, esse pode ser o primeiro passo de um diagnóstico mais rápido e preciso.

Em relação aos dados apenas em 210 nm, o comportamento apresentado foi muito semelhante quando se faz uma comparação visual dos escores. Assim, é interessante analisar a matriz de confusão dos dois modelos para averiguar qual apresenta melhores resultados. A matriz de confusão está indicada na Tabela 3.

Tabela 3 - Matrizes de confusão obtido para o somatório dos dados e em 210 nm através do algoritmo de PLS-DA. Em negrito estão os acertos feitos pelos modelos. Classe real na vertical e classe predita na horizontal.

		Treinamento					Teste			
		C	VF	VI	Exatidão		C	VF	VI	Exatidão
Somatório	C	**5**	0	1		C	**1**	1	1	
	VF	0	**7**	0	95%	VF	0	**4**	0	63%
	VI	0	0	**7**		VI	0	2	**2**	
		C	VF	VI	Exatidão		C	VF	VI	Exatidão
210 nm	C	**5**	1	0		C	**0**	3	0	
	VF	0	**6**	1	75%	VF	0	**4**	0	54%
	VI	0	3	**4**		VI	0	2	**2**	

Fonte: O autor (2021).

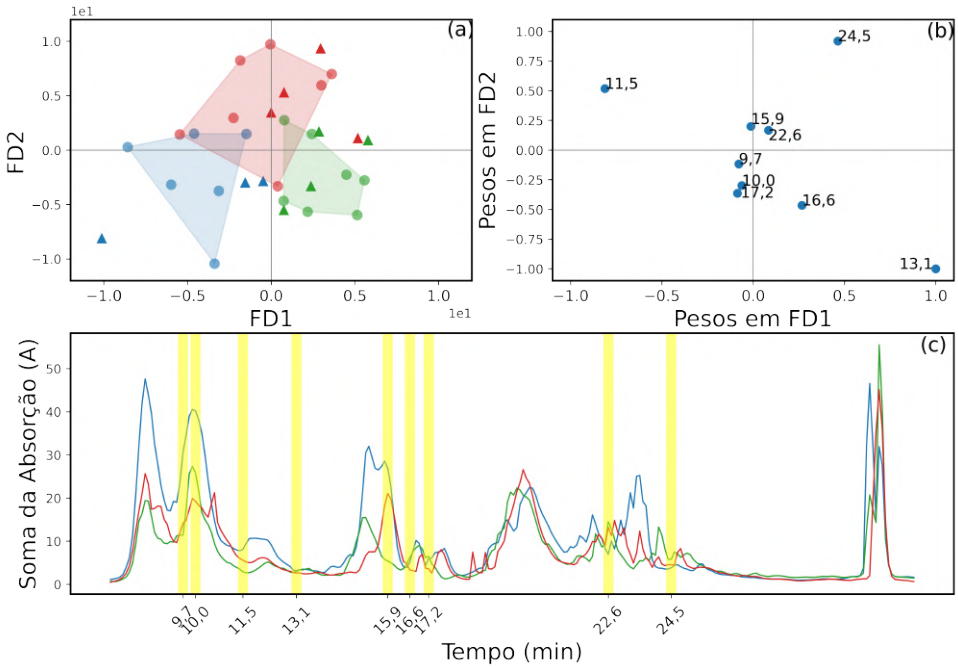
A exatidão obtida tanto no treinamento como no teste foram mais baixas no modelo de 210 nm do que no obtido no somatório. Ou seja, ao utilizar todo o espectro de 200 a 400 nm, o modelo tem menos chances de cometer erros através do PLS-DA. Isso pode indicar a presença de informações químicas detectadas em comprimentos de onda maiores. Ao também observar os pesos obtidos no modelo de PLS-DA na Figura 16, o pico um pouco antes de 32 minutos apresenta o maior peso na classificação das amostras utilizando

o somatório, mas praticamente não tem peso quando utilizando o λ de 210 nm. Tal comportamento pode ser um indicativo de algum marcador presente que interage com a radiação na região de comprimentos de onda mais alto no espectro.

Mesmo com o somatório, os resultados do PLS-DA através de somatório ainda deixa a desejar em relação à exatidão. As exatidãos entre o treinamento e teste variam muito, de 95% no treinamento e 63% no teste. Essa característica pode ser um sinal de sobreajuste do modelo, porém no presente caso ocorre devido ao número de amostras uma vez que o número de erros foram apenas dois do grupo controle no teste e 1 no treinamento.

Como o cromatograma possui muitas informações químicas, sendo cada pico uma substância diferente, alguns desses podem causar o confundimento entre as classes, uma vez que o PLS-DA utiliza todo o espectro disponível. Assim, o modelo de GA-LDA pode ser uma boa alternativa para selecionar os picos que realmente influenciam na separação das amostras. Considerando também que os dados cromatográficos são muito numerosos, a seleção de variáveis permite também a redução dos dados necessários para a análise, diminuindo o esforço computacional na predição de novas amostras. A Figura 17 apresenta os gráficos de escores, pesos e os tempos de retenção selecionados pelo GA-LDA.

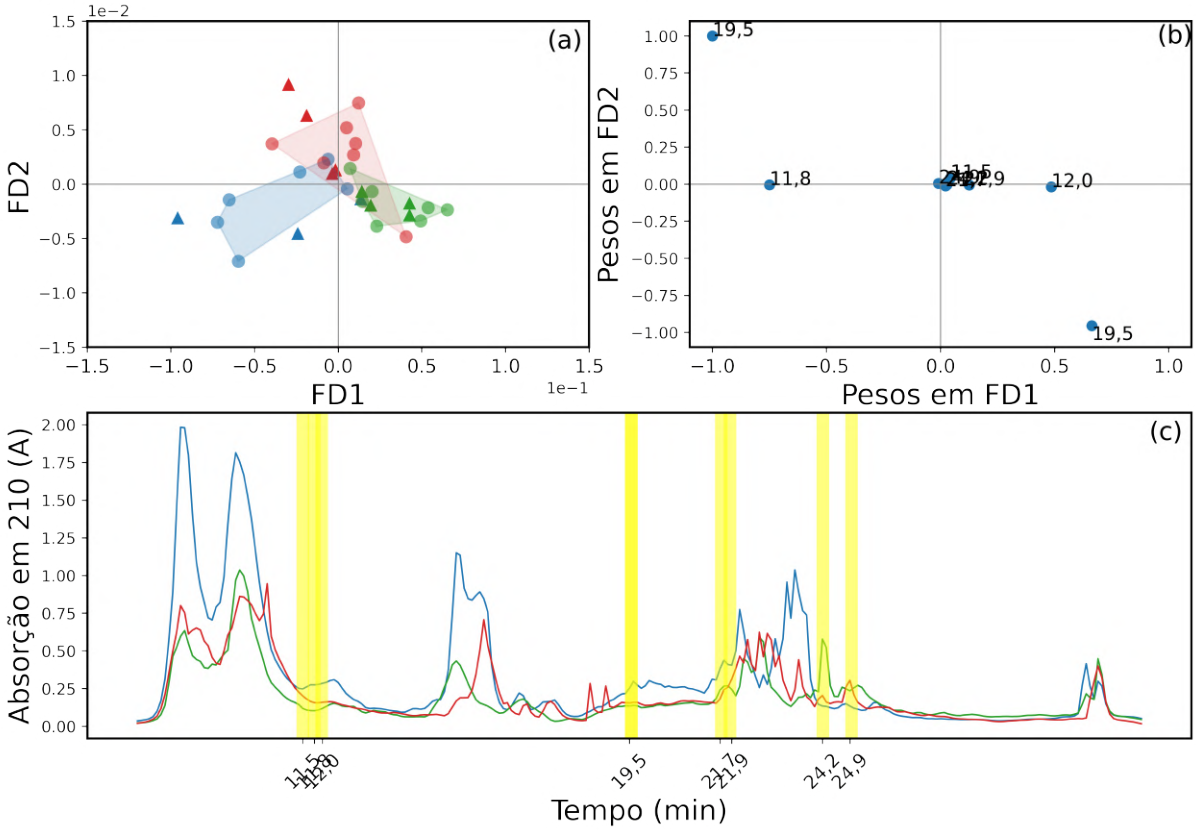
Figura 17 - Gráfico de escores do somatório obtidos pelo GA-LDA, onde: C treino (●), C teste (▲) VF treino (●), VF teste (▲), VI treino (●) e VI teste (▲). Gráficos de pesos (b). Tempos de retenção selecionados pelo modelo de GA-LDA em destaque em amarelo em comparação com a médias das classes (c), onde C (—), VF (—) e VI (—).



Fonte: O autor (2021).

O algoritmo genético selecionou 10 variáveis para fazer a classificação dos dados em duas funções discriminantes. O gráfico de escores Figura 17(a) apresenta uma melhor separação visual quando comparado ao PLS-DA(Figura 16(a)). Os tempos de retenção que apresentaram os maiores pesos na classificação de acordo com a Figura 17(b) foram os tempo de 11,5 min, 13,1 min e 24,5 min. Ou seja, com apenas 10 variáveis é possível fazer a classificação dos dados. O modelo para os dados em 210 nm também foi avaliado e está apresentado na Figura 18.

Figura 18 - Gráfico de escores em 210 nm obtidos pelo GA-LDA, onde: C treino (●), C teste (▲) VF treino (●), VF teste (▲), VI treino (●) e VI teste (▲). Gráficos de pesos (b). Tempos de retenção selecionados pelo modelo de GA-LDA em destaque em amarelo em comparação com a médias das classes (c), onde C (—), VF (—) e VI (—).



Fonte: O autor (2021).

Utilizando apenas o λ de 210 nm, a seperação entre as classes é um pouco prejudicada, como observado na Figura 18(a). Apesar do modelo ter selecionado menos tempos de retenção (nove), o desempenho foi reduzido. O algoritmo genético selecionou tempos de retenção muito próximos entre si, que podem até corresponder ao mesmo metabólito. Na tentativa de aumentar ao máximo a diferença entre as classes, o LDA acaba atribuindo pesos oposto pra informações muito parecidas, como os tempos de retenção próximos de 19,5 min, que estão em oposto no gráfico de pesos (Figura 18(b)). Além disso, a grandeza da FD1 é de uma ordem maior que a FD2 (Figura 18(a)), ou seja, praticamente toda a variação dos dados se encontram na VL1, o que prejudica a classificação. Assim, apenas com os dados em 210 nm há a perda de muita informação química que ajudem na discriminação das classes. As matrizes de confusão obtidas pelos modelos de LDA para as três classes estão dispostas na Tabela 4.

Tabela 4 - Matrizes de confusão obtido para o somatório dos dados e em 210 nm através dos algoritmos de PLS e LDA. Em negrito estão os acertos feitos pelos modelos. Classe real na vertical e classe predita na horizontal.

		Treinamento					Teste			
		C	VF	VI	Exatidão		C	VF	VI	Exatidão
Somatório	C	**5**	0	1		C	**2**	1	0	
	VF	0	**6**	1	80%	VF	0	**4**	0	81%
	VI	1	1	**5**		VI	0	1	**3**	
		C	VF	VI	Exatidão		C	VF	VI	Exatidão
210 nm	C	**4**	0	2		C	**2**	1	0	
	VF	0	**6**	1	80%	VF	0	**4**	0	90%
	VI	0	1	**6**		VI	0	0	**4**	

Fonte: O autor (2021).

O desempenho em relação a exatidão da LDA em relação ao PLS no somatório foi superior. Apesar da exatidão no conjunto de treinamento ter sido menor, o modelo conseguiu manter a taxa da acertos semelhante no conjunto de teste, indicando que não há sobreajuste nem subajuste na classificação. Como já citado anteriormente, o modelo em 210 nm não apresentou um resultado interessante, apesar da alta taxa de acerto no grupo de teste. A diferença de 10 % entre as exatidãos do grupo de treinamento e teste pode ser um indicativo de subeajuste, o que não é ideal para um modelo de discriminação.

Talvez o problema que impeça uma taxa de acerto maior nos modelos seja porque até agora se buscou a discriminação de três classes diferentes. Dentro de um cenário onde já se há o diagnóstico de varicocele no paciente, busca-se apenas inferir sobre a sua capacidade de fertilização. Como o objetivo é o diagnóstico da infertilidade, decidiu-se avaliar o desempenho dos modelos apenas com as classes de varicocele fértil e infértil, removendo assim o grupo de controle. Para ser mais objetivo e fluido na descrição dos dados aqui apresentados, os gráficos de escores, pesos e possíveis variáveis apresentadas para o PLS-DA e GA-LDA estão apresentados no APÊNDICE B e C, respectivamente. As matrizes de confusão dos modelos calculados com apenas as classes Vf e VI estão apresentados na Tabela 5.

Tabela 5 - Matrizes de confusão obtido para o somatório dos dados e em 210 nm através do algoritmo de GA-LDA. Em negrito estão os acertos feitos pelos modelos. Classe real na vertical e classe predita na horizontal.

		Treinamento				Teste			
PLS-DA	Somatório		VF	VI	Exatidão		VF	VI	Exatidão
		VF	**6**	1	93%	VF	**2**	2	75%
		VI	0	**7**		VI	0	**4**	
	210 nm		VF	VI	Exatidão		VF	VI	Exatidão
		VF	**6**	1	71%	VF	**4**	0	75%
		VI	3	**4**		VI	2	**2**	
LDA	Somatório		VF	VI	Exatidão		VF	VI	Exatidão
		VF	**7**	0	100%	VF	**4**	0	100%
		VI	0	**7**		VI	0	**4**	
	210 nm		VF	VI	Exatidão		VF	VI	Exatidão
		VF	**7**	0	100%	VF	**4**	0	100%
		VI	0	**7**		VI	0	**4**	

Fonte: O autor (2021).

Como pode ser observado na Tabela 5, de modo geral, o algoritmo de LDA com a seleção de variáveis apresentou um resultado bem melhor quando comparado ao PLS. Em linhas gerais, pode-se afirmar que a remoção do grupo de controle para a classificação apenas com os grupos com varicocele aumentou o número de acertos promovidos pelos modelos. Deste modo, a seleção dos picos cromatográficos é a melhor abordagem para esse tipo de dado.

O modelo de LDA não cometeu nenhum erro, tanto no somatório quando em 210 nm. Além do excelente resultado, isso também indica que apenas com os comprimentos de onda mais baixos é possível fazer a discriminação entre fértil e infértil nesse conjunto de dados. Em termos de praticidade, a obtenção dos dados também poderia ser realizada em equipamentos mais simples e baratos, como o HPLC com detector de ultravioleta, que lê não todo o espectro, mas alguns comprimentos de onda específicos.

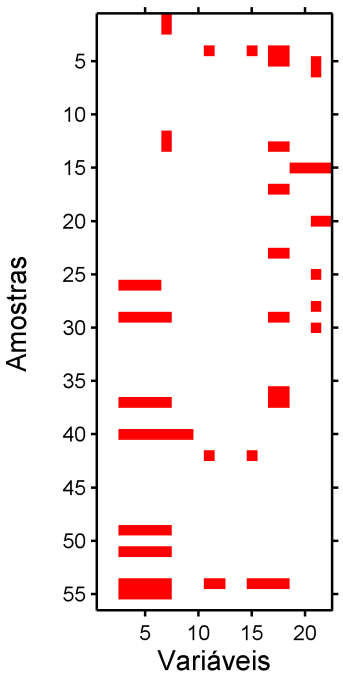
Os dados cromatográficos foram capazes de fornecer informações úteis para a diferenciação e classificação de pacientes com varicocele fértil e varicocele infértil. Isso indica que no soro do sêmen há evidências sobre o status de infertilidade, o que pode acelerar o

diagnóstico e promover tratamentos mais assertivos.

4.2 DADOS CLÍNICOS

Apesar dos resultados promissores das análises cromatográficas, aqui foi analisado se é possível usar um modelo para predizer infertilidade usando os dados clínicos realizados nos pacientes. Essa estratégia pode ser útil como uma triagem clínica e auxiliar os estudos metabonômicos ao indicar quais informações clínicas são mais importantes para este estudo. Na Figura 19 estão representados os dados faltantes indicados em quadrados vermelho.

Figura 19 - Mapa da matriz de dados. Os dados assinalados em vermelho indicam um dado faltante.



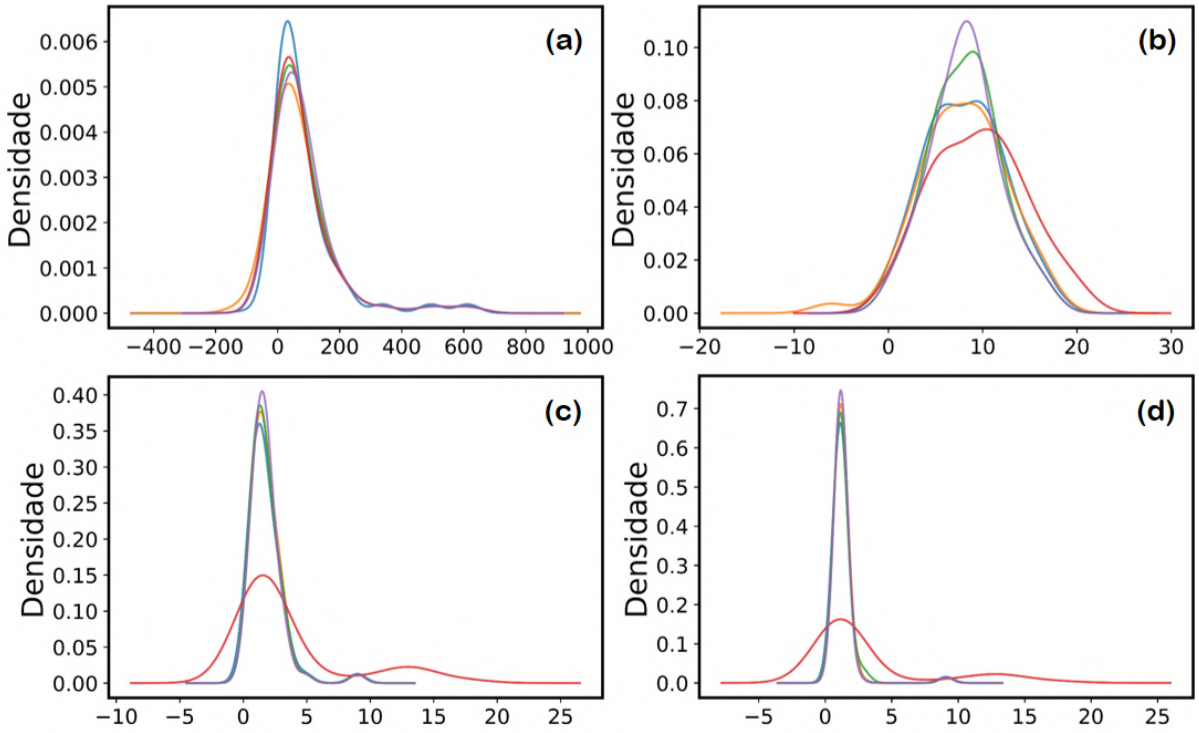
Fonte: O autor (2021).

Das 1232 observações esperadas, 82 não foram obtidas, o que equivale a 6,65% do total. Considerando os dados da Tabela 2, apenas as variáveis referentes ao tamanho dos testículos, testosterona, LH e FSH não possuem nenhum dado faltante. Sem a estratégia de inserção de dados, a análise multivariada só seria possível com estas variáveis. O teste de Little MCAR apresentou um valor de ρ igual a 0,829 para 252 graus de liberdade, indicando que o mecanismo dos dados faltantes é de fato MCAR, o que permite que a inserção seja feita com mais segurança, sem adicionar tendência aos dados [86].

4.2.1 **Validação dos métodos de inserção dos dados faltantes**

A inserção dos dados foi realizada utilizando os algoritmos BPCA, KNN, SVD e pela substituição pela média. Na Figura 20 estão apresentadas as densidades dos dados antes e depois da inserção para cada um dos modelos. Foram apresentadas apenas as quatro variáveis com mais dados faltantes.

Figura 20 - Gráficos para a densidade dos dados para as variáveis (a)Motilidade, (b)Kruger, (c)TSH e (d)T4 Livre antes e após a inserção de dados. Em que: Dados Originais (—), BPCA (—), KNN (—), SVD (—) e Média (—).

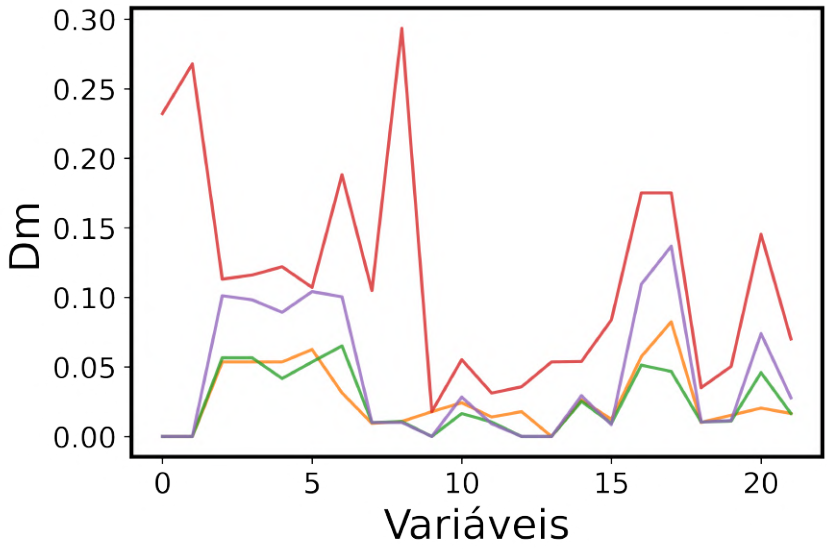


Fonte: O autor (2021).

O gráfico de densidade nos permite visualizar como os valores estão distribuídos para cada variável e como os resultados diferem dos dados originais, onde a densidade é obtida ignorando os valores ausentes. É evidente a diferença entre os dados originais e os dados produzidos pelo SVD (—). Os resultados mais próximos do original são os encontrados em KNN (—) e BPCA (—).

Antes de tirar conclusões do resultado do gráfico de densidade é necessário avaliar diferenças entre as distribuições dos dados após a inserção dos dados e os dados originais, usando o teste Kolmogorov-Smirnov. O valor do parâmetro D para cada variável, comparada com os dados iniciais, estão representadas na Figura 21.

Figura 21 - Valores de Dm para dos modelos de inserção para cada variável. Em que: BPCA (—), KNN (—), SVD (—) e Média (—).



Fonte: O autor (2021).

Como é possível observar na Figura 21, o algoritmo SVD (—) foi o que apresentou o maior valor de D para todas as variáveis, inclusive nas variáveis onde não haviam valores faltantes. Esse comportamento pode ser explicado pelo fato do SVD desconstruir toda a matriz considerando os dados faltantes como zero. Por isso, ao reconstruir a matriz de dados, pode carregar erros para as outras variáveis. Logo, não pode ser considerado como um bom candidato para fazer a inserção de valores nesse conjunto de dados.

Os valores de D para inserção feita com a média (—) também foi considerado alto quando comparada com os outros algoritmos. Esse resultado é esperado, pois sabe-se que os valores seguem uma distribuição e que a inserção feita apenas com a média pode levar a tendências que não estariam presentes nos dados originais.

Para os algoritmos de KNN (—) e BPCA (—), os valores de D apresentados foram bem próximos, onde podemos perceber que o KNN se apresentou melhor em algumas variáveis, o BPCA foi melhor em outras e em certas variáveis, não houve diferença. Ambos os métodos seriam ideais para o conjunto de dados e não apresentariam discrepâncias nas modelagens futuras. Porém, ao avaliar os valores produzidos pela BPCA, foram observados que haviam alguns dados faltantes inseridos com valores negativos. O BPCA assume que as amostras possuem uma distribuição bayseana, e para manter essa premissa como verdadeira, é possivel que seja feita a inserção de valores negativos. Porém, como

o conjunto de dados se trata de respostas positivas, não há significado físico em qualquer valor negativo. Por isso, o algoritmo KNN foi o selecionado para continuar com as modelagens, por inserir valores que não levaram a tendências no conjunto de dados e por terem um significado físico. Na tabela Tabela 6 estão apresentados os valores médios de D das variáveis de acordo com os métodos de inserção.

Tabela 6 - Média de D para cada um dos algoritmos utilizados para a inserção de dados.

algoritmo	média valor de D
BPCA	0,0267
KNN	0,0244
SVD	0,1148
Média	0,0435

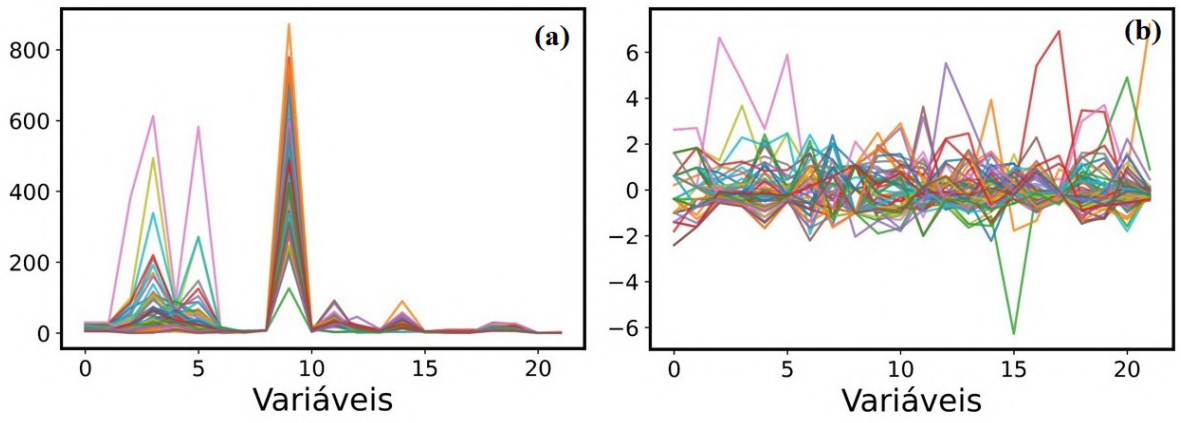
Fonte: O autor (2021).

Quanto menor o valor de Dm, mais próxima é a similaridade entre dois conjuntos de dados. Nota-se que o melhor resultado é encontrado no KNN, o que reforça a escolha sobre o método de inserção.

4.2.2 **Pré-Processamento dos dados**

Após a inserção dos dados faltantes, foi avaliada a necessidade do pré-processamento nos dados. Os dados antes e após o pré-processamento foram plotados e apresentados na Figura 22.

Figura 22 - Dados antes (a) e após o pré-processamento (b).



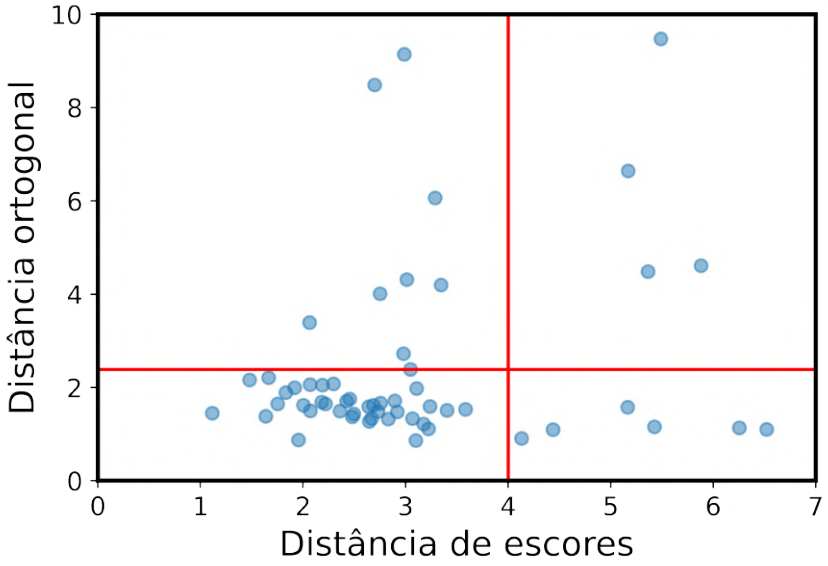
Fonte: O autor (2021).

De acordo com a Figura 22(a) nota-se que os dados possuem diferentes ordens de grandeza. Tal característica pode levar a uma tendência nas modelagens futuras, onde uma variável com maior ordem pode maior influência na distribuição das amostras, o que pode mascarar a importância de uma variável com menor variância, sendo necessária a aplicação de um pré-processamento. Os dados são compostos por variáveis discretas com valores contínuos, logo, a melhor escolha para pré-processamento é fazer o autoescalonamento. Como pode-se perceber, os dados após o autoescolonamento (Figura 22 (b) apresentam a mesma ordem de grandeza e não levará a tendência voltadas a maior importância para as variáveis com maior variância. Por esse motivo todas as análises realizadas serão feitas com os dados autoescalonados.

4.2.3 **Detecção de amostras anômalas e Análise Exploratória**

A detecção de amostras anômalas foi realizada utilizando o algoritmo da PCA robusta (ROBPCA). A Figura 23 apresenta o gráfico que dispõe a distância ortogonal e a distância de escores para cada amostra.

Figura 23 - Distância ortogonal e dos escores das amostras.



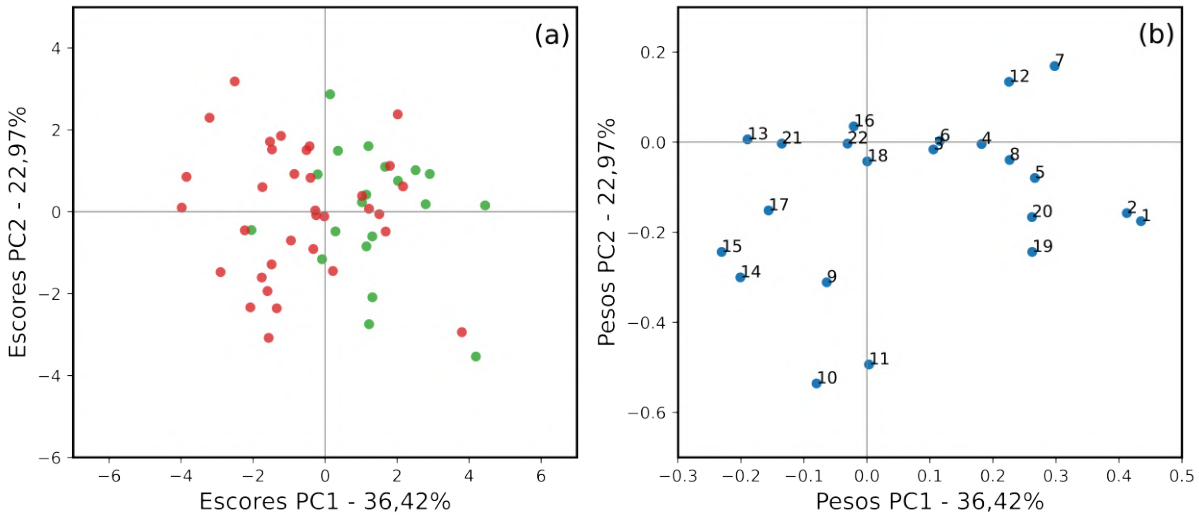
Fonte: O autor (2021).

As amostras estão presentes nas 4 regiões distintas do gráfico. Apesar de quatro amostras estarem no primeiro quadrante, estas não foram consideradas anômalas, uma vez que as distâncias ortogonais e de escores são elevadas suficiente ao ponto de causar

uma rotação no modelo. Além disso, essas 4 amostras foram analisadas quanto ao perfil em comparação com a média de cada classe e não foi observada nenhuma indicação de resultado anômalo. Neste caso, nenhuma amostra foi considerada anômala e as etapas posteriores foram realizadas com todas as 56 amostras.

Assim, foi realizada uma análise exploratória utilizando os escores e pesos calculados pela ROBPCA. Na Figura 24 estão apresentados os gráficos de escores e pesos da PC1 e PC2 obtidos pela ROBPCA.

Figura 24 - Escores (a), em que:VF (●) e VI (●) e pesos (b) obtidos pelo cálculo da ROBPCA.
.



Fonte: O autor (2021).

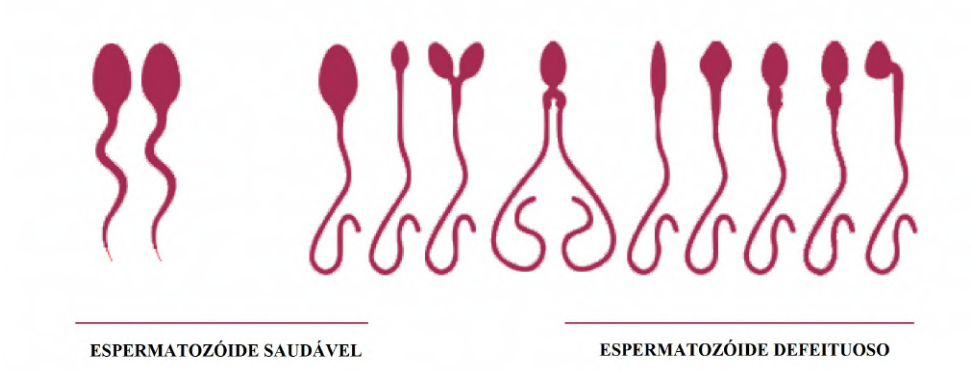
Pela análise exploratória, nota-se que há uma tendência de separação das classes de VF (●) e VI (●) no gráfico de escores (Figura 24a), principalmente na direção da PC1. Apesar das fronteiras não estarem bem definidas, é possível perceber que a classe VF encontra-se a direita da PC1, enquanto a classe VI fica localizada mais a esquerda. No gráfico de pesos (Figura 24b) é possível inferir que as variáveis (Tabela 2) que apresentam valores positivos de peso de PC1 apresentam maiores valores para a classe de pacientes VF quando comparado aos de VI. De forma similar, as variáveis (Tabela 2) que apresentam valores negativos de peso em PC1 apresentam maiores valores para a classe de pacientes VI quando comparado aos de VF.

De início, pode-se avaliar o comportamento das variáveis 1, 2, 19 e 20 em relação distribuição no gráfico de pesos (Figura 24b), onde se encontram mais a direita do gráfico. Estas variáveis carregam informações semelhantes em relação ao tamanho dos testículos

dos indivíduos, sendo 1 e 2 o tamanho avaliado com um orquidômetro de Prader e 19 e 20 através da ultrassonografia. Esse comportamento é justificado pelo fato do tamanho do testículo ser negativamente afetado pela presença da varicocele [87], mesmo sem afetar o status de fertilidade [88] . Como todos os indivíduos avaliados aqui possuem varicocele, os pesos da variáveis nos indicam que os pacientes férteis possuem em média o volume testicular maior do que aqueles inférteis. Apesar de não haver estudos que expliquem esse comportamento especificamente para a varicocele, uma recente investigação realizada por Bellurkar *et. al* (2020) com 354 homens, concluiu-se que o tamanho médio dos testículos está significativamente correlacionado com os parâmetros seminais, que por sua vez estão relacionados com a infertilidade. Logo, esses parâmetros podem ser um bom indicativo sobre a qualidade seminal do paciente, mesmo antes de serem realizados testes mais avançados.

Um comportamento semelhante é encontrado nas variáveis 5 e 7, que correspondem à motilidade progressiva e à morfologia de Kruger, respectivamente. São duas variáveis que estão relacionadas por avaliarem a qualidade dos espermatozoides. A morfologia definida pelo critério de Kruger indica a porcentagem de espermatozoides saudáveis que têm potencial de fertilização. Um porcentagem menor que 4% de células saudáveis denota uma grande chance do indivíduo ser infértil [90]. Na Figura 25 estão representadas as formas dos espermatozoides avaliados no parâmetro de Kruger.

Figura 25 - Representação das formas dos espermatozoides avaliados no parâmetro de Kruger.



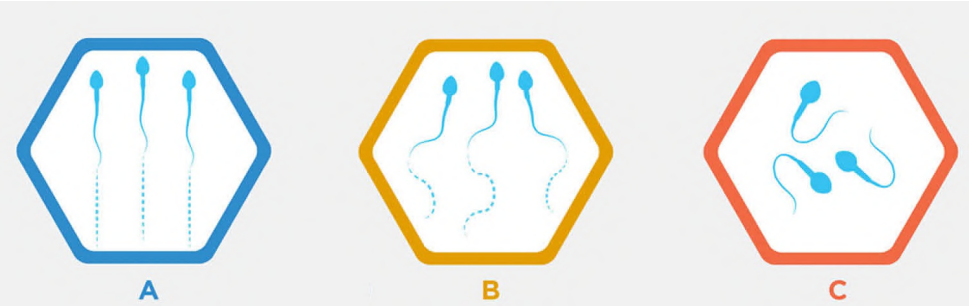
Fonte: Nova IVF Fertility (2020) [91]. Modificado pelo autor.

Os padrões para os espermatozoides mais saudáveis são os que se encontram à esquerda Figura 25, enquanto os menos saudáveis se encontram a direita. A morfologia influencia diretamente a motilidade do esperma e a sua capacidade de fecundação. A

motilidade é o parâmetro utilizado para avaliar como os espematozóides se movimentam (Figura 26), sendo dividido em graus [92]:

- **Grau A** (progressivo rápido): se movimentam para frente e rapidamente em linha reta;
- **Grau B** (progressivo lento): se movimentam para frente, mas em uma linha curva ou torta (motilidade linear lenta ou não linear);
- **Grau C** (não progressivo): os espermatozoides movem suas caudas, mas não avançam (apenas motilidade local);
- **Grau D** (imóvel): os espermatozoides não se movem.

Figura 26 - Graus de motilidade dos espermatozoides.



Fonte: Health Jade (2021) [93]. Modificado pelo autor.

Ambas as variáveis estão relacionadas com a capacidade de fertilização presente no esperma. Logo, é de se esperar que elas sejam mais pronunciadas no grupo fértil, como observado na Figura 24.

Partindo agora para lado oposto, temos as variáveis mais pronunciadas em VI, são elas: FSH (13), LH (14), SHBG (15), TSH (17) e USG Varicocele Direita (21). O LH e o FSH são dois hormônios relacionados com o pleno funcionamento das gônadas masculinas. O LH, através das células de Leydig, estimula a produção de esteroides sexuais, enquanto o FSH atua nas células de Sertoli para estimular a garantir a produção de espermatozoides [94].

Pelas funções citadas, é de se esperar que os níveis de FSH e LH sejam mais pronunciados em VF, uma vez que eles estimulam positivamente os parâmetros seminais.

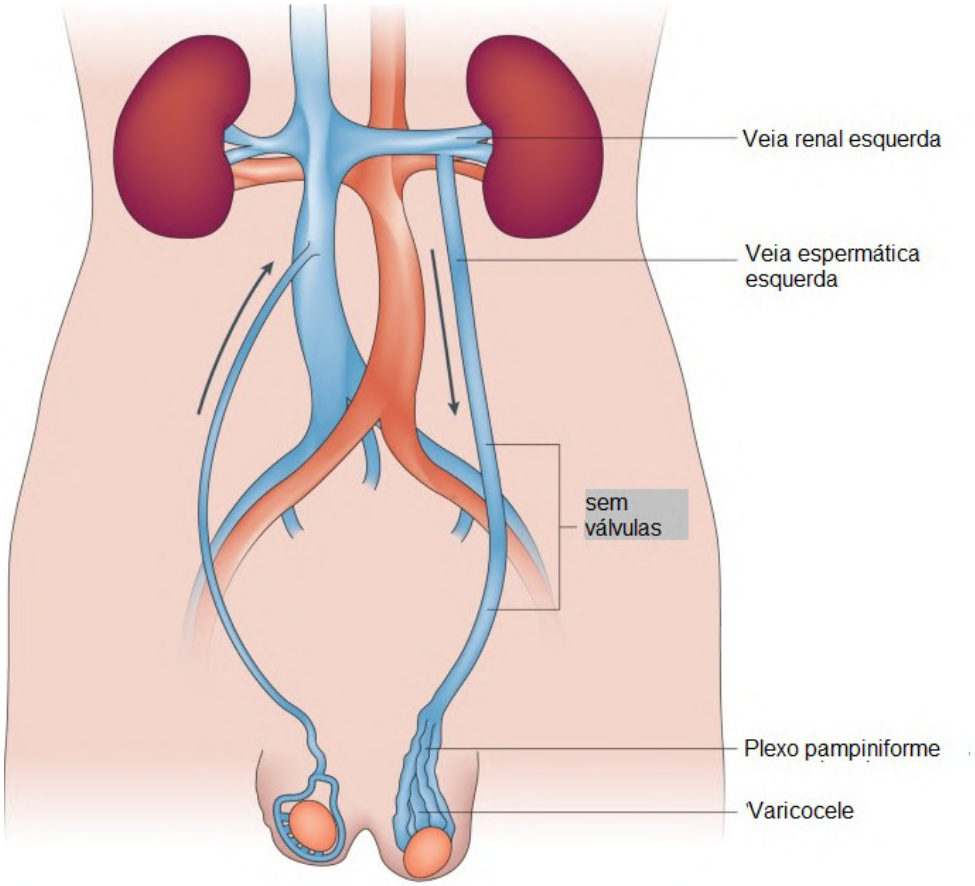
Porém, o contrário é encontrado, onde os níveis estão mais pronunciados em VI. Apesar de curioso, o resultado está de acordo com outro encontrados na literatura [95, 96]. Acontece que em função da insuficiência nos parâmetros testiculares, o corpo tenta trabalhar em um efeito compensatório a fim de recuperar os níveis normais de produção espermática [97]. Assim, é importante perceber que dentro de um quadro de varicocele, uma maior concentração desses hormônios na corrente sanguínea pode ser um indicativo de infertilidade.

A SHBG, outra variável numericamente pronunciada no grupo VI, é a chamada globulina ligadora de hormônios sexuais. A SHBG é secretada pelo fígado e está relacionada diretamente com a concentração de testosterona livre do sangue. Isso porque, a testosterona pode estar presente em três formas diferentes: a livre, a ligada à albumina e a ligada à SHBG. Como esta última ligação é muito forte, o hormônio quando ligado à globulina não está biologicamente disponível. Assim, um maior nível de SBHG na corrente sanguínea, como observado na classe VI, pode levar diminuição da testosterona disponível. Esta por sua vez influencia na qualidade seminal e status de fertilidade [98].

O comportamento da variável TSH pode ser justificado pelo papel indireto que esta possui no mecanismo de produção de espermatozoides. O TSH, hormônio estimulador da tireoide, é responsável por estimular a produção de T3 e T4 [99]. Os hormônios produzidos na tireoide agem nos testículos de várias maneiras, incluindo nas células de Leydig e Sertoli e nas células germinativas. O excesso de T3 e T4 resulta em alterações da função testicular, incluindo anormalidades do sêmen, como redução do volume e da densidade, motilidade e morfologia dos espermatozoides [100]. Como pode ser observado também no Apêndice B, o TSH possui uma correlação, mesmo que leve, negativa com os parâmetros seminais.

Por fim, o último parâmetro apresentado em relação ao grupo VI é o tamanho da varicocele direita. Antes de entender as justificativas desse comportamento, é preciso entender como funciona o mecanismo que causa a varicocele e o porquê dele ser diferente a depender do testículo esquerdo ou direito. Já se foi citado que a varicocele é uma dilatação da veia espermática, porém, ela costuma ocorrer em 90% dos casos no testículo esquerdo, quando a varicocele é unilateral [14]. Na Figura 27 está representado uma ilustração do caso.

Figura 27 - Anatomia das veias espermáticas.



Fonte: Jensen C. F. S. *et al.* (2017) [14]. Modificado pelo autor.

Em termos práticos, a varicocele acontece nas veias do plexo pampiniforme, de onde o sangue venoso sai dos testículos e segue para as veias espermáticas. A veia espermática esquerda forma um ângulo de 90° com a veia renal, onde tal arranjo proporciona o aumento da pressão, o que por sua vez facilita o refluxo sanguíneo, causando o dilatamento. Já a veia espermática direita forma um ângulo oblíquo com a veia cava, o que facilita o escoamento do sangue, já que não há aumento na pressão como ocorre na esquerda. Por isso, a frequência e intensidade do dilatamento são mais severas no testículo esquerdo [14].

Deste modo, segundo a análise exploratória, o tamanho da varicocele direita apresenta um peso maior na dispersão entre as classes de amostras. Provavelmente, o tamanho da dilatação do testículo esquerdo não seja tão diferente entre as duas classes, uma vez que todos os indivíduos estudados possuem a doença. Nesse caso, o resultado leva a crer que a dilatação no testículo direito é maior nos casos onde a infertilidade está presente.

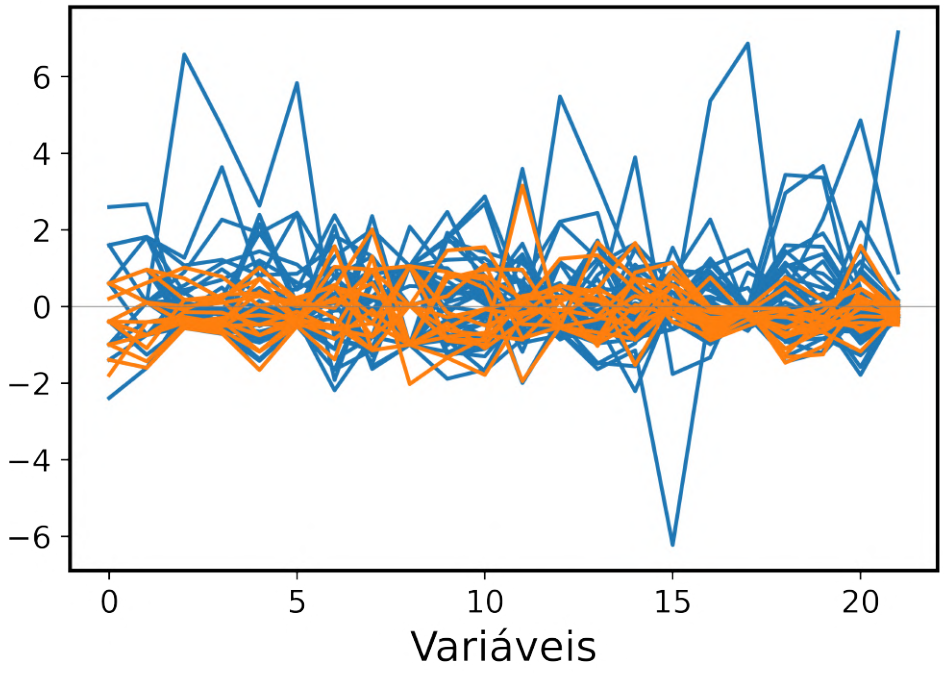
Mesmo a PCA apresentando uma tendência de separação clara, esta é uma aná-

lise exploratória que não nos permite inferir em qual classe a amostra está inserida. A construção de modelos classificatórios nos fornecerá informações sobre amostra futuras e possibilitará a predição de que classe esta amostra pertence.

4.2.4 **Análise Classificatória**

A primeira etapa para a análise classificatória é a separação do conjunto de dados entre o conjunto de treinamento e o conjunto de teste, selecionados com o algoritmo de Kenard-Stone. Na Figura 28 estão indicadas as amostras de treinamento e de teste. Como o conjunto de dados é pequeno, as amostras foram separadas apenas entre treinamento (70%) e teste (30%). Assim, para a classe VF, foram utilizadas 14 amostras para treinamento e 6 para teste, e 21 e 9 pra a classe VI, respectivamente.

Figura 28 - Amostras de treinamento (—) e teste (—).

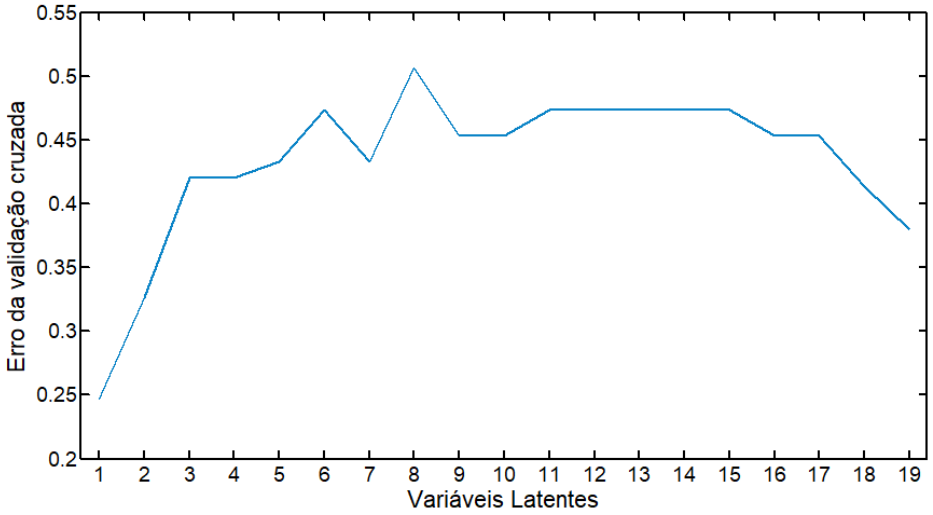


Fonte: O autor (2021).

4.2.4.1 Análise Classificatória por PLS-DA

A primeira classificação foi realizada utilizando o algoritmo de PLS-DA. A fim de definir o número ideal de variáveis latentes no modelo, foi realizada uma validação cruzada. Os erros para cada número de VL estão indicados no gráfico na Figura 29.

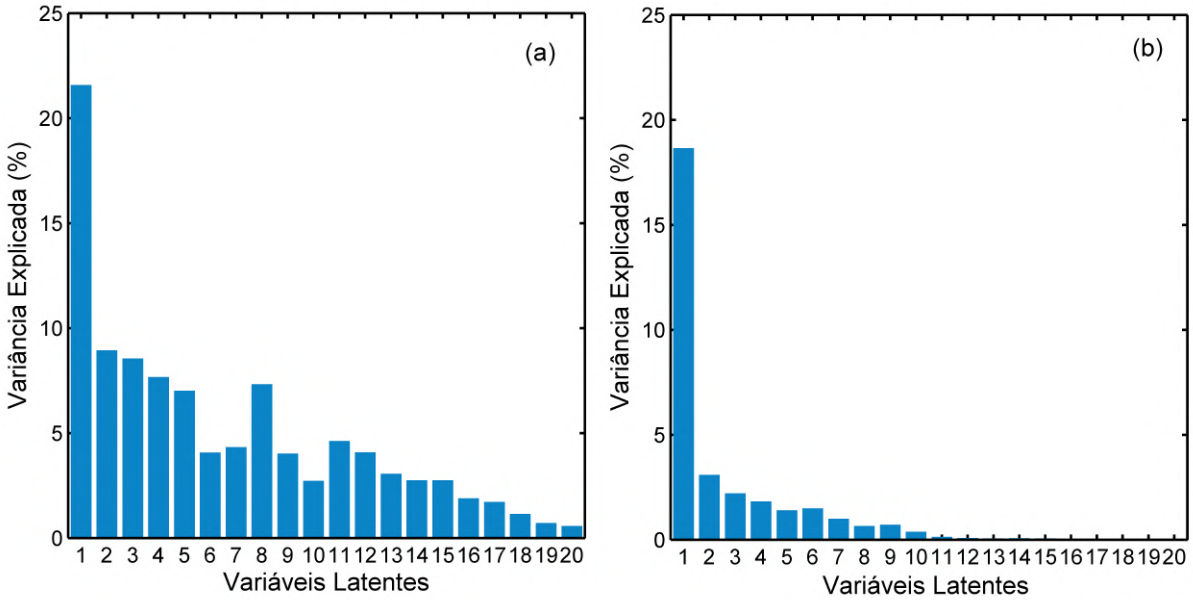
Figura 29 - Erro calculado para cada quantitativo de Variáveis Latentes.



Fonte: O autor (2021).

Como pode ser observado, o menor erro encontrado na validação cruzada dos dados foi utilizando apenas uma variável latente. Desde modo, a classificação por PLS-DA foi realizada apenas com uma VL, com variância explicada de 22%. A justificativa pelo uso de apenas uma variável latente está indicada na Figura 30.

Figura 30 - Porcentagem da variância explicada para cada variável latente na matriz de dados (a) e na matriz de identificação de classe (b).

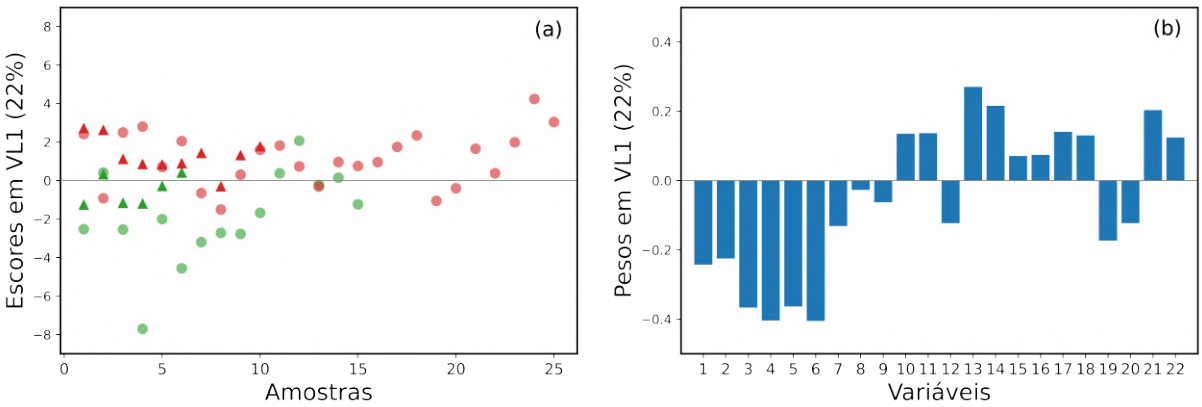


Fonte: O autor (2021).

Com exposto na Figura 30, apesar da pouca variância explicada pela primeira

variável latente (22%), esta é suficiente para diferenciar e classificar ambas as classes. Deste modo, a classificação foi realizada utilizando apenas uma VL. O gráfico dos escores e pesos para o modelo estão indicados na Figura 31.

Figura 31 - Análise classificatória dos grupos de varicocele fértil (VF) e varicocele infértil (VI) PLS_DA. Gráfico dos escores (a), em que: VF treino (●), VF teste (▲). VI treino (●) e VI teste (▲) e pesos(b).



Fonte: O autor (2021).

Segundo a distribuição das amostras na Figura 31(a), há um bom padrão de separação entre as classes, o que indica que os dados clínicos possuem poder de discriminação em relação a fertilidade. Nota-se também que o modelo foi capaz de classificar corretamente boa parte das amostras de teste em destaque. Em relação aos pesos na VL, o comportamento é semelhante ao encontrado na análise exploratória, onde as variáveis de três a cinco mais pronunciadas em VF e as variáveis 13, 14 e 21 mais pronunciadas na classe VI.

As variáveis de 3 a 6 são os parâmetros seminais e, como pode ser observado, possuem um peso maior em relação a classe fértil. De fato, é de se esperar que os pacientes férteis possuam parâmetros seminais melhores do que aqueles encontrados nos inférteis. Nota-se também, de acordo com o APÊNDICE D que estas variáveis são positivamente correlacionadas, ou seja, elas carregam as mesmas informações em relação aos dados. As justificativas em relação as variáveis 13, 14 e 21 são as mesmas dadas na discussão da análise exploratória, o que reforça o igual comportamento dos dados em duas abordagens diferentes. A matriz de confusão e as figuras de mérito do modelo classificatório calculado estão indicados na Tabela 7.

Tabela 7 - Tabela de confusão e figuras de mérito para o método de classificação PLS-DA entre os grupos de VF e VI.

Treinamento						
	VF	VI	sensibilidade	especificidade	precisão	Exatidão
VF	10	5	0,667	0,840	0,714	77,5%
VI	4	21	0,840	0,667	0,807	
Teste						
	VF	VI	sensibilidade	especificidade	precisão	Exatidão
VF	3	3	0,50	1	1	81,2%
VI	0	10	1	0,50	0,769	

Fonte: O autor (2021).

Como indicado na Tabela 7, no conjunto de treinamento o modelo foi capaz de classificar as amostras com 77,5 % de exatidão. Na classe VF, 5 amostras foram classificadas como VI, enquanto 4 amostras da classe VI foram inseridas no grupo VF. No conjunto de teste, o modelo conseguiu classificar corretamente todas as amostras do conjunto VI, mas errou metade das amostras em VF.

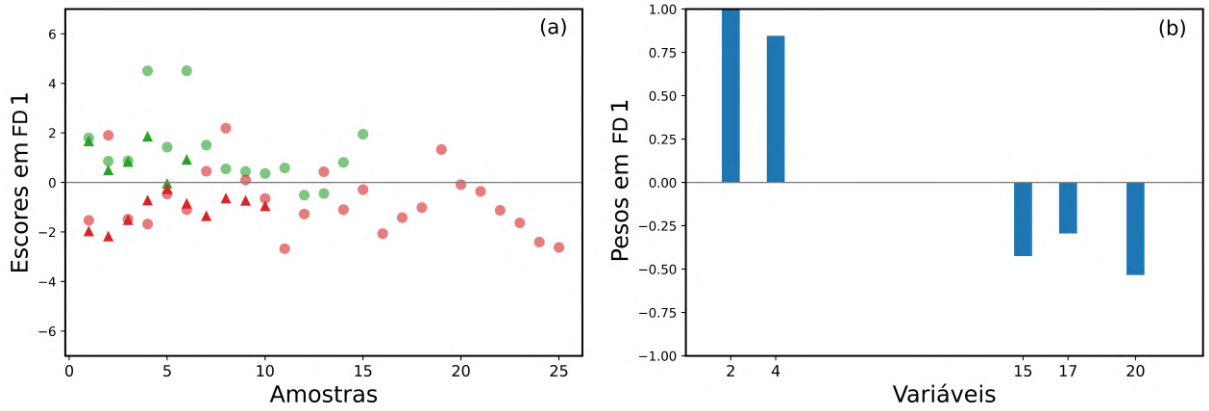
Nesse conjunto de dados específicos, existe um tipo de resultado ideal esperado em relação a classificação. Considerando que um paciente seja infértil e classificado como fértil, esse tipo de erro é menos desejável, uma vez que, hipoteticamente, o paciente não seguiria com nenhum tratamento. Já se o contrário ocorresse, um paciente fértil classificado como infértil, futuros tratamentos iriam provar o contrário. Assim, o modelo construído consegue muito bem classificar corretamente o grupo VI, evitando que eles sejam classificados como férteis. Já o grupo VF, apesar de metade estar classificada erroneamente, indica que o modelo tende a designar como infértil, o que no pior cenário levaria apenas a um diagnóstico mais tardio.

Como foi observado tanto na análise exploratória como na classificatória, existem variáveis que não contribuem para a diferenciação entre as classes, o que pode dificultar a discriminação. Assim, apesar do resultados satisfatórios, é possível ainda refinar as variáveis e alcançar classificações mais assertivas através da seleção de variáveis.

4.2.4.2 Seleção de variáveis e Classificação por LDA

A classificação dos grupos Varicocele Fértil e Varicocele Infértil foi realizada utilizando o método de seleção de variáveis algoritmo Genético (GA), seguido de uma Análise Discriminante Linear (LDA). Os resultados obtidos estão indicados na Figura 32. A tabela de confusão e as figuras de mérito dos grupos de treinamento e teste estão apresentados na Tabela 8.

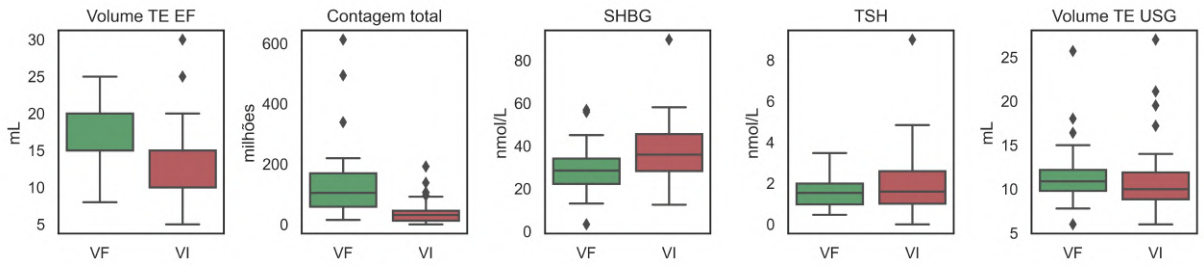
Figura 32 - Análise classificatória dos grupos de varicocele fértil (VF) e varicocele infértil (VI) por GA-LDA. Gráfico dos escores (a), em que: VF treino (●), VF teste (▲). VI treino (●) e VI teste (▲) e pesos(b).



Fonte: O autor (2021).

Nota-se de acordo com a Figura 32(a) que visualmente houve uma boa separação entre as classes. As variáveis selecionadas pelo GA foram as 2, 4, 15, 17 e 20. Na Figura 33 se encontram os *boxplot* de cada uma das variáveis selecionadas.

Figura 33 - Gráfico boxplot de cada uma das variáveis selecionadas pelo GA.



Fonte: O autor (2021).

As variáveis 2 e 4 possuem os maiores pesos do modelo, indicando que ambas são as que mais contribuem para a separação das classes. Eles correspondem respectivamente

ao tamanho do testículo esquerdo e a contagem total de espermatozoides progressivos. De fato, como já foi exposto nas discussões da análise exploratória, o tamanho do testículo esquerdo (o mais afetado pela varicocele) parece ser um parâmetro interessante sobre o status de fertilidade. No caso do conjunto de dados, o testículo esquerdo é maior naqueles pacientes férteis, mesmo que possuam varicocele (Figura 33). A variável quatro, que é um dos parâmetros seminais, também foi importante para a diferenciação dos grupos, onde a maior contagem de espematozóides se encontram no grupo fértil (Figura 33).

As variáveis 15 e 17, SHBG e TSH, respectivamente, apesar de possuírem um menor peso na classificação (Figura 32), também contribuem para o diagnóstico. Como já citado, ambas são parâmetros hormonais e estão ligadas com a produção e disponibilidade dos espematozóides. Como observado no gráfico de pesos, elas se pronunciam mais no grupo VI. Como observado também a Figura 33, o SHBG é maior no grupo VI.

Por fim, a variável 20 (USG Testículo Esquerdo) possuem peso negativo e mais pronunciado na classe VI. Esse comportamento é contrário ao encontrado na análise exploratória, onde a variável estava mais pronunciada no grupo VF. É importante lembrar que a LDA trabalha de forma a aumentar a distância entre as classes e aproximar as amostras do mesmo grupo. Aqui, a variável 20, apesar de apresentar um comportamento de início estranho, é responsável por diminuir a dispersão entre as amostras da mesma classe. Esse comportamento é interessante para garantir uma correta classificação de amostras futuras. Nota-se também que na Figura 33 que a diferença entre as classe é baixa, apesar da distribuição de valores em VF ser maior para essa variável.Na Tabela 8 está os resultados da matriz de confusão obtida para este modelo.

Tabela 8 - Tabela de confusão e figuras de mérito para o método de classificação GA-LDA entre os grupos de VF e VI.

Treinamento						
	VF	VI	sensibilidade	especificidade	precisão	Exatidão
VF	13	2	0,87	0,88	0,81	87,5%
VI	3	22	0,88	0,87	0,92	
Teste						
	VF	VI	sensibilidade	especificidade	precisão	Exatidão
VF	5	1	0,83	1	1	93,8%
VI	0	10	1	0,83	0,90	

Fonte: O autor (2021).

Os resultados obtidos pela classificação através do LDA foram superiores aos obtidos por PLS, mesmo utilizando apenas 5 variáveis, conforme indicado na Tabela 8. A exatidão obtida para o grupo de treino e teste foi de cerca de 87 % e 93 %, respectivamente. Aqui, o número de falsos negativos em relação ao grupo VF também caiu, mas as amostras VI continuam sendo classificadas corretamente, o que é uma vantagem para o tipo de diagnóstico que se almeja.

Apesar dos bons resultados apresentados na seleção de variáveis pelo AG é importante avaliar quanto a reprodutibilidade das mesmas. O modelo construído de LDA foi considerado satisfatório mas as variáveis de maior peso, que estão relacionadas com o tamanho, dependem muito de quem está fazendo o exame. Esse tipo de dado quando aplicado em análises multivariadas pode levar a tendências e interpretações não muito assertivas.

Os dados clínicos que foram utilizados para a construção de modelos de classificação mostraram-se capazes de fornecer informação suficiente sobre o status de fertilidade de pacientes com varicocele. Apesar disso, algumas variáveis não apresentaram muita influência nessa classificação. Aqui, vale ressaltar que isso não significa que elas não são importantes no diagnóstico, mas que dentro desse conjunto específico de dados elas não apresentaram peso suficiente de separação quando comparadas à outras variáveis. São elas: o pH, a albumina, a testosterona e a testosterona livre, onde os valores podem ser observados nos boxplots disponível no APÊNDICE E.

Começando pelo pH, a falta de influência desta variável pode estar associada com algumas razões, principalmente pela forma como ela foi medida. Segundo a metodologia utilizada, o pH foi determinado utilizando fitas de pH, o qual só permite resultados absolutos entre 0 e 14. Como o corpo humano trabalha para que o pH seja mantido na faixa neutra, alterações nessa medida seriam mínimas e não detectadas por uma simples fita de pH. Essa variável merece um foco especial pois ela pode ser um indicativo do estresse oxidativo, conhecida consequência da varicocele.

Um estudo publicado por Ghabili e colaboradores (2009), onde se propõe um mecanismo causador da infertilidade em casos de varicocele, indica que o aumento de espécies oxidativas de oxigênio são responsáveis pela degradação ácida dos produtos testiculares. Como consequência dessa acidificação, a motilidade dos espermatozoides é negativamente afetada, e, como o pH se encontra em desequilíbrio, algumas proteínas antioxidantes importantes para o pleno funcionamento das células também não funcionam. Logo, somando todo esse cenário com o refluxo presente, ocorre a diminuição do pH no meio. Por isso, vale aqui destacar que o uso de um pH-metro seria ideal para esse tipo de avaliação.

Em relação à albumina, além de ser um antioxidante, parte da testosterona que não está biologicamente disponível está associada com esta proteína. Em um estudo realizado por Boeri *et al.* (2019), foi verificado que baixos níveis de albumina no sangue estão relacionados não só com a baixa qualidade seminal mas também com níveis hormonais não controlados. De certo modo, nos modelos construídos neste trabalho, as principais influências nas classificações estão ligadas com a resposta da baixa albumina. Talvez esse comportamento possa justificar a sua ausência de peso como variável.

Em relação a testosterona, sabe-se que esse hormônio apresenta uma menor concentração em pacientes com varicocele, como provado no estudo apresentado por Ando *et al.*(1984). A principal razão se dá pelos danos que a varicocele causa nas células de Leydig, que são responsável pela produção de hormônios andrógenos. Porém, são necessários estudos mais aprofundados que justifiquem a causa dessa baixa influência entre homens férteis e inférteis com varicocele, uma vez que a testosterona também está associada com a infertilidade [104].

5 CONCLUSÕES E PERSPECTIVAS FUTURAS

A partir de amostras de soro de sêmen foi possível realizar um estudo metabonômico através de dados cromatográficos e dados clínicos para a discriminação de homens com varicocele quanto à fertilidade. O DLLME foi o método de extração que melhor conseguiu extrair informações da matriz.

Foi verificado que se obtem melhores resultados de classificação quando o grupo de controle não é incluído na modelagem matemática. O algoritmo de LDA foi o que melhor classificou os dados cromatográficos, sendo a taxa de acerto de 100% tanto pra o somatório dos comprimentos de onda quanto para os dados em 210 nm. A inserção dos dados faltantes foi melhor alcançada utilizando o algoritmo de KNN sequencial, sendo possível então realizar as análises multivariadas nos dados completos. A análise exploratória dos dados clínicos indicou que os parâmetros físicos seminais e hormonais foram os principais responsáveis pela diferenciação entre as classes. Os estudos classificatórios com os dados clínicos apresentaram resultados semelhantes, sendo necessário estudos mais aprofundados sobre o papel específico e os efeitos sinérgicos de cada variável nas diferentes classes. Assim, os estudos apresentados aqui, apesar de preliminares, apresentam elevado potencial para funcionarem como triagem para o diagnóstico de infertilidade em homens com varicocele. Podendo então, após o diagnóstico inicial de varicocele, ser realizada a triagem inicial a partir dos dados clínicos e, em caso, positivo para infértil, proceder com a validação a partir do cromatograma do soro do sêmen. Assim, o tempo de diagnóstico será reduzido, quando comparado com o método tradicional hoje disponível. Logo, mesmo com os resultados promissores aqui encontrados, as abordagens necessitam de maiores validações e melhorias que podem ser alcançadas com:

- Aumento do número de amostras analisadas;
- Realizar o estudo metabolômico de soro de sêmen com detector de espectroscopia de massas a fim de determinar a estrutura dos metabólitos que diferenciam as classes envolvidas e propor rotas de ação da doença.
- Como a infertilidade na varicocele é uma condição que depende de vários mecanismos e fatores, o uso de outras ciências ômicas, como a genômica e a proteômica podem prover pistas mais claras sobre as principais causas e rotas de ação da doença.

- Balancear o número de amostras entre as classes, tanto para os dados cromatográficos quanto para os dados clínicos

REFERÊNCIAS

[1] LARSON, J. *La Vie Bohème*. [S.l.]: Arif Mardin, 1996.

[2] MASCARENHAS, M. N. et al. National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. *PLoS Med*, Public Library of Science, v. 9, n. 12, p. e1001356, 2012.

[3] BORGES, L. d. S. et al. Avaliação da concordância diagnóstica entre métodos não invasivos e endoscopia na investigação de infertilidade. *Revista Brasileira de Ginecologia e Obstetrícia*, SciELO Brasil, v. 27, n. 7, p. 401–406, 2005.

[4] FONSECA, R. P.; MACEDO, L. C. Varicocele: a principal causa da infertilidade masculina. *Saúde e Pesquisa*, v. 8, n. 1, p. 167–174, 2015.

[5] PASTUSZAK, A. W.; WANG, R. Varicocele and testicular function. *Asian journal of andrology*, Wolters Kluwer–Medknow Publications, v. 17, n. 4, p. 659, 2015.

[6] CHAN, S. Male infertility: diagnosis and treatment. *Canadian Family Physician*, College of Family Physicians of Canada, v. 34, p. 1735, 1988.

[7] NETO, F. T. L. et al. 1h nmr-based metabonomics for infertility diagnosis in men with varicocele. *Journal of assisted reproduction and genetics*, Springer, v. 37, n. 9, p. 2233–2247, 2020.

[8] BAIGORRI, B. F.; DIXON, R. G. Men's health: Varicocele: A review. In: THIEME MEDICAL PUBLISHERS. *Seminars in interventional radiology*. [S.l.], 2016. v. 33, n. 3, p. 170.

[9] ALSAIKHAN, B. et al. Epidemiology of varicocele. *Asian journal of andrology*, Wolters Kluwer–Medknow Publications, v. 18, n. 2, p. 179, 2016.

[10] LEVINGER, U. et al. Is varicocele prevalence increasing with age? *Andrologia*, Wiley Online Library, v. 39, n. 3, p. 77–80, 2007.

[11] ZYLBERSZTEJN, D. S.; ESTEVES, S. C. Varicocele. In: *Male Infertility*. [S.l.]: Springer, 2020. p. 391–407.

[12] MASSON, P.; BRANNIGAN, R. E. The varicocele. *Urologic Clinics*, Elsevier, v. 41, n. 1, p. 129–144, 2014.

[13] DUBIN, L.; AMELAR, R. D. The varicocele and infertility. In: *Male infertility*. [S.l.]: WB Saunders Co Philadelphia, 1977. p. 57–67.

[14] JENSEN, C. F. S. et al. Varicocele and male infertility. *Nature Reviews Urology*, Nature Publishing Group, v. 14, n. 9, p. 523, 2017.

[15] AGARWAL, A. et al. Role of oxidative stress in pathogenesis of varicocele and infertility. *Urology*, Elsevier, v. 73, n. 3, p. 461–469, 2009.

[16] HSIUNG, R.; NIEVA, H.; CLAVERT, A. Scrotal hyperthermia and varicocele. In: *Temperature and Environmental effects on the testis*. [S.l.]: Springer, 1991. p. 241–244.

[17] GOLDSTEIN, M.; EID, J.-F. Elevation of intratesticular and scrotal skin surface temperature in men with varicocele. *The Journal of urology*, Wolters Kluwer Philadelphia, PA, v. 142, n. 3, p. 743–745, 1989.

[18] KHALAFALLA, K. et al. Varicocele and testicular hyperthermia: infrared digital thermographic measurement of scrotal and inguinal temperatures among varicocele patients and normal controls. *Fertility and Sterility*, Elsevier, v. 112, n. 3, p. e362–e363, 2019.

[19] NAUGHTON, C. K.; NANGIA, A. K.; AGARWAL, A. Varicocele and male infertility: part ii: pathophysiology of varicoceles in male infertility. *Human reproduction update*, Oxford University Press, v. 7, n. 5, p. 473–481, 2001.

[20] HOSSEINIFAR, H. et al. Study of sperm protein profile in men with and without varicocele using two-dimensional gel electrophoresis. *Urology*, Elsevier, v. 81, n. 2, p. 293–300, 2013.

[21] LEE, J.-D.; JENG, S.-Y.; LEE, T.-H. Increased expression of hypoxia-inducible factor-1α in the internal spermatic vein of patients with varicocele. *The Journal of urology*, Elsevier, v. 175, n. 3, p. 1045–1048, 2006.

[22] SHEEHAN, M. M.; RAMASAMY, R.; LAMB, D. J. Molecular mechanisms involved in varicocele-associated infertility. *Journal of assisted reproduction and genetics*, Springer, v. 31, n. 5, p. 521–526, 2014.

[23] WANG, H. et al. Hypoxia-induced apoptosis in the bilateral testes of rats with left-sided varicocele: a new way to think about the varicocele. *Journal of andrology*, Wiley Online Library, v. 31, n. 3, p. 299–305, 2010.

[24] MAJZOUB, A. et al. Scrotal hyperthermia, hormonal disturbances, testicular hypoperfusion, and backflow of toxic metabolites in varicocele. In: *Varicocele and Male Infertility*. [S.l.]: Springer, 2019. p. 27–35.

[25] CAMOGLIO, F. S. et al. Varicocele and retrograde adrenal metabolites flow. *Urologia internationalis*, Karger Publishers, v. 73, n. 4, p. 337–342, 2004.

[26] STEENO, O.; KOUMANS, J.; MOOR, P. D. Adrenal cortical hormones in the spermatic vein of 95 patients with left varicocele. *Andrologia*, Wiley Online Library, v. 8, n. 2, p. 101–104, 1976.

[27] TURNER, T.; LOPEZ, T. Testicular blood flow in peripubertal and older rats with unilateral experimental varicocele and investigation into the mechanism of the bilateral response to the unilateral lesion. *The Journal of urology*, Elsevier, v. 144, n. 4, p. 1018–1021, 1990.

[28] MEDICINE, P. C. of the American Society for R. et al. Diagnostic evaluation of the infertile male: a committee opinion. *Fertility and sterility*, Elsevier, v. 103, n. 3, p. e18–e25, 2015.

[29] OMENN, G. S. et al. Evolution of translational omics: lessons learned and the path forward. National Academies Press, 2012.

[30] VAILATI-RIBONI, M.; PALOMBO, V.; LOOR, J. J. What are omics sciences? In: *Periparturient Diseases of Dairy Cows*. [S.l.]: Springer, 2017. p. 1–7.

[31] Pereira Braga, C.; ADAMEC, J. Metabolome analysis. In: RANGANATHAN, S. et al. (Ed.). *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press, 2019. p. 463 – 475. ISBN 978-0-12-811432-2.

[32] CANUTO, G. A. et al. Metabolômica: definições, estado-da-arte e aplicações representativas. *Química Nova*, SciELO Brasil, v. 41, n. 1, p. 75–91, 2018.

[33] VUCKOVIC, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Analytical and bioanalytical chemistry*, Springer, v. 403, n. 6, p. 1523–1548, 2012.

[34] FIGUEIREDO, E.; BORGES, K.; QUEIROZ, M. Preparo de amostras para análise de compostos orgânicos. *Editora Gen LTC, Rio de Janeiro*, 2015.

[35] BURATO, J. Soares da S. et al. Recent advances and trends in miniaturized sample preparation techniques. *Journal of Separation Science*, Wiley Online Library, v. 43, n. 1, p. 202–225, 2020.

[36] ANASTASSIADES, M. et al. Fast and easy multiresidue method employing acetonitrile extraction/partitioning and "dispersive solid-phase extraction" for the determination of pesticide residues in produce. *Journal of AOAC international*, Oxford University Press, v. 86, n. 2, p. 412–431, 2003.

[37] PRESTES, O. 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. *Química Nova*, SciELO Brasil, v. 32, n. 6, p. 1620–1634, 2009.

[38] CASTRO-PUYANA, M.; HERRERO, M. Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, Elsevier, v. 52, p. 74–87, 2013.

[39] CHEVOLLEAU, S.; BOUVILLE, A.; DEBRAUWER, L. Development and validation of a modified quechers protocol coupled to uhplc-apci-ms/ms for the simple and rapid quantification of 16 heterocyclic aromatic amines in cooked beef. *Food chemistry*, Elsevier, v. 316, p. 126327, 2020.

[40] RUBERT, J. et al. Evaluation of mycotoxins and their metabolites in human breast milk using liquid chromatography coupled to high resolution mass spectrometry. *Analytica chimica acta*, Elsevier, v. 820, p. 39–46, 2014.

[41] CASADO, N. et al. An improved and miniaturized analytical strategy based on μ-quechers for isolation of polyphenols. a powerful approach for quality control of baby foods. *Microchemical Journal*, Elsevier, v. 139, p. 110–118, 2018.

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

[43] MARTINS, M. L. et al. Microextração líquido-líquido dispersiva (dllme) fundamentos e aplicações. 2012.

[44] ZHAO, X.-E. et al. Sensitive and accurate determination of neurotransmitters from in vivo rat brain microdialysate of parkinson's disease using in situ ultrasound-assisted derivatization dispersive liquid–liquid microextraction by uhplc-ms/ms. *RSC advances*, Royal Society of Chemistry, v. 6, n. 110, p. 108635–108644, 2016.

[45] CAMPILLO, N. et al. Glyoxal and methylglyoxal determination in urine by surfactant-assisted dispersive liquid–liquid microextraction and lc. *Bioanalysis*, Future Science, v. 9, n. 4, p. 369–379, 2017.

[46] HUANG, Y. et al. Three-phase solvent bar liquid-phase microextraction combined with high-performance liquid chromatography to determine sarcosine in human urine. *Journal of separation science*, Wiley Online Library, v. 41, n. 15, p. 3121–3128, 2018.

[47] KUEHNBAUM, N. L.; BRITZ-MCKIBBIN, P. New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chemical reviews*, ACS Publications, v. 113, n. 4, p. 2437–2468, 2013.

[48] ZHANG, Y.; XIE, W.-P. Evaluation of hplc–dad selectivity by discrimination power and mean list length for the identification of unknown drugs. *Chromatographia*, Springer, v. 77, n. 23-24, p. 1613–1622, 2014.

[49] PRAGST, F.; HERZLER, M.; ERXLEBEN, B.-T. Systematic toxicological analysis by high-performance liquid chromatography with diode array detection (hplc-dad). *Clinical Chemistry and Laboratory Medicine (CCLM)*, De Gruyter, v. 42, n. 11, p. 1325–1340, 2004.

[50] PORTER, S. E. et al. Analysis of four-way two-dimensional liquid chromatography-diode array data: application to metabolomics. *Analytical chemistry*, ACS Publications, v. 78, n. 15, p. 5559–5569, 2006.

[51] ORTIZ-VILLANUEVA, E. et al. Chemometric evaluation of hydrophilic interaction liquid chromatography stationary phases: resolving complex mixtures of metabolites. *Analytical Methods*, Royal Society of Chemistry, v. 9, n. 5, p. 774–785, 2017.

[52] JIMÉNEZ-CARVELO, A. M. et al. Classification of olive oils according to their cultivars based on second-order data using lc-dad. *Talanta*, Elsevier, v. 195, p. 69–76, 2019.

[53] BELTRÁN, N. et al. Feature extraction and classification of chilean wines. *Journal of Food Engineering*, Elsevier, v. 75, n. 1, p. 1–10, 2006.

[54] LUCA, S. D. et al. Simultaneous quantification of caffeine and chlorogenic acid in coffee green beans and varietal classification of the samples by hplc-dad coupled with chemometrics. *Environmental Science and Pollution Research*, Springer, v. 25, n. 29, p. 28748–28759, 2018.

[55] BOCCARD, J.; RUDAZ, S. 4.20 - analysis of metabolomics data—a chemometrics perspective. In: BROWN, S.; TAULER, R.; WALCZAK, B. (Ed.). *Comprehensive Chemometrics (Second Edition)*. Second edition. Oxford: Elsevier, 2020. p. 483 – 505. ISBN 978-0-444-64166-3.

[56] ESBENSEN, K.; GELADI, P. 2.02 - principal component analysis: Concept, geometrical interpretation, mathematical background, algorithms, history, practice. In: BROWN, S.; TAULER, R.; WALCZAK, B. (Ed.). *Comprehensive Chemometrics (Second Edition)*. Second edition. Oxford: Elsevier, 2009. p. 3 – 15.

[57] GELADI, P.; LINDERHOLM, J. 2.03 - principal component analysis. In: BROWN, S.; TAULER, R.; WALCZAK, B. (Ed.). *Comprehensive Chemometrics (Second Edition)*. Second edition. Oxford: Elsevier, 2020. p. 17 – 37.

[58] HUBERT, M.; ROUSSEEUW, P. J.; BRANDEN, K. V. Robpca: a new approach to robust principal component analysis. *Technometrics*, Taylor & Francis, v. 47, n. 1, p. 64–79, 2005.

[59] HUBERT, M. Robust methods for high-dimensional data. In: ELSEVIER. *Comprehensive Chemometrics [Recurso electrónico]: chemical and biochemical data analysis. Volume 1*. [S.l.], 2020. p. 149–171.

[60] BRO, R.; SMILDE, A. K. Principal component analysis. *Analytical methods*, Royal Society of Chemistry, v. 6, n. 9, p. 2812–2831, 2014.

[61] BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, Wiley Online Library, v. 28, n. 4, p. 213–225, 2014.

[62] BARKER, M.; RAYENS, W. Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*, Wiley Online Library, v. 17, n. 3, p. 166–173, 2003.

[63] FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

[64] PARK, C. H.; PARK, H. A comparison of generalized linear discriminant analysis algorithms. *Pattern Recognition*, v. 41, n. 3, p. 1083 – 1097, 2008. Part Special issue: Feature Generation and Machine Learning for Robust Multimodal Biometrics.

[65] LAVINE, B.; WHITE, C. G.; DAVIDSON, C. E. 3.33 - genetic algorithms for variable selection and pattern recognition. In: BROWN, S.; TAULER, R.; WALCZAK, B. (Ed.). *Comprehensive Chemometrics (Second Edition)*. Second edition. Oxford: Elsevier, 2020. p. 673 – 700.

[66] FILHO, P. A. da C.; POPPI, R. J. Algoritmo genético em química. *Química Nova*, v. 22, n. 3, p. 405, 1999.

[67] BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. part 1: linear models. pls-da. *Analytical Methods*, Royal Society of Chemistry, v. 5, n. 16, p. 3790–3798, 2013.

[68] FILHO, G. P. et al. Residi: Um sistema de decisao inteligente para infraestruturas residenciais via sensores e atuadores sem fio.

[69] KOTU, V. *Model Evaluation in Data Science (ed. Kotu, V. & Deshpande, B) 263–279*. [S.l.]: Morgan Kaufmann, 2019.

[70] WASSAN, J. T.; ZHENG, H. et al. Measurements of accuracy in biostatistics. In: *Encyclopedia of Bioinformatics and Computational Biology*. [S.l.]: Elsevier, 2018. p. 685–690.

[71] LEE, J. Y. Uses of clinical databases. *The American journal of the medical sciences*, Elsevier, v. 308, n. 1, p. 58–62, 1994.

[72] MARSHALL, J.; CHAHIN, A.; RUSH, B. Review of clinical databases. In: *Secondary Analysis of Electronic Health Records*. [S.l.]: Springer, 2016. p. 9–16.

[73] IAVINDRASANA, J. et al. Clinical data mining: a review. *Yearbook of medical informatics*, Georg Thieme Verlag KG, v. 18, n. 01, p. 121–133, 2009.

[74] RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

[75] HUSSON, F. et al. Imputation of mixed data with multilevel singular value decomposition. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 28, n. 3, p. 552–566, 2019.

[76] MENG, F.; CAI, C.; YAN, H. A bicluster-based bayesian principal component analysis method for microarray missing value estimation. *IEEE journal of biomedical and health informatics*, IEEE, v. 18, n. 3, p. 863–871, 2013.

[77] BERETTA, L.; SANTANIELLO, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, BioMed Central, v. 16, n. 3, p. 197–208, 2016.

[78] KIM, K.-Y.; KIM, B.-J.; YI, G.-S. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, Springer, v. 5, n. 1, p. 1–9, 2004.

[79] ABAYOMI, K.; GELMAN, A.; LEVY, M. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 57, n. 3, p. 273–291, 2008.

[80] CALIXTO, E. *Gas and oil reliability engineering: modeling and analysis*. [S.l.]: Gulf Professional Publishing, 2016.

[81] WIKIPéDIA. *Teste Kolmogorov-Smirnov — Wikipédia, a enciclopédia livre*. 2020. [Online; accessed 24-junho-2020]. Available at: <`https://pt.wikipedia.org/w/index.php?title=Teste_Kolmogorov-Smirnov&oldid=58590781`>.

[82] KIM, J.-W.; PACHEPSKY, Y. A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for swat streamflow simulation. *Journal of hydrology*, Elsevier, v. 394, n. 3-4, p. 305–314, 2010.

[83] NGUYEN, C. D.; CARLIN, J. B.; LEE, K. J. Diagnosing problems with imputation models using the kolmogorov-smirnov test: a simulation study. *BMC medical research methodology*, BioMed Central, v. 13, n. 1, p. 1–9, 2013.

[84] LIU, Y.; DE, A. Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International journal of statistics in medical research*, NIH Public Access, v. 4, n. 3, p. 287, 2015.

[85] SCHLEGEL, P. N. et al. Diagnosis and treatment of infertility in men: Aua/asrm guideline - part 1. *Fertility and Sterility*, v. 115, n. 1, p. 54–61, 2021. ISSN 0015-0282.

[86] LITTLE, R. J. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, Taylor & Francis, v. 83, n. 404, p. 1198–1202, 1988.

[87] TAHA, E. A.; WAHED, S. R. A.; MOSTAFA, T. Varicocele impact on testicular size of infertile men in unilateral or bilateral associated cases. *Human Andrology*, LWW, v. 1, n. 3, p. 76–78, 2011.

[88] BERG, W. T. et al. Thresholds for testicular size discrepancy in fertile men with and without varicocele. *Fertility and Sterility*, Elsevier, v. 114, n. 3, p. e95–e96, 2020.

[89] BELLURKAR, A. et al. Role of testicular size as a parameter for predicting infertility in indian males. *Journal of Human Reproductive Sciences*, Wolters Kluwer–Medknow Publications, v. 13, n. 2, p. 114, 2020.

[90] CHECK, J. et al. Evaluation of sperm morphology using kruger's strict criteria. *Archives of andrology*, Taylor & Francis, v. 28, n. 1, p. 15–17, 1992.

[91] NOVAIVFFERTILITY. *How To Read A Sperm Analysis Report*. 2017. [Online; accessed 31-junho-2021]. Available at: <`https://www.novaivffertility.com/blog/understanding-semen-analysis-report/`>.

[92] ORGANIZATION, W. H. *World health statistics 2010*. [S.l.]: World Health Organization, 2010.

[93] HEALTHJADE. *Sperm motility*. 2019. [Online; accessed 31-junho-2021]. Available at: <`https://healthjade.net/sperm-motility/`>.

[94] WDOWIAK, A. et al. Levels of fsh, lh and testosterone, and sperm dna fragmentation. *Neuroendocrinol Lett*, v. 35, n. 1, p. 73–79, 2014.

[95] BABU, S. R. et al. Evaluation of fsh, lh and testosterone levels in different subgroups of infertile males. *Indian Journal of Clinical Biochemistry*, Springer, v. 19, n. 1, p. 45–49, 2004.

[96] MADBOULY, K. et al. Clinical, endocrinological and histopathological patterns of infertile saudi men subjected to testicular biopsy: a retrospective study from a single center. *Urology annals*, Wolters Kluwer–Medknow Publications, v. 4, n. 3, p. 166, 2012.

[97] MIĆIĆ, S. et al. Seminal plasma hormone profile in infertile men with and without varicocele. *Archives of andrology*, Taylor & Francis, v. 17, n. 3, p. 173–178, 1986.

[98] ALI, M.; PAREKH, N. Male age and andropause. In: *Male Infertility*. [S.l.]: Springer, 2020. p. 469–477.

[99] RAJENDER, S. et al. Thyroid, spermatogenesis, and male infertility. *Front Biosci (Elite Ed)*, v. 3, p. 843–55, 2011.

[100] VIGNERA, S. L.; VITA, R. Thyroid dysfunction and semen quality. *International journal of immunopathology and pharmacology*, SAGE Publications Sage UK: London, England, v. 32, p. 2058738418775241, 2018.

[101] GHABILI, K. et al. Hypothesis: intracellular acidification contributes to infertility in varicocele. *Fertility and sterility*, Elsevier, v. 92, n. 1, p. 399–401, 2009.

[102] BOERI, L. et al. Serum albumin levels are associated with sex steroids hormones and sperm concentration impairment in primary infertile men–results of a cross-sectional study. *European Urology Supplements*, Elsevier, v. 18, n. 1, p. e325, 2019.

[103] ANDO, S. et al. Physiopathologic aspects of leydig cell function in varicocele patients. *Journal of andrology*, Wiley Online Library, v. 5, n. 3, p. 163–169, 1984.

[104] OHLANDER, S. J.; LINDGREN, M. C.; LIPSHULTZ, L. I. Testosterone and male infertility. *Urologic Clinics*, Elsevier, v. 43, n. 2, p. 195–202, 2016.

APÊNDICE A - TABELA INICIAL DOS DADOS CLÍNICOS

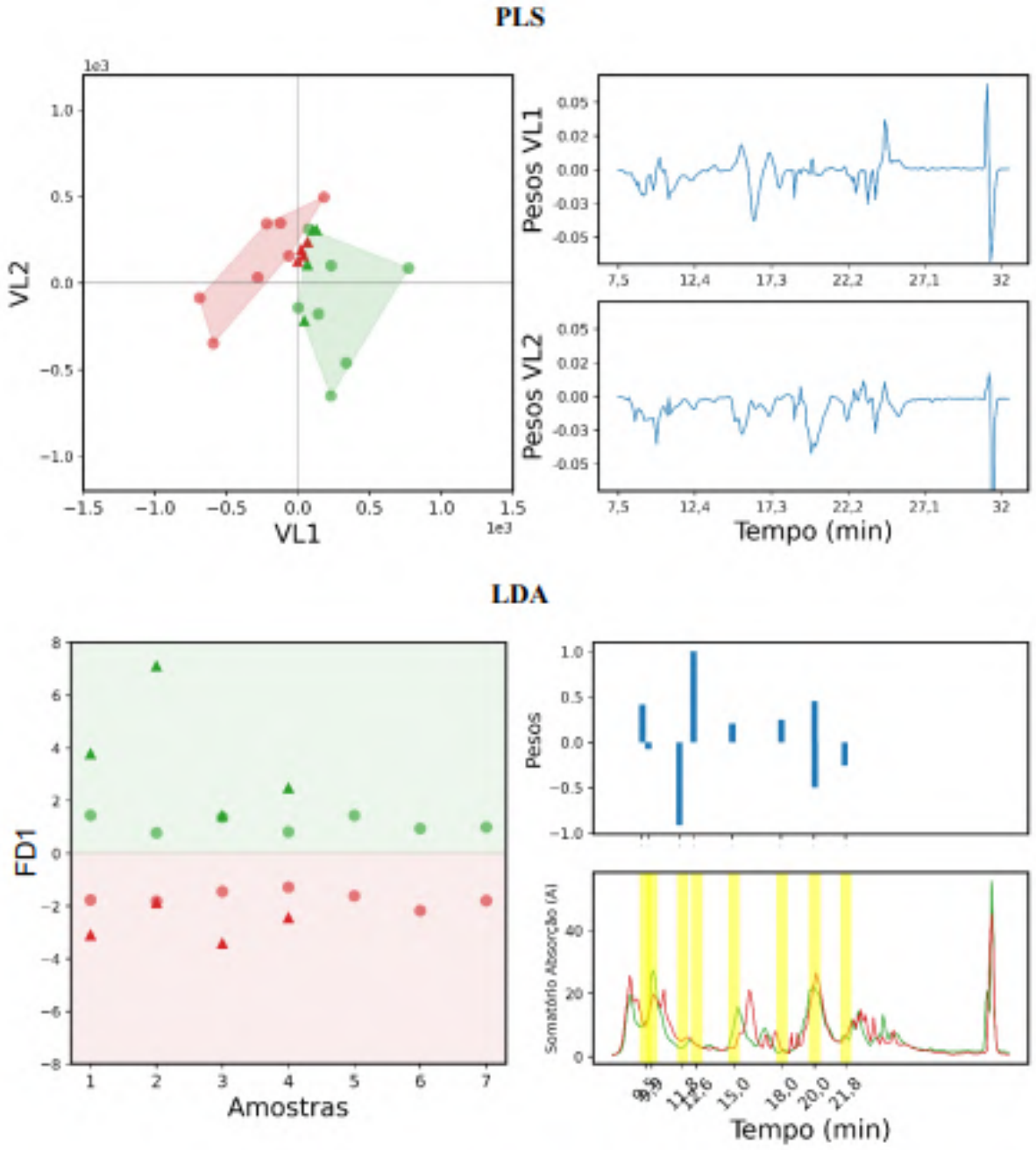
Classe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30*
C	42	NaN	31	0	0	NaN	NaN	14	42,8	34,24	35	11,984	NaN	0,8	7,5	227	5,34	22,1	4,6	2,5	24,6	4	1,42	1,04	13,5	12,4	NaN	0	NaN	0
C	32	NaN	25,1	0	0	NaN	NaN	11	38,2	183,36	45	82,512	20	4,8	8	271	6,9	64,6	2,9	NaN	17,8	4,6	1,51	1,21	12	10,7	0,14	1	0,19	1
C	34	NaN	26,7	0	0	NaN	NaN	15	7,2	56,16	90	50,544	2	7,8	8	329	6,5	61,2	5,5	3	30,3	4,7	0,78	1,14	19,8	10	0,18	0	0,15	0
C	35	NaN	29,2	0	0	NaN	NaN	15	19	57	61	34,77	NaN	3	8	434	10,2	29,9	10,6	3,8	22,7	4,8	1,02	1,17	16,8	15,3	0,21	1	0,29	1
C	35	NaN	26,8	0	0	NaN	NaN	13	115,4	577	85	490,45	10	5	7,5	479	10,8	47,1	3	3,1	25,6	4,8	0,78	1,08	18,7	17,5	NaN	0	NaN	0
C	38	28	37,6	0	0	25	25	14	53	212	70	148,4	9	4	7,5	187	4,7	29	2,6	2,9	15,5	4,8	1,94	0,93	16	13,6	NaN	0	0,21	1
C	38	32	30,2	0	0	18	18	14	137,4	439,68	55	241,824	6	3,2	8,5	192	5,1	24,8	3,9	2,9	15,4	4,5	2,39	1,07	9,3	8,6	NaN	1	NaN	1
C	38	34	29,3	0	0	25	25	14	108,8	369,92	85	314,432	7	3,4	8	256	6,29	23,4	2,4	2,3	18,6	4,7	1,11	0,89	17,4	15	NaN	NaN	0,21	1
C	36	34	28,3	0	0	25	25	12	343,4	1304,92	97	1265,7724	3	3,8	8,5	283	6,64	19,6	5,1	3,1	22,4	4,5	1,39	0,92	17,7	16	0,13	1	0,16	1
C	38	36	28,3	0	0	25	25	13	59	342,2	70	239,54	3	5,8	8	570	10,2	22	2,6	4,1	40,4	4,8	2,16	1,26	20	19	0,2	1	0,2	1
C	39	30	29,7	0	0	25	25	15	245	367,5	85	312,375	8	1,5	7	248	7,06	35	5,8	1,35	10,7	4,9	2,5	1,35	18,1	15,3	0,19	0	0,3	1
C	32	22	36,4	0	0	15	15	14	16	80	70	56	18	5	8	303,9	6,29	41	1,6	1,9	28,4	4,6	1,13	1,32	9,8	9,8	0,9	0	0,8	0
C	27	26	28,9	0	0	25	25	17	18	54	70	37,8	21	3	8	316	10,2	40,4	2,5	4	16,1	3,5	0,89	0,91	17,1	20	0,17	1	0,23	1
C	27	28	21,5	0	0	25	25	14	39,6	158,4	55	87,12	NaN	4	9	359	7,14	21	4,7	5,5	29,1	5	0,82	1,47	14,7	16	0,14	0	0,14	0
C	31	25	23,8	0	0	18	18	17	15	75	60	45	18	5	7,5	402	8,98	27,6	6,1	3,2	25,4	4,7	2,69	1,03	15,6	13,5	0,18	0	0,18	0
C	41	28	26,7	0	0	35	35	14	19,4	93,12	45	41,904	4	4,8	8,5	473	9,73	39	5	5,9	32,5	4,5	1,34	1,02	16,9	15,2	0,16	0	0,13	1
C	32	27	25,7	0	0	15	12	13	18	90	30	27	11	5	8	446	6,5	22,2	3,7	3,2	54,3	4,5	1,29	1,15	12,6	11,5	0,2	1	0,2	1
C	31	33	33,6	0	0	30	30	13	194,2	815,64	92	750,3888	4	4,2	8,5	201	4,8	32,8	NaN	NaN	20,1	4,4	1,32	1,26	22	22	NaN	0	NaN	0
C	32	30	27,4	0	0	25	25	12	390	1053	64	673,92	16	2,7	7,8	753	14	35,7	0,9	3,03	46	4,3	1,75	1,39	25,2	27,1	NaN	NaN	0,15	1
C	36	33	23,5	0	0	20	15	15	8	24	90	21,6	NaN	3	7,5	229,3	5,74	31	2,7	3,3	15,4	5	2,73	1,06	19	14,2	0,14	0	0,18	1
C	28	27	27,9	0	0	20	20	15	85,6	256,8	75	192,6	9	3	6	362,6	8,79	33	4,6	5	19,9	4,8	1,1	0,93	12,6	11,1	0,2	1	0,24	1
C	32	25	24,2	0	0	20	20	15	18	108	60	64,8	14	6	8	432	6,11	46	4,7	2,1	53,7	4,9	1,76	1,16	15,2	14,6	NaN	0	NaN	0
C	38	31	28,5	0	0	20	15	16	10	20	40	8	12	2	7	276	6,07	39,6	4,7	6,9	23	4,9	2,16	0,88	12,5	9,6	0,16	0	0,21	0
C	29	26	21,4	0	0	20	20	16	10,6	53	40	21,2	5	5	8	267	5,1	26	2,5	3,5	32,2	4,5	2,4	1,18	11,6	11,6	0,15	1	0,21	1
VF	34	31	25,5	1	0	25	25	16	50	50	85	42,5	NaN	1	8	231	5,2	38,3	3,6	3,9	22,3	4,7	2	0,83	15,3	12,2	0,16	1	0,32	1
VF	30	NaN	33,7	1	0	18	18	15	85	170	40	68	NaN	2	7,5	495	12	21,3	3,5	3,1	22,6	4,7	1,23	1,32	16,8	12	0,19	0	0,32	1
VF	37	37	34	1	0	20	15	16	7,6	22,8	40	9,12	5	3	7	222	6,1	35,2	5,3	5,4	13,1	4,7	1,58	1,12	12,8	10,2	0,26	1	0,27	1
VF	40	37	30	1	0	20	25	16	12	42	25	10,5	6	3,5	7	468	NaN	33,4	4,2	5,1	NaN	4,6	NaN	NaN	14,9	16,4	0,23	0	0,3	1
VF	33	33	28,7	1	0	15	15	12	26	117	47	54,99	8	4,5	8	302	5,89	85	3,8	2,5	28,6	5,1	NaN	NaN	10,8	9,2	NaN	0	0,21	1
VF	35	29	27,1	2	0	20	20	15	27,4	109,6	80	87,68	12	4	7,5	343,7	6	92	1,6	3	38,7	4,6	2,63	0,91	14,7	11,3	NaN	NaN	0,25	1
VF	43	34	27,1	2	0	15	15	15	383	612,8	95	582,16	5	1,6	8	246	4,6	32	3,9	2,9	34,2	4,4	1,37	0,96	11,4	8,7	0,19	1	0,26	1
VF	33	32	25,7	2	0	20	15	13	60,2	210,7	70	147,49	9	3,5	7	347	7,2	24,6	3,4	2,9	26,9	4,9	1,15	1,15	11	11,5	0,2	0	0,42	1
VF	33	36	26,1	2	0	25	25	16	98,8	494	55	271,7	11	5	7,5	322	6,1	55	2,3	3,5	31,7	4,8	1,85	1,08	18,3	15	0,26	1	0,3	1
VF	32	32	19	2	0	20	15	14	78,8	338,84	80	271,072	1	4,3	8	620	9,1	28,9	2,5	3,5	55,9	4,7	2,13	1,26	17,7	9,8	0,08	0	0,39	1
VF	36	29	29,8	2	0	20	20	12	70	105	61	64,05	7	1,5	7	354	8,39	31	11,2	4,9	22,3	4,6	0,97	1,22	15,6	9,9	0,2	1	0,27	1
VF	34	27	21	3	0	15	8	14	40	112	55	61,6	NaN	2,8	7,5	401	8,4	37,5	11,1	5,3	27,1	4,9	0,68	0,98	10,1	8,5	0,14	1	0,35	1
VF	36	33	28,7	3	0	20	15	19	21	58,8	90	52,92	NaN	2,8	8	243	6,2	25,1	2,4	1,1	18,5	1,4	NaN	NaN	16,2	10,4	0,1	0	0,47	1
VF	41	NaN	NaN	3	0	25	25	NaN	88	220	57	125,4	14	2,5	7	779	12	49,6	3,18	6,22	56,9	4,6	1,53	1,34	29,2	25,7	0,18	1	0,27	1
VF	24	NaN	26,27	3	0	15	10	12	53,6	160,8	65	104,52	9	3	8	436	9,8	56,4	3,6	3,4	27,6	4,4	0,46	1,06	NaN	NaN	NaN	NaN	NaN	NaN
VF	31	38	28,7	3	0	20	15	13	36	72	32	23,04	10	2	8	510	10,9	36,5	8,9	8,1	31,2	4,5	0,53	0,93	13,8	10	0,27	1	0,28	1
VF	39	34	27	3	0	15	10	16	5	15	33	4,95	6	3	8	529	10,9	15,7	15,3	3,8	31,1	4,8	NaN	NaN	9,2	6	0,34	1	0,43	1
VF	26	22	20,4	3	0	20	20	13	30	75	38	28,5	9	2,5	8	658	12,5	43,5	4,3	4,9	39,9	4,6	3,47	1,4	20,7	18	0,25	1	0,34	1
VF	28	27	26,1	2	1	12	15	16	12,4	37,2	40	14,88	11	3	7	247	5,2	27	2,6	1,4	21,1	5,4	0,95	1,3	9,2	7,8	0,21	1	0,32	1
VF	35	38	21,5	2	1	20	20	16	40	100	24	24	9	2,5	7	601	10,8	27,2	3,7	2,8	45,1	4,7	1,39	1,08	10,7	10,9	NaN	NaN	NaN	1
VF	27	NaN	20,7	3	2	20	20	16	14	70	50	35	15	5	7,5	503	9,9	34,9	1,12	3,6	3,4	4,7	0,57	1,12	16,1	12,4	0,23	1	0,37	1
VI	28	28	25,2	1	0	15	15	13	9	18	26	4,68	7	2	8	872	10,1	45,4	4	4,62	89,8	3,7	0	2,37	15,2	9,2	0,33	1	3,4	1
VI	34	32	34	1	0	25	25	13	8	16	35	5,6	6	2	7	126,4	3,53	35,9	4,5	1,44	12,6	4,5	NaN	NaN	11,5	12,8	0,17	1	0,24	1
VI	24	31	30,6	2	0	20	20	12	2	5	40	2	9	2,5	8	472	11,9	45	23	9,6	21,6	4,5	9	9,1	9	9,1	0,25	1	0,36	1
VI	23	19	21,2	2	0	10	5	15	0,5	2,5	5	0,125	3	5	6,5	220	3,1	31	8,1	2,36	47,3	5,2	1,76	1,35	6,5	6,9	NaN	NaN	0,46	1

Classe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30*
VI	42	34	23,3	2	0	5	5	15	NaN	NaN	NaN	NaN	0	2	7	530	9,29	28	15,5	5,5	39,9	5	4,84	1,11	9,1	6,1	0,24	1	0,4	1
VI	37	34	37,7	2	0	20	10	14	35	31,5	26	8,19	10	0,9	7	311	7,26	22,2	6,1	6	24,4	4,3	2,65	1,16	15,3	10	0,18	0	0,4	1
VI	32	37	26,8	2	0	15	10	16	2	3	10	0,3	2	1,5	8	742	13	52,5	9,1	7,1	47,5	4,5	1,41	1,06	15,5	9,3	NaN	NaN	0,46	1
VI	38	34	20,6	2	0	20	15	12	NaN	NaN	NaN	NaN	NaN	4,5	8	373	7,8	38,5	11,6	3,1	28,5	4,7	NaN	NaN	12,1	10,8	0,25	1	0,37	1
VI	26	24	24,6	3	0	20	20	14	48	192	46	88,32	17	4	7	449	8,17	39,8	3,2	4,3	38	4,7	1,35	1,23	11,1	11,6	NaN	NaN	0,3	1
VI	29	29	NaN	3	0	20	15	14	11	71,5	34	24,31	5	6,5	7	393	6,22	29	4,2	2,9	49,2	5,2	1,1	1,29	18,1	17,2	0,25	1	0,47	1
VI	34	32	24,69	3	0	20	15	12	0,8	2,8	43	1,204	4	3,5	8	756	11,1	2,9	5,2	6,8	58	4,9	3,3	1,4	12,6	9,4	0,22	1	0,25	1
VI	43	42	20,2	3	0	20	20	11	7	31,5	20	6,3	6	4,5	8	518	9,32	20,6	6	2,4	42,8	4,3	1,66	1,18	15,1	21,1	0,73	1	0,78	1
VI	29	28	26	3	0	15	10	13	13	19,5	36	7,02	5	1,5	8	276	7,64	20,4	5,7	4,18	14,8	4,5	3,02	1,27	19,7	19,5	0,29	1	0,38	1
VI	29	31	25,25	3	0	12	6	14	5	12,5	15	1,875	6	2,5	8	604	10,1	3,5	3,5	8	45	4,9	1,7	1,3	13,1	9	0,28	1	0,3	1
VI	32	31	24	1	1	12	10	14	8	48	35	16,8	10	6	7	585	11,1	38	3,5	3,8	38,6	4,6	NaN	NaN	12	10	0,33	1	0,31	1
VI	30	29	24	2	1	15	15	14	NaN	NaN	NaN	NaN	NaN	4	7	375	5,24	28,4	8,5	5,5	58,2	4,5	NaN	NaN	8,2	8,7	0,41	1	0,4	1
VI	37	35	29,7	2	1	25	15	15	24	96	40	38,4	12	4	7	391	8,08	23,3	4,53	7,96	30,8	4,5	0,9	1,06	20	14	0,24	1	0,4	1
VI	37	36	35,1	2	1	15	15	13	3	4,5	30	1,35	10	1,5	7,5	293	5,2	41,8	6,3	4,3	38,6	4,4	0,63	0,9	10,4	11,8	0,37	1	0,48	1
VI	22	24	NaN	3	1	15	15	15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	533	10,2	19,4	1,95	3,22	33,7	5,1	2,96	1,39	15,1	12,6	0,25	1	0,39	1
VI	28	27	24,4	3	1	12	10	12	6	36	64	23,04	4	6	8	563	8,8	32,6	6,59	3,6	50,2	4,7	0,94	1,18	10,7	10,9	0,4	1	0,41	1
VI	32	27	23,9	3	1	15	7	16	4	8	5	0,4	3	2	7	314	NaN	39	5,2	3,03	NaN	4,6	1,73	0,72	12,9	10,1	0,24	1	0,34	1
VI	40	32	29,4	3	1	15	12	16	10	35	24	8,4	10	3,5	7,5	296	5,52	37,6	9,1	4,8	32,4	4,9	0,82	1,21	10	9,8	0,2	1	0,24	1
VI	36	28	33,6	3	1	8	12	15	16	32	48	15,36	6	2	8	422	7,09	32,2	6,61	3,99	44,6	4,4	0,74	1,2	7,5	10	0,27	1	0,43	1
VI	21	23	21,7	3	1	15	10	12	5	20	11	2	2	4	7	752	16,9	30	5,6	4,7	28	5	2,35	1,21	8,2	8,7	0,19	1	0,43	1
VI	34	NaN	29,9	1	2	15	10	12	16	24	30	7,2	14	1,5	7,5	343	6,3	20,8	8,2	5,5	33,6	5	1,76	1,22	7,2	6	0,3	1	0,38	1
VI	28	29	26,1	3	2	30	30	14	0,4	0,8	20	0,16	10	2	8,5	597	9,6	60,8	3,3	2,8	48,5	4,8	0,76	0,97	27	27	0,3	1	0,29	1
VI	24	32	21,7	3	2	20	15	12	7	31,5	55	17,325	12	4,5	7	575	11,6	50	3,3	2,6	35,3	4,5	0,6	0,8	12,5	11,5	0,3	1	0,32	1
VI	33	25	30	3	2	15	10	12	NaN	NaN	NaN	NaN	NaN	1,5	7	314	6,5	39	10,4	3,9	28,3	4,6	1,34	1,16	9,6	9,3	0,23	1	0,29	1
VI	39	35	29,9	3	2	20	10	14	23	138	29	40,02	13	6	7	337,2	8,01	34	2,9	2,5	22,6	4,5	1,58	1,27	18,6	12	0,26	1	0,32	1
VI	34	41	26,4	3	2	15	10	16	NaN	NaN	NaN	NaN	NaN	4,5	7,5	700	13,4	35	16,1	7,3	40,9	4,5	2,81	1,06	12,1	8,5	0,25	1	0,33	1
VI	26	30	19,1	3	2	12	12	16	17	13,6	23	3,128	12	0,8	7	760	17,5	40	8,4	7,8	28,3	4,8	0,78	1,08	7,2	7,8	0,13	1	0,29	1
VI	33	34	32,6	3	2	15	10	13	21	27,3	38	10,374	16	1,3	8	418	10,4	2,7	4,3	3,5	23	4,3	1,6	0,65	16,9	13,4	0,28	1	0,29	1
VI	34	32	26,5	3	2	10	5	15	NaN	NaN	NaN	NaN	NaN	4	8	489	NaN	NaN	22,7	7,1	NaN	NaN	NaN	NaN	9,3	8,4	0,2	1	0,24	1
VI	27	23	15,4	1	3	10	10	14	NaN	NaN	NaN	NaN	NaN	3	7	299	3,4	21,1	46,2	11,2	46,2	4,6	2,53	1,08	11,3	9,7	0,47	1	0,6	1
VI	45	37	21	1	3	5	5	14	0,4	1,6	10	0,16	5	4	7	312	5,74	36,9	11,4	3,9	36	4,5	2,88	1,32	6,6	6,3	0,35	1	0,41	1

*

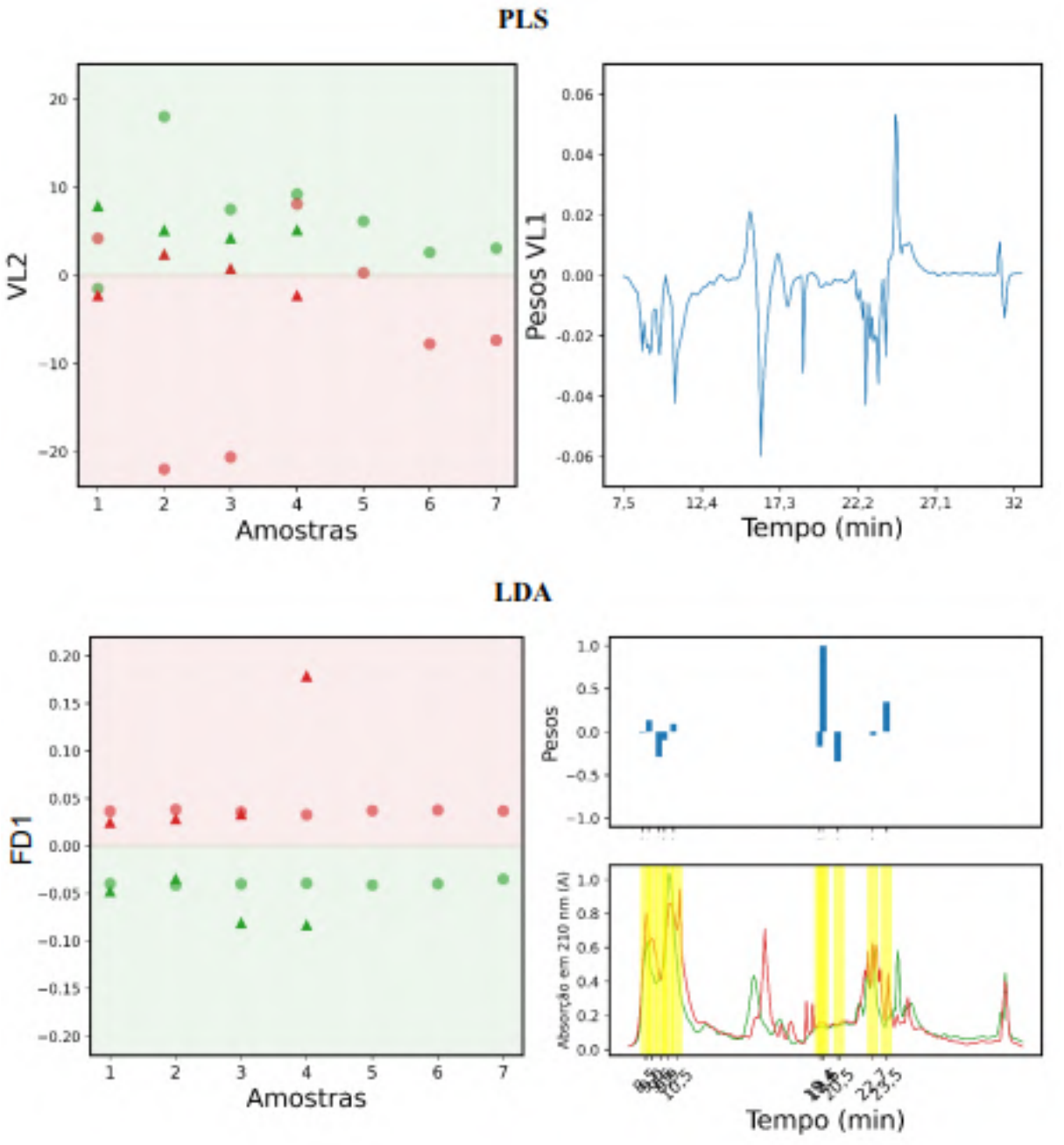
1 Idade Paciente
2 Idade Parceira
3 Índice de massa corpórea (IMC)
4 Grau Varicocele Esquerda
5 Grau Varicocele Direita
6 Tamanho testículo direito
7 Tamanho testículo esquerdo
8 Idade Puberdade
9 Concentração
10 Contagem Total
11 Motilidade Progressiva
12 Total de espermatozóides progressivos
13 Kruger
14 Volume
15 pH
16 Testosterona
17 Testosterona Livre
18 Estradiol
19 Hormônio folículo-estimulante (FSH)
20 Hormônio luteinizante (LH)
21 Globulina ligadora de hormônios sexuais (SHBG)
22 Albumina
23 Hormônio estimulante da tireoide (TSH)
24 T4 Livre
25 Ultrassonografia Testículo Direito
26 Ultrassonografia Testículo Esquerdo
27 Ultrassonografia varicocele Direita (Tamanho)
28 Ultrassonografia varicocele Direita (Refluxo)
29 Ultrassonografia varicocele esquerda (Tamanho)
30 Ultrassonografia varicocele esquerda (Refluxo)

APÊNDICE B – GRÁFICOS DE ESCORES E PESOS PARA OS MODELOS CLASSIFICATÓRIO UTILIZANDO AS CLASSES VF E VI PARA O SOMATÓRIO DOS DADOS CROMATOGRÁFICOS



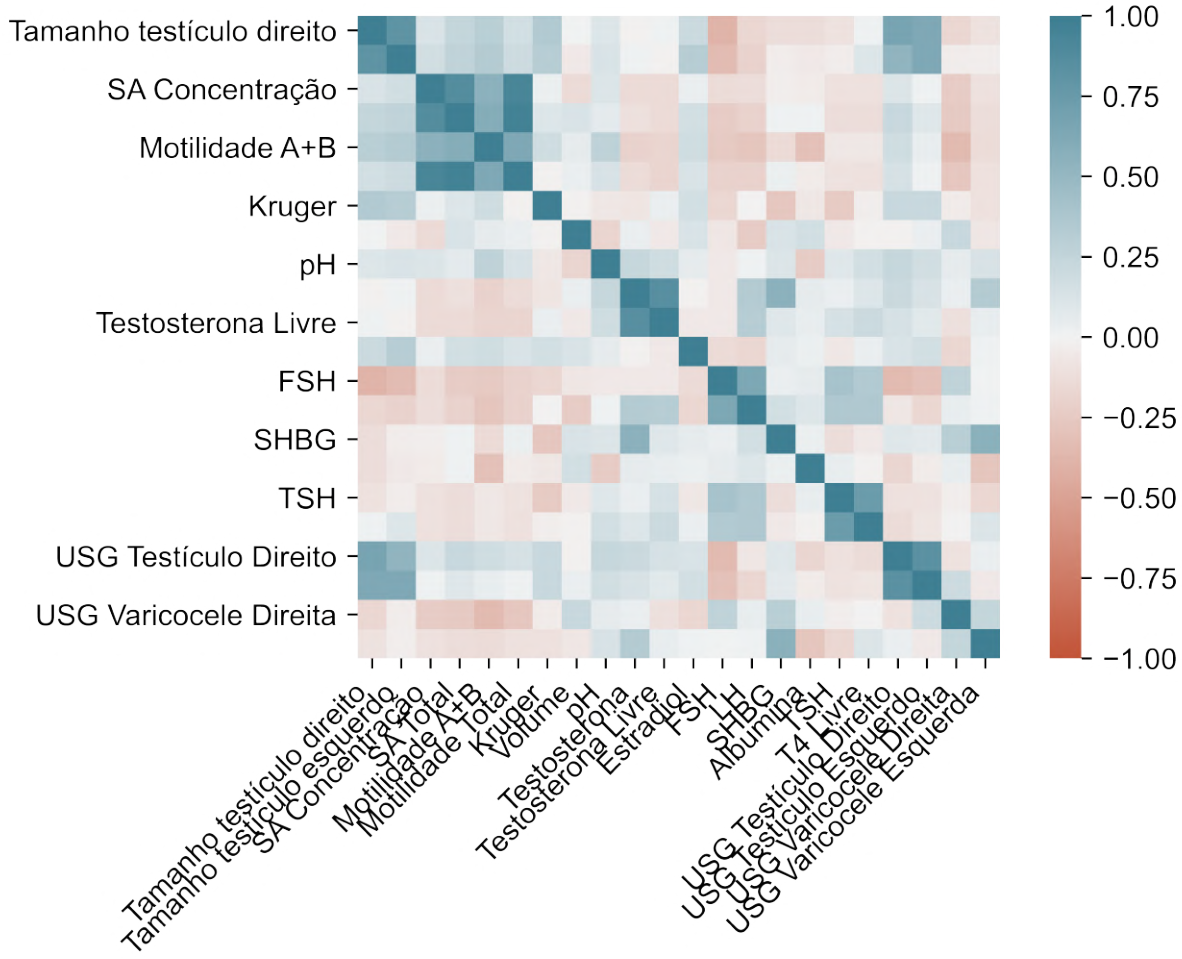
Fonte: O autor (2021).

APÊNDICE C – GRÁFICOS DE ESCORES E PESOS PARA OS MODELOS CLASSIFICATÓRIO UTILIZANDO AS CLASSES VF E VI EM 210 nm DOS DADOS CROMATOGRÁFICOS



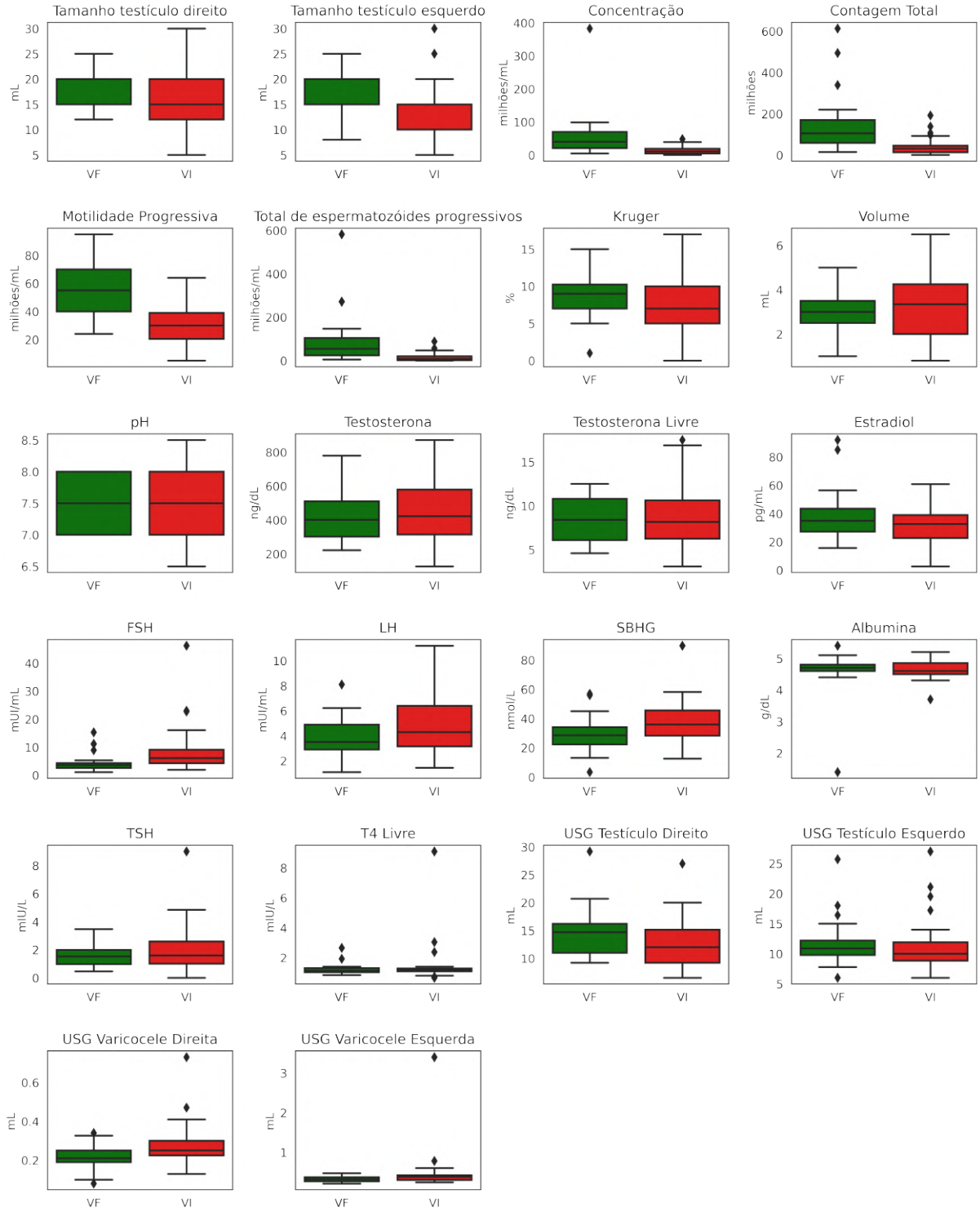
Fonte: O autor (2021).

APÊNDICE D – HEATMAP DE CORRELAÇÃO DOS DADOS CLÍNICOS



Fonte: O autor (2021).

APÊNDICE E – BOXPLOTS DE TODAS AS VARIÁVEIS DO CONJUNTO DE DADOS CLÍNICOS



Fonte: O autor (2021).