

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Física

Vinícius Aguiar Cardoso Godinho Baliza

**ESPECTROSCOPIA DE APRENDIZADO DE
MÁQUINA APLICADA AO DIAGNÓSTICO DOS
EFEITOS DA COVID-19 NA REPRODUÇÃO
HUMANA**

Belo Horizonte
2025

Vinícius Aguiar Cardoso Godinho Baliza

**ESPECTROSCOPIA DE APRENDIZADO DE MÁQUINA
APLICADA AO DIAGNÓSTICO DOS EFEITOS DA COVID-19
NA REPRODUÇÃO HUMANA**

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Física.

Orientador: Prof. Dr. Juan Carlos González Pérez

Coorientador: Dr. Paulo Henrique Ribeiro Amaral

Belo Horizonte

2025

Dados Internacionais de Catalogação na Publicação (CIP)

B186e Baliza, Vinícius Aguiar Cardoso Godinho.

Espectroscopia de aprendizado de máquina aplicada ao diagnóstico dos efeitos da covid-19 na reprodução humana / Vinícius Aguiar Cardoso Godinho Baliza. – 2025.
80 f. : il.

Orientador: Juan Carlos González Pérez.

Coorientador: Paulo Henrique Ribeiro Amaral.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Departamento de Física.
Bibliografia: f. 64-75.

1. COVID-19 (Doença). 2. Aprendizado do computador. 3. Reprodução humana. 4. Análise espectral. 5. Inteligência artificial. I. Título. II. González Pérez, Juan Carlos. III. Amaral, Paulo Henrique Ribeiro. IV. Universidade Federal de Minas Gerais, Departamento de Física.

CDU – 535.33(043)

Ficha catalográfica elaborada por Romário Martins Ribeiro – CRB6 3595
Biblioteca Professor Manoel Lopes de Siqueira – Departamento de Física - UFMG



UNIVERSIDADE FEDERAL DE MINAS GERAIS

ATA

ATA DA SESSÃO DE ARGUIÇÃO DA 766ª DISSERTAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA, DEFENDIDA POR VINÍCIUS AGUIAR CARDOSO GODINHO BALIZA orientado pelo professor Juan Carlos González Pérez e coorientado pelo Dr. Paulo Henrique Ribeiro Amaral para obtenção do grau de **MESTRE EM FÍSICA, área de concentração Física**. Às oito horas e trinta minutos do dia quatro de julho de dois mil e vinte e cinco reuniu-se, por videoconferência, a Comissão Examinadora composta pelos professores **Juan Carlos González Pérez** (Orientador - Departamento de Física/UFMG), **Paulo Henrique Ribeiro Amaral** (Coorientador - Departamento de Física/UFMG), **Talita de Oliveira Farias** (Instituto de Ciências Biológicas/UFMG), **Guilherme Mattos Jardim Costa** (Instituto de Ciências Biológicas/UFMG) e **Emilson Ribeiro Viana Junior** (Departamento de Física/UTFPR-Curitiba), para dar cumprimento ao Artigo 37 do Regimento Geral da UFMG, submetendo **VINÍCIUS AGUIAR CARDOSO GODINHO BALIZA** à arguição de seu trabalho de dissertação, que recebeu o título de "**Espectroscopia de Aprendizado de Máquina Aplicada ao Diagnóstico dos Efeitos da Covid-19 na Reprodução Humana**". O candidato fez uma exposição oral de seu trabalho durante aproximadamente 50 minutos. Após esta, os membros da comissão prosseguiram com a sua arguição, e apresentaram seus pareceres individuais sobre o trabalho, concluindo pela aprovação do candidato.

Belo Horizonte, 04 de julho de 2025.

Prof. Juan Carlos González Pérez
Orientador do estudante
Departamento de Física/UFMG

Dr. Paulo Henrique Ribeiro Amaral
Coorientador do estudante
Departamento de Física/UFMG

Dra. Talita de Oliveira Farias
Instituto de Ciências Biológicas/UFMG

Prof. Guilherme Mattos Jardim Costa
Instituto de Ciências Biológicas/UFMG

Prof. Emilson Ribeiro Viana Junior
Departamento de Física/ UTFPR-Curitiba

CANDIDATO: Vinícius Aguiar Cardoso Godinho Baliza



Documento assinado eletronicamente por **Emilson Ribeiro Viana Junior, Usuário Externo**, em 08/07/2025, às 15:40, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Juan Carlos Gonzalez Perez, Professor do Magistério Superior**, em 08/07/2025, às 17:08, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Vinícius Aguiar Cardoso Godinho Baliza, Usuário Externo**, em 08/07/2025, às 17:29, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Paulo Henrique Ribeiro Amaral, Usuário Externo**, em 08/07/2025, às 22:46, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Talita de Oliveira Farias, Usuária Externa**, em 09/07/2025, às 14:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Guilherme Mattos Jardim Costa, Professor do Magistério Superior**, em 16/07/2025, às 16:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4365072** e o código CRC **AD9265DD**.

Agradecimentos

Agradeço ao professor Dr. Juan Carlos González Pérez pela oportunidade e orientação ao longo do desenvolvimento deste trabalho e ao Dr. Paulo Henrique Ribeiro Amaral pela coorientação. Estendo meus agradecimentos ao Victor Hugo e ao João Carlos, que também integraram o nosso grupo de pesquisa. A colaboração de todos foi fundamental pelo suporte, compartilhamento de conhecimentos e divisão das tarefas, tornando possível a realização deste projeto. Também agradeço ao professor Guilherme Matos e ao médico Marcelo Furtado pelo fornecimento das amostras utilizadas neste estudo, bem como à Dra. Maria Ivonete pelo trabalho de medição das amostras. Agradeço, ainda, a todos os demais membros que contribuíram, direta ou indiretamente, com a metodologia aplicada neste estudo.

Sou grato à FAPEMIG pelo financiamento da minha bolsa de pesquisa e à UFMG por ter oferecido o espaço e as oportunidades que pude usufruir desde a graduação até o mestrado.

Agradeço à minha mãe, Ângela, e ao meu irmão, Matheus, pelo apoio, ensinamentos e parceria ao longo de toda a minha vida. Agradeço também ao meu pai, Marcelo, por tudo o que pude aprender com ele. Por fim, expresso a minha gratidão aos meus amigos Evely, Fred, Mel, Renê, Sarah e Vic, que têm sido minha companhia, online e presencial, há mais de 10 anos.

Resumo

Infecções virais, incluindo a COVID-19, causam efeitos adversos no sistema reprodutor masculino e podem levar a problemas de fertilidade, mesmo para pacientes já recuperados da doença. Por outro lado, os métodos de análise do sêmen humano indicados pela Organização Mundial da Saúde (OMS) normalmente não são capazes de identificar diferenças significativas nos parâmetros espermáticos, tais como morfologia, motilidade e quantidade de espermatozoides, de pacientes recuperados da COVID-19 com respeito àqueles que nunca contraíram a doença. Dessa forma, buscamos propor uma metodologia inovadora que permita distinguir amostras seminais de indivíduos que nunca contraíram COVID-19 daqueles que se recuperaram da doença. Para isso, duas abordagens envolvendo aplicação de aprendizado de máquina foram desenvolvidas. A primeira delas contempla uma análise mais profunda dos parâmetros espermáticos de ejaculado de pacientes e a segunda utiliza medidas de espectroscopia, mais especificamente, de elipsometria, para caracterizar as mesmas amostras de ejaculado. Em ambas metodologias foi empregado aprendizado de máquina para realizar a classificação das amostras a partir da análise dos seus dados medidos. Diferentes algoritmos de classificação (KNN, LR, SVM, MLP, DT e XGBoost) foram treinados, otimizados e validados para desempenhar essa tarefa. O melhor resultados alcançado para a primeira abordagem, com o algoritmo SVM, demonstrou um bom desempenho com acurácia de 77%. A partir de um estudo de interpretabilidade, chegamos à conclusão que a motilidade progressiva e a imotilidade dos espermatozoides são os principais parâmetros para diferenciar as amostras de cada classe. O segundo método, com medidas de elipsometria, exibiu uma ótima performance com uma acurácia de 90% nos resultados. Nesse sentido, é notável que a metodologia apresentada aqui funciona, sendo capaz de detectar diferenças nas amostras analisadas e classificá-las com precisão. Portanto, este estudo demonstra a aplicabilidade de uma ferramenta com potencial para ser utilizada por urologistas e centros de reprodução assistida para avaliação de parâmetros espermáticos e sua implicação na infertilidade masculina. Além disso, abrem-se as portas para novas pesquisas que visem aprofundar e evoluir nesta área que ainda é pouco explorada.

Palavras-chave: COVID-19, Aprendizado de máquina, Reprodução humana, Espectroscopia, Inteligência artificial

Abstract

Viral infections, including COVID-19, cause adverse effects on the male reproductive system and may lead to fertility issues, even in patients who have already recovered from the disease. However, the methods for human semen analysis recommended by the World Health Organization (WHO) are usually not able to identify significant differences in the sperm parameters, such as morphology, motility, and sperm count, of patients recovered from COVID-19 compared to those who have never contracted the disease. Therefore, we propose a novel method capable of distinguishing seminal samples from individuals who have never contracted COVID-19 from those who have recovered from the disease. Thus, two approaches involving the application of machine learning were developed. The first one involves a more in-depth analysis of the sperm parameters in patients' ejaculate, while the second adopts spectroscopic measurements, specifically ellipsometry, to further characterize the same semen samples. In both methodologies, machine learning was used to classify the samples employing their measured data. Different classification algorithms (KNN, LR, SVM, MLP, DT e XGBoost) were trained, optimized, and validated to perform this task. The best result achieved with the first approach, using the SVM algorithm, demonstrated good performance, with an accuracy of 77%. By conducting an interpretability evaluation, we concluded that progressive motility and sperm immotility are the main parameters for distinguishing the samples of each class. The second method, using ellipsometry measurements, showed excellent performance with an accuracy of 90%. In this sense, it is evident that the methodology presented here is effective, being capable of detecting differences in the analyzed samples and classifying them with precision. Therefore, this study demonstrates the applicability of a tool with the potential to be used by urologists and fertility clinics to evaluate sperm parameters and their implications for male infertility. Furthermore, it paves the way for additional research to deepen and advance this still underexplored field.

Keywords: COVID-19, Machine learning, Human reproduction, Spectroscopy, Artificial intelligence

Lista de ilustrações

Figura 1 – Gráfico da parte real e imaginária da constante dielétrica complexa em função do logaritmo da frequência angular da luz $\log(\omega)$	20
Figura 2 – Esquema demonstrando a luz incidente, refletida e transmitida ao incidir em um meio.	22
Figura 3 – Diferentes possibilidades para a polarização da luz.	22
Figura 4 – Reflexão de ondas de luz polarizadas p e s.	23
Figura 5 – Representação da incidência, reflexão e transmissão da luz p-polarizada e s-polarizada.	24
Figura 6 – Interferência óptica em um filme fino sobre um substrato.	25
Figura 7 – Modelo óptico mais detalhado para um sistema com estrutura: ambiente/filme fino/substrato.	25
Figura 8 – Modelo óptico para um sistema de multicamadas.	26
Figura 9 – Esquema das medidas de elipsometria.	27
Figura 10 – Esquema do sistema RCE - PSC _{RA}	29
Figura 11 – Esquema do sistema de multicamadas utilizado nas medidas de elipsometria.	31
Figura 12 – Esquema da reflexão difusa na superfície rugosa.	33
Figura 13 – Esquema da reflexão difusa na superfície da fita de mesmo índice de refração do substrato.	33
Figura 14 – Demonstração de uma curva ROC para um AUC de 0,94	38
Figura 15 – Demonstração do processo de validação cruzada com 5 divisões.	39
Figura 16 – Esquema do processo de otimização e validação na validação cruzada K-fold utilizado neste trabalho.	39
Figura 17 – Esquema do processo de otimização e validação na validação cruzada Nested utilizado neste trabalho.	41
Figura 18 – Distribuição dos parâmetros espermáticos	50
Figura 19 – Matriz de correlação para todos os parâmetros espermáticos.	51
Figura 20 – SHAP para os parâmetros espermáticos.	52
Figura 21 – Espectro médio (linha sólida) e desvio padrão (região sombreada) obtidos de Ψ , Δ , dPol e uR para ambas as classes de amostras.	54
Figura 22 – Gráfico do índice de refração em função do comprimento de onda para diferentes ângulos para as lamínulas (a) limpa e (b) com fita mágica Scotch.	56
Figura 23 – Gráfico do índice de refração em função do comprimento de onda para o ângulo de 70° medido com água deionizada em uma montagem vidro/água DI/vidro/fita.	57

Figura 24 – Histograma normalizado das probabilidades	59
Figura 25 – Curva ROC obtida para o modelo SVM + uR com ângulo de incidência de 50°.	60

Lista de tabelas

Tabela 1 – Descrição e vantagens de cada um dos modelos de aprendizado de máquina utilizados neste trabalho.	35
Tabela 2 – Descrição das diferentes categorias de pré-processamento de dados. . .	36
Tabela 3 – Descrição, intervalo do resultado e equação das métricas de avaliação. .	37
Tabela 4 – Valores medianos (intervalo interquartil - IIQ) dos parâmetros semi-nais dos pacientes controle e dos pacientes recuperados de COVID-19, comparado com os valores de referência da OMS.	49
Tabela 5 – Métricas de desempenho e valor ótimo de threshold obtidos para os modelos de classificação dos parâmetros espermáticos.	52
Tabela 6 – Métricas de desempenho e valor ótimo de threshold obtidos com a validação cruzada K-fold para todos os modelos de MLS.	58
Tabela 7 – Métricas de desempenho obtidas com a validação cruzada nested para os modelos de melhor desempenho na validação cruzada Nested.	60
Tabela 8 – Parâmetros e espaços de busca para os modelos de aprendizado de máquina.	77
Tabela 9 – Hiperparâmetros e espaços de busca para os modelos de aprendizado de máquina.	78
Tabela 10 – Parâmetros otimizados para cada um dos modelos.	79
Tabela 11 – Hiperparâmetros otimizados para cada um dos modelos.	80

Sumário

1	INTRODUÇÃO	13
2	OBJETIVOS	15
3	EFEITOS DA COVID-19 NA REPRODUÇÃO HUMANA	16
4	ELIPSOMETRIA ESPECTROSCÓPICA DE ÂNGULO VARIÁVEL	18
4.1	Princípios de óptica	18
4.1.1	Interação entre a luz e o meio	18
4.1.2	Índice de refração complexo	18
4.1.3	Constante dielétrica complexa	19
4.1.4	Função dielétrica	19
4.1.5	Leis de reflexão e refração da luz	21
4.1.6	Polarização da luz	21
4.1.7	Interferência e tratamento para multicamadas	24
4.2	Princípios da elipsometria espectral	26
4.2.1	Medidas da elipsometria	26
4.2.2	Elementos ópticos de um elipsômetro	28
5	O PROBLEMA DAS REFLEXÕES INCOERENTES NOS MODELOS ÓPTICOS PARA ELIPSOMETRIA	30
5.1	Natureza do problema	30
5.2	Soluções discutidas na literatura	31
6	APRENDIZADO DE MÁQUINA	34
6.1	Algoritmos	34
6.2	Otimização de hiperparâmetros	35
6.3	Técnicas de tratamento de dados	36
6.4	Métricas de desempenho	36
6.5	Métodos de validação	37
6.5.1	Validação cruzada K-fold	38
6.5.2	Validação cruzada Nested	40
6.6	Interpretabilidade	40
7	ESPECTROSCOPIA DE APRENDIZADO DE MÁQUINA APLICADA AO DIAGNÓSTICO MÉDICO	42

8	METODOLOGIA	44
8.1	Coleta de amostras	44
8.2	Análise das amostras de sêmen	44
8.3	Aprendizado de Máquina aplicado aos parâmetros espermáticos . .	45
8.4	Medidas de espectroscopia	46
8.5	O problema das reflexões incoerentes	47
8.6	Aprendizado de Máquina aplicado aos dados de espectroscopia . . .	47
8.7	Análise estatística dos dados	48
9	RESULTADOS E DISCUSSÃO	49
9.1	Análise das amostras de sêmen	49
9.2	Aprendizado de Máquina aplicado aos parâmetros espermáticos . .	51
9.3	Medidas de espectroscopia	54
9.4	O problema das reflexões incoerentes	55
9.5	Aprendizado de Máquina aplicado aos dados de espectroscopia . . .	58
10	CONCLUSÕES	62
	REFERÊNCIAS	64
	APÊNDICES	76
	APÊNDICE A – TABELAS DE ESPAÇOS DE BUSCA PARA OS PARÂMETROS E HIPERPARÂMETROS	77
	APÊNDICE B – TABELAS DE PARÂMETROS E HIPERPARÂ- METROS OTIMIZADOS	79

1 Introdução

A COVID-19, doença infecciosa causada pelo vírus SARS-CoV-2, foi declarada pela Organização Mundial da Saúde (OMS) como pandemia em março de 2020. A pandemia foi um marco na nossa história recente e gerou impactos em várias esferas da sociedade como economia, relações sociais e, principalmente, na saúde global. A infecção viral pode causar diversos sintomas incluindo febre, dor de garganta, dores musculares, coriza, diarreia e problemas respiratórios. Algumas complicações também pode ser manifestadas, como perda de olfato e paladar e, em casos graves, a doença pode levar à morte. Desde o primeiro caso reportado até março de 2025, quase 780 milhões de casos e mais de 7 milhões de mortes foram registrados no mundo. Sendo que a quantidade de mortes no Brasil corresponde a cerca de 10% do número total [1, 2]. É importante ressaltar ainda que uma quantidade significativa de casos de infecção e mortes por COVID-19 está subnotificada. Além disso, o contexto socioeconômico e regional dos indivíduos infectados está relacionado à gravidade dos impactos causados pela doença [3]. Apesar de a pandemia ter sido controlada, mesmo após a aplicação de mais de 13 bilhões de doses de vacina, com 67% da população mundial tendo recebido ao menos uma dose, a ocorrência de um número significativo de casos de infecção e mortes ainda persiste [1].

Os sintomas e impactos da COVID-19 no sistema respiratório são bem conhecidos. Porém, vários estudos evidenciaram a presença e manifestação do vírus em diferentes órgãos do corpo humano. Diversas anormalidades associadas à doença já foram observadas nos sistemas renal, nervoso, hepático, tegumentar (pele e pelos), vascular, gastrointestinal, cardíaco e endócrino [4–6]. Destaca-se ainda que uma série de efeitos adversos, associados ao vírus SARS-CoV-2, foram observados no sistema reprodutor humano, incluindo alterações nos testículos, epidídimo (estrutura próxima aos testículos) e em vários parâmetros espermáticos [7–9]. Os efeitos da COVID-19 no trato reprodutor humano serão melhor discutidos no Capítulo 3. Tal tema é fundamental para a presente pesquisa, visto que aqui serão analisadas amostras de sêmen, e seus parâmetros espermáticos, de pacientes que contraíram a doença e se recuperaram.

Além da análise dos parâmetros espermáticos das amostras de sêmen, também procederemos com medidas de espectroscopia por uma técnica óptica chamada elipsometria espectroscópica. A elipsometria é uma técnica capaz de medir a espessura e propriedades ópticas de superfícies, interfaces, filmes finos, materiais nanoestruturados e também amostras biológicas. Essa abordagem tem a vantagem de ser não invasiva e não destrutiva, sendo capaz de extrair informações do material sem perturbá-lo. O conceito da elipsometria está na propriedade da luz sofrer alteração na sua polarização ao interagir com uma superfície. De forma geral, o seu funcionamento consiste na incidência de luz linearmente

polarizada em um ângulo oblíquo sobre a amostra a ser medida. Ao ser refletida, a luz se torna elipticamente polarizada, o que justifica o nome “elipsometria” [10–12]. O equipamento utilizado para fazer as medidas neste trabalho, chamado de elipsômetro, possui algumas funcionalidades específicas. Ele tem a capacidade de realizar medições em diversos ângulos e em múltiplos comprimentos de onda, que vão desde o ultravioleta (UV) até o infravermelho próximo (NIR). Dessa forma, a aplicação da técnica nos permitiu coletar uma quantidade muito grande de dados experimentais das amostras de sêmen que estamos estudando.

Neste trabalho, propomos a aplicação de inteligência artificial (IA), mais especificamente, aprendizado de máquina, para analisar e classificar os dados coletados dos parâmetros espermáticos e das medidas de elipsometria. A ideia de usar aprendizado de máquina, ou *Machine Learning* (ML), vem do fato de a técnica ser capaz de receber uma grande quantidade de dados, com potencial para detectar padrões e variações discretas nos dados. Um trabalho que os humanos podem não ter capacidade de realizar [13, 14]. Recentemente, ML tem sido aplicado de forma bem-sucedida para resolver problemas de física, química e biologia [15]. Aplicações de aprendizado de máquina relacionadas à COVID-19 também alcançaram bons resultados. Uma ampla variedade de estudos tem demonstrado a capacidade dos modelos de ML em realizar previsões [16], diagnóstico médico [17], investigação de infertilidade humana [18], bem como avaliação da qualidade e seleção de espermatozoides para otimizar procedimentos de reprodução assistida [19, 20]. Além disso, a abordagem de usar aprendizado de máquina em conjunto com medidas de espectroscopia já foi aplicada por nosso grupo. Com essa técnica, conseguimos uma boa precisão na classificação entre amostras de sêmen de camundongos saudáveis e infectados com o vírus da Zika [21, 22], assim como no diagnóstico de pacientes com COVID-19 [16, 17].

2 Objetivos

O objetivo do presente trabalho é o desenvolvimento de um método que permita a distinção de amostras de sêmen de pacientes que foram diagnosticados com COVID-19 e posteriormente se recuperaram, daqueles que nunca haviam adquirido a doença, tornando possível o diagnóstico dos efeitos da doença na saúde reprodutiva humana. Para isso, buscamos demonstrar a capacidade de modelos de aprendizado de máquina em classificar adequadamente as amostras de sêmen. A partir dessa abordagem, busca-se investigar correlações entre a infecção pelo SARS-CoV-2 e alterações associadas à fertilidade masculina, contribuindo para a compreensão dos impactos da COVID-19 na saúde reprodutiva. Ambos os métodos citados anteriormente serão demonstrados. O primeiro deles é o emprego de ML na análise dos parâmetros espermáticos. O segundo é a abordagem da espectroscopia associada ao aprendizado de máquina, que aqui chamaremos de “espectroscopia de aprendizado de máquina”, traduzido do termo usual em inglês: *machine learning spectroscopy* (MLS). Ao comprovar o potencial da aplicação dos dois procedimentos, estaremos demonstrando um método inovador, rápido e de baixo custo capaz de auxiliar na triagem e diagnóstico de problemas de infertilidade, como também poderá ajudar a entender os impactos da doença na reprodução humana.

3 Efeitos da COVID-19 na reprodução humana

Cerca de 17,5% da população mundial apresenta problemas de infertilidade. Apesar de ser um problema de grande relevância, a prevenção, o diagnóstico e o tratamento permanecem inacessíveis para uma grande parte da população [23]. Nas últimas duas décadas, muitas pesquisas têm voltado sua atenção para investigar relações entre infecções virais e problemas causados ao sistema reprodutor masculino. Uma variedade de mais de 30 diferentes vírus já foi encontrada no sêmen [24]. Porém, essas infecções costumam ser assintomáticas, o que dificulta a sua detecção. Ademais, o vírus pode permanecer no sêmen por um tempo prolongado, maior do que em outros fluidos do corpo humano, como sangue, urina e saliva. Isso pode indicar que o trato reprodutor masculino, além de ser vulnerável a infecções virais, também funciona como um reservatório, sendo um ambiente favorável para a replicação e permanência do vírus [24, 25]. Uma série de efeitos adversos está relacionada à presença de vírus no sistema reprodutor masculino. Os efeitos incluem: inflamação testicular (orquite), apoptose (morte celular programada) de células germinativas, fibrose, disrupção da espermatogênese, diminuição da síntese de testosterona, apoptose das células de Leydig e comprometimento da barreira hemato-testicular das células de Sertoli (células suporte para a produção de espermatozoides). Essas implicações podem levar a diversas consequências, como: infertilidade, infecção da linhagem germinativa, câncer nos órgãos genitais masculinos, transmissão sexual de doenças, transmissão para embriões e anormalidades congênitas [21, 24–26].

A já conhecida vulnerabilidade do sistema reprodutor masculino a infecções virais levou a uma diversidade de pesquisas sobre a presença e efeitos do vírus SARS-CoV-2 nesse contexto [27]. Em relação à detecção do vírus no trato reprodutor, os resultados não são claros. Enquanto alguns estudos não conseguiram detectar a presença do vírus em todas ou na maioria das amostras de sêmen analisadas [7, 8], existem outras pesquisas que confirmaram a presença do vírus em todas ou na maioria de suas amostras [26, 28]. É importante ressaltar que o método utilizado para identificar o vírus no material analisado possui uma grande importância nos resultados obtidos [26, 28]. Apesar dos diferentes resultados para a presença do vírus, estudos observaram o surgimento da Enzima Conversora de Angiotensina 2 (ECA2) em células do sistema reprodutor masculino, sendo esse um receptor presente na superfície celular ao qual o SARS-CoV-2 pode se ligar para entrar nas células. Isso sugere, portanto, que existe um caminho para que o vírus consiga infectar as células e causar possíveis danos [8, 29]. A respeito dos efeitos do vírus, há um consenso de que, independentemente de sua presença ou não no sistema genital, diversos efeitos adversos

podem ocorrer e já foram observados [7]. Os impactos da COVID-19 incluem: necrose e apoptose das células espermatógenicas, lesões nas células de Sertoli, danos aos túbulos seminíferos, redução e inibição das células de Leydig, inflamação linfocitária, orquite viral, contagem reduzida de espermatozoides, aumento da concentração de leucócitos e citocinas inflamatórias no sêmen, fibrose testicular, alterações vasculares e perda da barreira das células de Sertoli. Essa diversidade de efeitos negativos devidos à infecção pelo vírus SARS-CoV-2 tem potencial para levar a consequências graves aos órgãos genitais masculinos, com o risco de infertilidade. [7, 8, 26, 28].

Em um estudo que analisou vários parâmetros espermáticos de amostras de sêmen de pacientes, antes e depois da infecção por SARS-CoV-2, concluiu-se que houve alteração de alguns parâmetros após a doença. Uma redução significativa foi observada na concentração média de espermatozoides, na contagem total de espermatozoides e na porcentagem média de amostras com morfologia normal após a infecção pelo vírus [9]. Uma outra pesquisa procedeu com a análise dos parâmetros espermáticos, comparando um grupo de pacientes que estavam há três meses recuperados da COVID-19 com grupos controle de pacientes saudáveis. Nesse caso, não foram observadas variações significativas entre os grupos para concentração e contagem total de espermatozoides, motilidade progressiva e porcentagem de amostras com morfologia anormal [30]. Porém, apesar da recuperação dos parâmetros espermáticos observada a partir do trabalho de Paoli et al. (2022) [30], impactos negativos e taxas mais altas de aborto espontâneo, associados à infecção por SARS-CoV-2, foram relatados em procedimentos de reprodução assistida entre pacientes recuperados por até seis meses [31]. Dessa forma, o intuito deste trabalho é investigar a relação dos efeitos, diretos ou indiretos, da infecção pelo vírus da COVID-19 com possíveis problemas causados à reprodução humana. Para isso, utilizaremos novas metodologias, que incluem a análise dos parâmetros espermáticos e medidas de espectroscopia de amostras de sêmen através de aprendizado de máquina.

4 Elipsometria Espectroscópica de Ângulo Variável

Como abordado no Capítulo 1, elipsometria espectroscópica é uma técnica de medições ópticas que permite extrair informações sobre propriedades ópticas de uma amostra sem danificá-la. O elipsômetro usado aqui é capaz de fazer medidas em diversos comprimentos de onda e também em diferentes ângulos de incidência da luz. Com o objetivo de esclarecer os princípios físicos envolvidos nessa técnica e como eles são aplicados, este capítulo foi dividido entre: princípios de óptica e princípios da elipsometria espectroscópica.

4.1 Princípios de óptica

4.1.1 Interação entre a luz e o meio

A propagação da luz, uma onda eletromagnética, pode ser descrita de uma forma simples por ondas senoidais. Porém, ao tratar o caso para quando a luz entra em um determinado meio, o seu comportamento se torna mais complexo para interpretar. Quando lidamos com essa situação, a luz pode ser refletida ou transmitida pelo meio, e alguns fenômenos podem ocorrer. A luz pode mudar de direção, ter sua velocidade alterada e apresentar variação em sua amplitude. Cada um desses fenômenos pode ser descrito ao considerarmos as constantes ópticas do material, como índice de refração complexo ou constante dielétrica. Além disso, a polarização da luz também pode sofrer alterações [11,32].

4.1.2 Índice de refração complexo

Ao ser transmitida por um meio opticamente diferente, a luz sofre uma refração, alterando sua velocidade e sua direção de propagação. O índice de refração (n) é uma propriedade óptica do material e define qual seria a velocidade da luz que atravessa por ele, assim como é usado para calcular a nova direção de propagação. Essa propriedade é normalmente definida como:

$$n \equiv \frac{c}{v}, \quad (4.1)$$

sendo c a velocidade da luz no vácuo e v a velocidade da luz no meio. Porém, para uma descrição mais completa, precisamos levar em conta que o meio pode absorver parte da energia da luz durante sua transmissão. Para isso, introduzimos o índice de refração complexo (N), composto por uma parte real e outra parte complexa:

$$N \equiv n - ik, \quad (4.2)$$

em que i é o número complexo e k é o coeficiente de extinção, sendo ele o responsável por determinar o comportamento de absorção do material [32].

4.1.3 Constante dielétrica complexa

Apesar do índice de refração descrever a propagação da luz em um meio, o que determina essa propriedade é a polarização dielétrica gerada nesse meio. Quando uma radiação eletromagnética incide sobre um material dielétrico, a interação com o campo elétrico faz surgir forças opostas entre cargas elétricas positivas e negativas, o que promove uma separação espacial entre elas. Dessa forma, o resultado é uma polarização dielétrica. A magnitude da polarização em um material dielétrico é representada pela constante dielétrica (ε), também chamada de permissividade. Portanto, essa propriedade descreve como um material é afetado pela incidência de luz. A constante dielétrica pode ser escrita como:

$$\varepsilon = 1 + \frac{P}{\varepsilon_0 E} = 1 + \chi, \quad (4.3)$$

onde P é a polarização dielétrica sofrida pelo meio com a interação com o campo elétrico E e ε_0 é a permissividade no vácuo. A fração na equação é equivalente à susceptibilidade dielétrica χ . Existe uma relação direta entre a constante dielétrica e o índice de refração complexo, que pode ser definida por:

$$N^2 \equiv \varepsilon, \quad (4.4)$$

nesse caso, ε é um número complexo e também é definido por uma parte real e uma imaginária:

$$\varepsilon \equiv \varepsilon_1 - i\varepsilon_2, \quad (4.5)$$

em que, a partir das Eqs. (4.2), (4.4) e (4.5), obtemos $\varepsilon_1 = n^2 - k^2$ e $\varepsilon_2 = 2nk$. Dessa forma, ε_1 e ε_2 também são grandezas que podem ser usadas para descrever propriedades de um meio e estão relacionadas com o índice de refração e coeficiente de extinção [11, 32, 33].

4.1.4 Função dielétrica

Uma outra propriedade importante da constante dielétrica complexa ε é a sua dependência com a frequência angular ω da luz incidente, uma relação dada pela função

dielétrica. A Figura 1 mostra como é o comportamento das componentes real ε_1 e imaginária ε_2 dessa grandeza em função do $\log \omega$ [32].

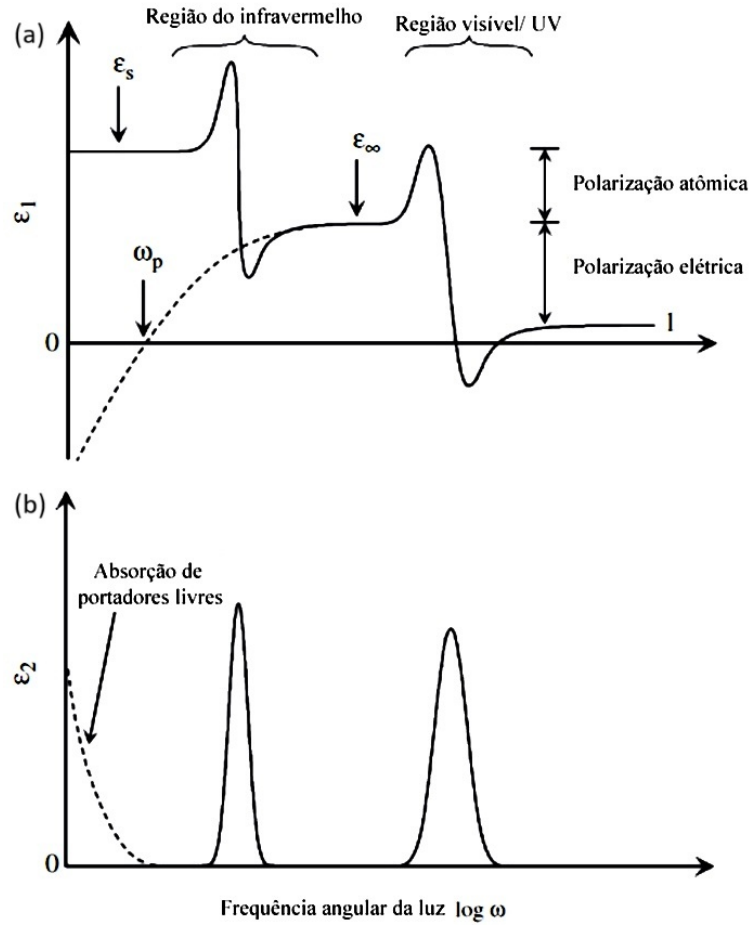


Figura 1 – Gráfico da parte real e imaginária da constante dielétrica complexa em função do logaritmo da frequência angular da luz $\log(\omega)$. Adaptado de [32].

Podemos observar no gráfico (a) da Figura 1 que existem duas regiões em que ocorrem picos do valor da parte imaginária ε_2 . Nesses pontos há uma coincidência entre a frequência angular da luz incidente e a frequência de oscilação atômica, ocorrendo uma ressonância. Pela relação apresentada anteriormente: $\varepsilon_2 = 2nk$, notamos que um aumento no valor da ε_2 leva a um incremento do valor do coeficiente de extinção k . Dessa forma, a luz é absorvida mais intensamente pelo meio nesses pontos de ressonância. Focando agora no gráfico (b) da Figura 1, notamos que nos pontos de ressonância também ocorrem grandes variações no valor da parte real ε_1 . Esses pontos representam, no primeiro caso, a perda da contribuição da polarização atômica, e no segundo caso, da polarização elétrica. Percebemos também que existem três importantes regiões onde ε_1 se mantém constante. Para valores de frequência angular baixa, o valor de ε_1 é representado pela constante dielétrica estática ε_s . Quando a frequência aumenta, ε_1 decresce e é definida como a constante para altas frequências ε_∞ . Para valores de ω ainda maiores, a polarização do meio não interfere mais na passagem da luz e, portanto, o valor da constante fica

$\varepsilon_1 = 1$, assim como seria no vácuo. Ademais, as linhas pontilhadas nos gráficos (a) e (b) representam o comportamento da constante dielétrica para materiais que possuem elétrons livres e buracos, sendo que o ponto ω_p define frequência de plasma, na qual ocorre a frequência natural de oscilação dos elétrons livres [32].

Dessa forma, a função dielétrica define o comportamento da constante dielétrica complexa em função da frequência angular da luz incidente, que é o mesmo que dizer que ε depende do comprimento de onda da luz, ou da sua energia. Isso justifica a importância de realizarmos as medidas de elipsometria em nossas amostras para diferentes comprimentos de onda. Com isso, estamos avaliando opticamente as amostras levando em consideração as variações de comportamento da sua constante dielétrica complexa em função do comprimento de onda da luz.

4.1.5 Leis de reflexão e refração da luz

Quando a luz incide sobre uma superfície, parte dela é transmitida (refratada) e outra parte é refletida. A Figura 2 demonstra ambas as possibilidades, sendo que θ_i é o ângulo de incidência da luz (seta amarela), θ_r é o ângulo de reflexão, θ_t é o ângulo de transmissão; N_i e N_t são os índices de refração complexos dos meios por onde a luz foi incidida e transmitida, respectivamente. Para esse caso, sabemos que a lei da reflexão diz que o ângulo de reflexão é igual ao de incidência, dessa forma:

$$\theta_i = \theta_r, \quad (4.6)$$

assim como, pela Lei de Snell, temos a seguinte relação entre a luz incidida e refratada:

$$N_i \sin\theta_i = N_t \sin\theta_t. \quad (4.7)$$

Com essas duas leis, podemos descrever as trajetórias da luz refletida e transmitida após interagir com uma superfície [11, 32].

4.1.6 Polarização da luz

A polarização da luz está relacionada ao comportamento do campo elétrico, tomando como referência a direção de propagação da luz. Existem quatro possibilidades de polarização da luz. Uma luz não polarizada significa que a direção do campo elétrico de cada fóton difere dos demais, não tendo uma polarização definida. Quando o campo elétrico de todos os fótons oscila apenas em uma única direção, a luz é linearmente polarizada. A luz circularmente polarizada é resultante de duas ondas com polarização linear perpendiculares entre si e com uma diferença de fase de $\pi/2$. Dessa forma, o vetor do campo elétrico gira com magnitude constante no plano perpendicular à direção de propagação, formando

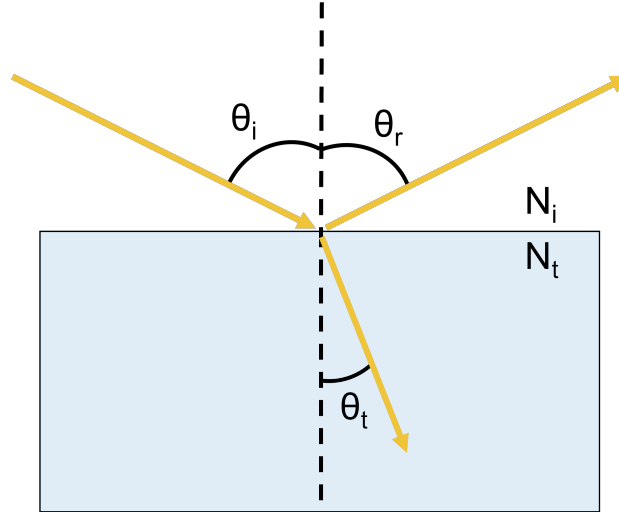


Figura 2 – Esquema demonstrando a luz incidente, refletida e transmitida ao incidir em um meio.

um círculo. Quando duas ondas possuem diferença de fase distinta de $\pi/2$ e direções de polarização perpendiculares, a luz resultante é elipticamente polarizada [11]. A Figura 3 mostra a representação dos possíveis tipos de polarização da luz.

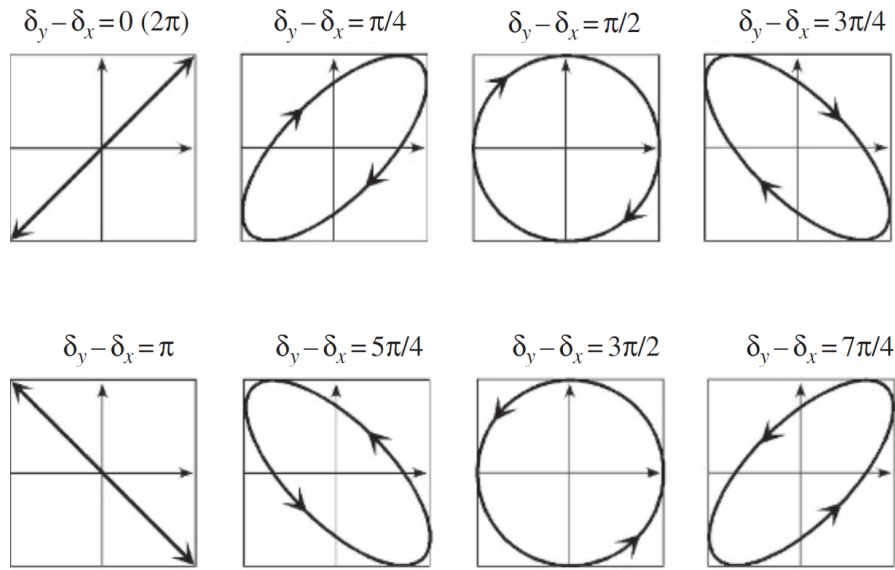


Figura 3 – Diferentes possibilidades para a polarização da luz levando em conta as diferenças de fase ($\delta_y - \delta_x$). Nessa figura as amplitudes das ondas em x e y são iguais ($E_{y0} = E_{x0}$). Adaptado de [32].

A luz elipticamente polarizada é importante aqui, sendo inclusive a razão pela qual a técnica que usamos é chamada de elipsometria. Quando uma luz linearmente polarizada é refletida por uma superfície, ela pode se tornar elipticamente polarizada. A elipsometria extrai propriedades das amostras medindo diferenças nessa polarização da luz, que por sua vez depende das propriedades do material ao qual ela interage [11].

Uma outra classificação usada na abordagem da polarização da luz é das ondas

com polarização p e s. A luz incidente, refletida ou transmitida pelas amostras pode ser classificada com polarização p quando a direção da oscilação do campo elétrico está na mesma direção do plano de incidência. Já a luz polarizada s tem seu campo elétrico oscilando no plano perpendicular ao plano de incidência, como mostrado na Figura 4. Cada um desses tipos de polarização da luz apresenta um comportamento específico [32].

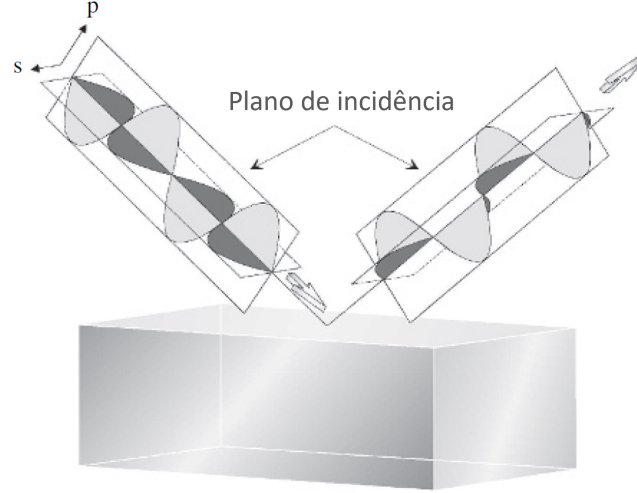


Figura 4 – Reflexão de ondas de luz polarizadas p e s. Adaptado de [32].

As polarizações da luz são representadas pelos vetores do campo elétrico E e da indução magnética B . Os índices i , r e t indicam, respectivamente, incidência, reflexão e transmissão na Figura 5. A partir desses dados, é possível chegar às amplitudes do coeficiente de reflexão (r) e do coeficiente de transmissão (t) tanto para a luz polarizada p:

$$r_p \equiv \frac{E_{rp}}{E_{ip}} = \frac{n_t \cos \theta_i - n_i \cos \theta_t}{n_t \cos \theta_i + n_i \cos \theta_t}, \quad (4.8)$$

$$t_p \equiv \frac{E_{tp}}{E_{ip}} = \frac{2n_i \cos \theta_i}{n_t \cos \theta_i + n_i \cos \theta_t}, \quad (4.9)$$

como também para a luz polarizada s [32]:

$$r_s \equiv \frac{E_{rs}}{E_{is}} = \frac{n_i \cos \theta_i - n_t \cos \theta_t}{n_i \cos \theta_i + n_t \cos \theta_t}, \quad (4.10)$$

$$t_s \equiv \frac{E_{ts}}{E_{is}} = \frac{2n_i \cos \theta_i}{n_i \cos \theta_i + n_t \cos \theta_t}. \quad (4.11)$$

A diferença entre as Eqs. (4.8) e (4.10) e também das Eqs. (4.9) e (4.11) demonstra como o tratamento de cada tipo de polarização deve ser diferente, visto que o comportamento de cada uma delas não é o mesmo. Importante destacar também que há uma dependência de todos os coeficientes em relação ao índice de refração do meio por onde a

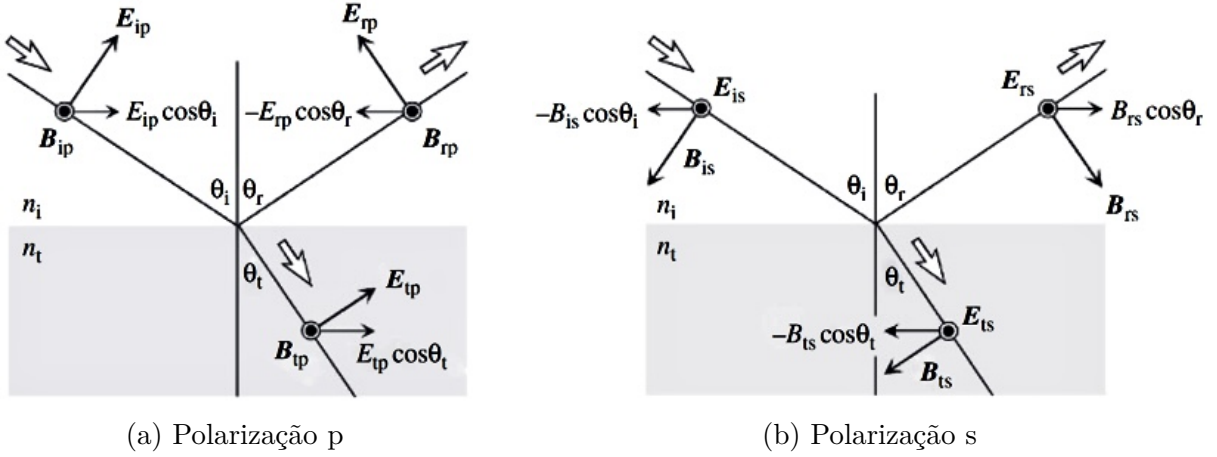


Figura 5 – Representação da incidência, reflexão e transmissão da luz (a) p-polarizada e (b) s-polarizada [32].

luz é transmitida (n_t). O que indica que as propriedades da luz refletida e transmitida, em qualquer polarização, são influenciadas pelas propriedades ópticas do meio.

4.1.7 Interferência e tratamento para multicamadas

A superposição de ondas de luz geradas em posições diferentes é chamada de interferência. Tal fenômeno pode ser observado ao incidir luz sobre um filme fino colocado sobre um substrato, como mostrado na Figura 6. Nessa figura, temos um modelo óptico para um sistema com estrutura ambiente/filme/substrato. Sendo que N_0 , N_1 e N_2 são os índices de refração complexos do ar, filme fino e substrato, respectivamente, e d é a espessura do filme fino. Nesse caso, vemos que a luz incidente é refletida tanto na superfície ar/filme como também na segunda interface filme/substrato. Sendo assim, as duas ondas de luz resultantes se sobrepõem e ocorre uma interferência entre elas. Porém, isso não acontece apenas uma vez. Quando lidamos com um filme fino com baixo coeficiente de absorção, várias reflexões internas ocorrem, o que é mostrado de forma mais detalhada na Figura 7 [32].

Usando as Eqs. (4.8) a (4.11) e aplicando a notação da Figura 7, obtemos as equações para o cálculo de cada uma das componentes dos coeficientes de reflexão e coeficientes de transmissão:

$$r_{jk,p} = \frac{N_k \cos \theta_j - N_j \cos \theta_k}{N_k \cos \theta_j + N_j \cos \theta_k}, \quad (4.12)$$

$$t_{jk,p} = \frac{2N_j \cos \theta_j}{N_k \cos \theta_j + N_j \cos \theta_k}, \quad (4.13)$$

$$r_{jk,s} = \frac{N_j \cos \theta_j - N_k \cos \theta_k}{N_j \cos \theta_j + N_k \cos \theta_k}, \quad (4.14)$$

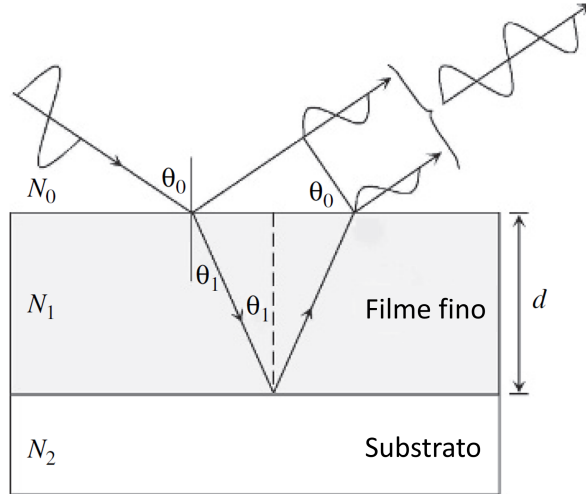


Figura 6 – Interferência óptica em um filme fino sobre um substrato. Adaptado de [32].

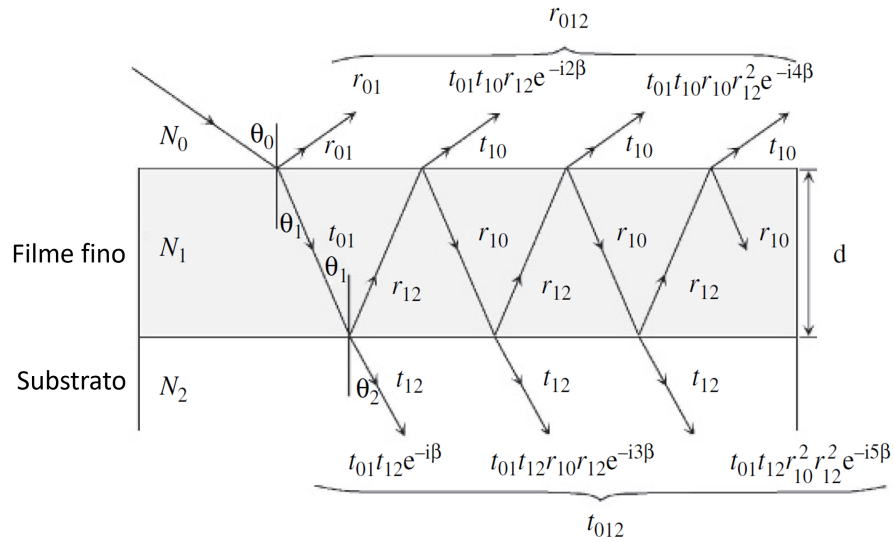


Figura 7 – Modelo óptico para um sistema com estrutura: ambiente/filme fino/substrato. Adaptado de [32].

$$t_{jk,s} = \frac{2N_j \cos \theta_j}{N_j \cos \theta_j + N_k \cos \theta_k}. \quad (4.15)$$

Aqui, usamos j e k como índices que podem ser 0, 1 ou 2, de acordo com qual componente se deseja obter [32].

Agora consideramos um sistema de multicamadas, sendo ar/1ª camada/2ª camada/substrato, ilustrado na Figura 8. Este sistema é o mais semelhante possível ao que usamos para as medições de elipsometria. Nas nossas medidas, a 1ª camada é uma lamínula de vidro transparente, a 2ª é a amostra que desejamos estudar e a 3ª é outra lamínula de vidro (substrato).

A partir das Eqs. (4.12) a (4.15) podemos obter a amplitude do coeficiente de reflexão total (r_{0123}) e também do coeficiente de transmissão total (t_{0123}). Ou seja, teremos

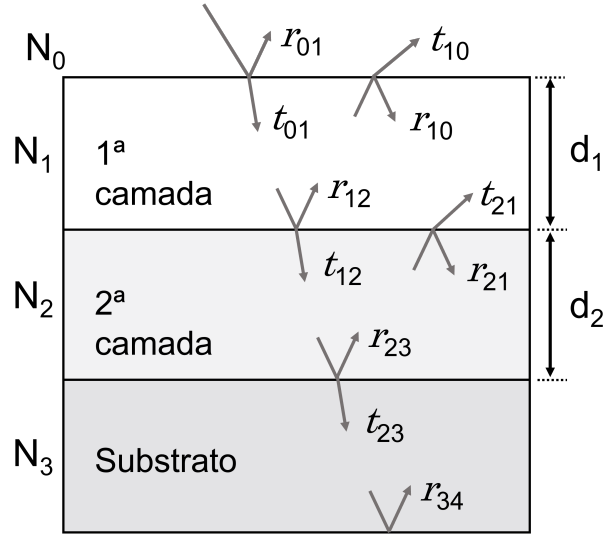


Figura 8 – Modelo óptico para um sistema de multicamadas. Adaptado de [32].

um coeficiente com informação para a inferência de todas as ondas de luz que se sobrepõem no modelo. Os coeficientes são dados por:

$$r_{0123} = \frac{r_{01} + r_{12}e^{-i2\beta_1} + [r_{01}r_{12} + e^{-i2\beta_1}]r_{23}e^{-i2\beta_2}}{1 + r_{01}r_{12}e^{-i2\beta_1} + [r_{12} + r_{01}e^{-i2\beta_1}]r_{23}e^{-i2\beta_2}}, \quad (4.16)$$

$$t_{0123} = \frac{t_{01}t_{12}t_{23}e^{-i(\beta_1+\beta_2)}}{1 + r_{01}r_{12}e^{-i2\beta_1} + [r_{12} + r_{01}e^{-i2\beta_1}]r_{23}e^{-i2\beta_2}}, \quad (4.17)$$

em que aqui o β_1 e β_2 podem ser calculados pela equação:

$$\beta_j = \frac{2\pi d_j}{\lambda} N_j \cos \theta_j, \quad (4.18)$$

sendo λ o comprimento de onda da luz [32].

É importante destacar que aqui estamos considerando um sistema ideal. Quando lidamos com essa questão na prática, as reflexões que ocorrem no interior não são coerentes como foram consideradas. Nesse caso, surge o problema das reflexões incoerentes que buscamos resolver neste trabalho, descrito no Capítulo 5.

4.2 Princípios da elipsometria espectroscópica

4.2.1 Medidas da elipsometria

Na elipsometria, luz é incidida sobre uma superfície e as medidas são extraídas a partir das mudanças na polarização da luz após ser refletida ou transmitida pela amostra. A Figura 9 demonstra em detalhe as ondas com polarizações p e s incidindo e refletindo em uma amostra. É possível observar que a polarização resultante é elíptica, com duas

propriedades que a define: Ψ e Δ . Ψ é definido pela razão entre as amplitudes das ondas de polarização p e s. Já Δ é a diferença de fase entre as duas ondas. Como vimos antes, a luz de polarização p e a de polarização s se comportam de formas diferentes. Isso justifica a alteração da polarização após a reflexão [32].

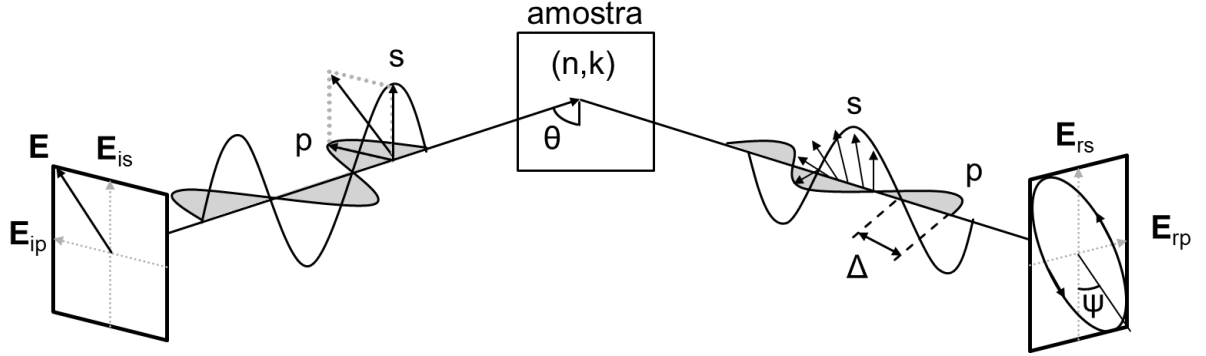


Figura 9 – Esquema das medidas de elipsometria. Adaptado de [32].

Dessa forma, as grandezas Ψ e Δ são as duas principais medidas dessa técnica, sendo consideradas parâmetros da elipsometria. A relação entre elas pode ser escrita a partir da razão entre os coeficientes de reflexão para p e s

$$\rho \equiv \tan(\Psi) e^{i\Delta} \equiv \frac{r_p}{r_s}, \quad (4.19)$$

e também pelos coeficientes de transmissão

$$\rho \equiv \tan(\Psi) e^{i\Delta} \equiv \frac{t_p}{t_s}. \quad (4.20)$$

Considerando a parte da reflexão, Ψ e Δ podem ser representados de uma forma mais direta a partir de:

$$\tan(\Psi) = \left| \frac{r_p}{r_s} \right|, \quad (4.21)$$

$$\Delta = \delta_p - \delta_s. \quad (4.22)$$

Uma outra medida da elipsometria é a despolarização (dPol). Quando consideramos uma situação ideal, tanto a luz incidente como também a refletida/transmitida são totalmente polarizadas. Porém, em medidas reais, pode ocorrer que uma parte da luz se torne não polarizada após interagir com a amostra. Nesse sentido, a medida dPol descreve a porcentagem da luz que ficou sem polarização. Dessa forma, ela é uma propriedade importante a ser considerada quando lidamos com amostras não ideais. Um dos motivos que pode levar a uma despolarização da luz é a presença de reflexões das camadas

internas do sistema medido, que produzem luz incoerente em relação àquela refletida pela superfície [34].

A última medida da elipsometria que vamos considerar é simplesmente a intensidade da luz refletida, sem considerar a polarização da luz. Aqui chamaremos essa grandeza de uR (*unpolarized reflection*). Para obtermos o valor de uR , temos que

$$uR = R \cdot I_i, \quad (4.23)$$

em que I_i é a intensidade da luz incidente e R é a reflectância da luz não polarizada, podendo ser escrita como

$$R = \frac{R_p + R_s}{2}, \quad (4.24)$$

na qual

$$R_p = |r_p|^2 \quad e \quad R_s = |r_s|^2. \quad (4.25)$$

Assim, conseguimos obter o valor de uR a partir dos coeficientes de reflexão para a luz com polarização p e s [32].

4.2.2 Elementos ópticos de um elipsômetro

O modelo de elipsômetro utilizado neste trabalho, da empresa J. A. Woollam Co., além de ser de ângulo variável, também possui elementos ópticos específicos. A classificação para o instrumento que utilizamos é de elipsometria com compensador rotativo, mais conhecido pelo termo em inglês *rotating-compensator ellipsometry* (RCE). Uma nomenclatura secundária, mais específica, que ele recebe, é expressa através dos símbolos dos elementos ópticos e informa a posição de cada um deles, incluindo a amostra. Nesse caso, lidamos com um $PSC_R A$, o que significa que seus elementos são, nessa ordem, polarizador (P), amostra (S), compensador rotativo (C_R) e analisador (A) [32]. A Figura 10 demonstra um esquema para o sistema de medidas utilizado neste trabalho.

Um polarizador e um analisador, na realidade, são o mesmo elemento. O que muda entre um e outro é a sua função, ou posição, no sistema de medidas. O polarizador, colocado logo após a fonte de luz, serve para extrair luz linearmente polarizada da luz não polarizada proveniente da fonte. Já o analisador, posicionado antes do detector, tem a função de determinar qual é o estado da polarização da luz que passa por ele [32].

Um compensador, que no esquema da Figura 10 é mostrado antes do analisador, pode ser usado para converter luz linearmente polarizada em luz elipticamente polarizada, ou o contrário. Dessa forma, ele permite que seja possível controlar a diferença de fase

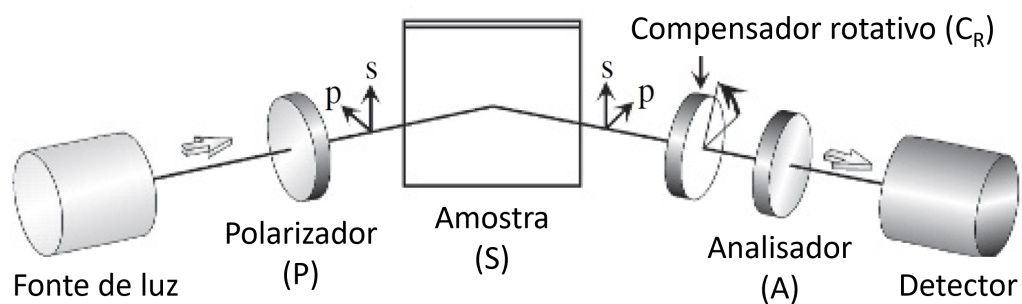


Figura 10 – Esquema do sistema RCE - $PSC_R A$. Adaptado de [32].

entre as ondas de luz com polarizações p e s. O fato de ele ser rotativo permite que a variação da diferença de fase seja dinâmica, o que torna as medições mais completas [32].

5 O problema das reflexões incoerentes nos modelos ópticos para elipsometria

O valor do índice de refração de um material está diretamente relacionado com as propriedades físico-químicas desse material, inclusive para um material biológico. Sabemos também que uma infecção bacteriana ou viral pode causar alterações nas propriedades das células biológicas afetadas. Dessa forma, a aplicação de métodos de medida do índice de refração de amostras biológicas tem ganhado importância, especialmente para o estudo da citobiologia e na realização de diagnóstico médico [22, 35–37]. Nesse sentido, conhecer o comportamento do índice de refração do material que analisamos é importante para o diagnóstico que desejamos fazer neste trabalho. Esta área de estudo ainda é recente e não são muitas as pesquisas que investigaram as propriedades ópticas dos tecidos biológicos. Um fator que dificulta bastante o desenvolvimento dessas investigações é a natureza turva das amostras biológicas. Porém, algumas metodologias já foram aplicadas e apresentaram bons resultados. Entre elas, a única capaz de obter, simultaneamente, as partes real e imaginária do índice de refração é a elipsometria [38]. Portanto, a ideia da presente pesquisa poderia incluir a aplicação da elipsometria para obter as constantes ópticas das amostras analisadas. Porém, enfrentamos o problema das reflexões incoerentes, que não permitiu modelar opticamente as amostras para obter suas propriedades físicas, logo, não foi possível analisar quantitativamente o índice de refração ou a constante dielétrica de forma direta. A solução, então, foi utilizar os dados brutos das medidas de elipsometria associado ao emprego de aprendizado de máquina na análise dos dados. Vale lembrar que as medidas da elipsometria são influenciadas pelas constantes ópticas das amostras medidas, como discutido no Capítulo 4. Desse modo, o que fazemos é uma análise indireta das propriedades ópticas das amostras de sêmen. A seguir será descrito qual foi o principal problema que enfrentamos para fazer a modelagem óptica do material estudado, bem como as soluções discutidas na literatura que foram consideradas na tentativa de solucionar o problema no nosso caso.

5.1 Natureza do problema

Quando um substrato espesso e transparente é utilizado em uma medida de espectroscopia, como a elipsometria, ocorre o problema da superposição de reflexões incoerentes. O problema acontece porque a luz refletida na superfície traseira do substrato possui uma diferença de fase em relação àquela refletida na superfície frontal, devido à diferença do caminho óptico percorrido em cada camada. Isso resulta em uma superposição

parcialmente incoerente da luz que chega ao detector do instrumento. Nesse sentido, não seria possível determinar corretamente a fase da luz refletida total, além de provocar uma depolarização da luz. Esse problema dificulta a modelagem óptica e a análise dos dados, prejudicando a precisão e eficiência das análises [39–41].

Porém, aqui ainda estamos tratando do problema decorrente de medidas em um único substrato. No nosso caso, estamos trabalhando com um sistema de multicamadas, sendo que a amostra biológica fica entre dois substratos transparentes. O sistema é mostrado na Figura 11. Portanto, estamos lidando com reflexões incoerentes provenientes não só da primeira camada, como também da amostra e do substrato inferior. Além disso, ainda ocorrem múltiplas reflexões internas em cada uma das camadas. O resultado é que a luz recebida pelo detector será uma mistura dessas várias reflexões, sendo cada uma incoerente em relação às outras [32, 39–41]. Nesse caso, estamos lidando com uma situação ainda mais difícil de solucionar em relação ao caso de apenas um substrato.

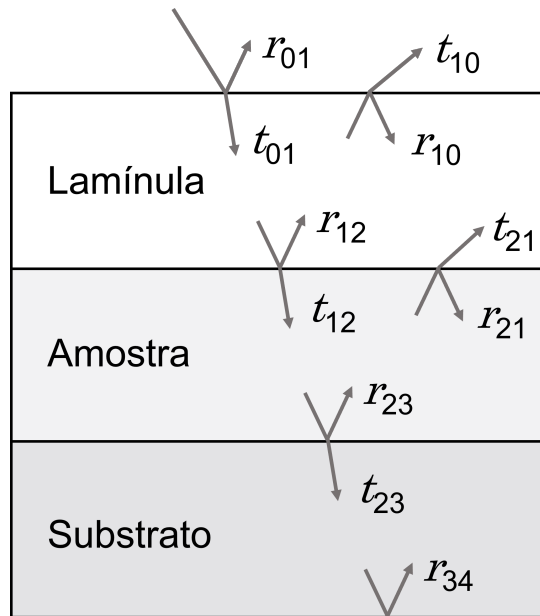


Figura 11 – Esquema do sistema de multicamadas utilizado nas medidas de elipsometria. Adaptado de [32].

5.2 Soluções discutidas na literatura

Existem diversas possibilidades de solução do problema das reflexões incoerentes que são discutidas e demonstradas de forma eficaz na literatura [40]. Todos eles estão focados em resolver o problema considerando apenas um único substrato, sendo então ideias que abrangem parcialmente nosso problema. Dessa forma, procuramos pensar e adaptar cada solução ao nosso caso. Entretanto, a maior parte delas não foi possível aplicar neste trabalho.

A primeira solução que apresentamos aqui teve bons resultados para eliminar as reflexões da superfície superior do substrato. O método consistiu em conectar um prisma com 60° de inclinação na superfície do substrato de vidro. Para evitar reflexões na interface do prisma com o vidro, foi aplicado um fluido entre eles com o mesmo índice de refração de ambos [42]. Para o nosso caso, esse método é de difícil aplicação porque fazemos medidas em diversos ângulos, bem como em diferentes posições. Realizar o alinhamento adequado do feixe de luz com o prisma para cada medida seria uma tarefa bastante custosa. Portanto, não consideramos tal aplicação.

Uma outra solução que poderia ser promissora é a aplicação de um diafragma que filtraria a luz refletida, deixando passar para o detector somente as componentes que realmente interessam para a análise. Dessa forma, as reflexões incoerentes que ocorrem nas diferentes superfícies do sistema poderiam ser eliminadas das medidas [41, 43, 44]. Porém, estamos lidando com um feixe de luz muito estreito, assim como o substrato utilizado é muito fino. Nesse caso, a separação entre os raios refletidos é muito pequena [41]. Portanto, fazer o bloqueio de componentes não desejáveis da luz refletida necessitaria de uma instrumentação extremamente precisa e, mais uma vez, teríamos o desafio do alinhamento em cada medida.

Uma outra forma de contornar esse obstáculo é fazer medidas de transmissão, em vez de reflexão. O que reduziria bastante o número de reflexões incoerentes que chegam ao detector [45]. Contudo, como mostrado na Figura 11, nossas amostras são montadas horizontalmente. Não existe possibilidade de fazer medidas de transmissão dessa forma. Para isso, seria necessário colocar a amostra na vertical, o que causaria o extravasamento do líquido. Somado a isso, a configuração do elipsômetro que utilizamos não permite medidas em transmissão. No entanto, este método merece atenção, visto que, em uma pesquisa futura, pode ser realizada uma adaptação na montagem das amostras para possibilitar as medidas por transmissão.

Um método bastante promissor aplicado para eliminar as reflexões incoerentes envolve tornar a superfície traseira do substrato rugosa. Desse modo, a luz seria refletida ou espalhada de forma não especular. Portanto, uma grande parte dela não chegaria ao detector. A Figura 12 ilustra como isso acontece. Entretanto, essa é uma técnica destrutiva e pode não ser adequada para materiais muito finos ou frágeis [41, 46–48]. Ademais, essa técnica resolveria o problema apenas da superfície traseira do substrato inferior para nossas medições.

Por último, apresentamos o método que melhor funcionou no nosso caso, mas ainda assim apresenta o mesmo problema de resolver apenas a reflexão da superfície traseira do substrato inferior. A técnica é bastante simples e consiste na aplicação de uma fita adesiva que possua o mesmo índice de refração do substrato. Quando a luz chega à fita que está conectada ao substrato, a reflexão ocorre de forma difusa, e boa parte dela não é recebida

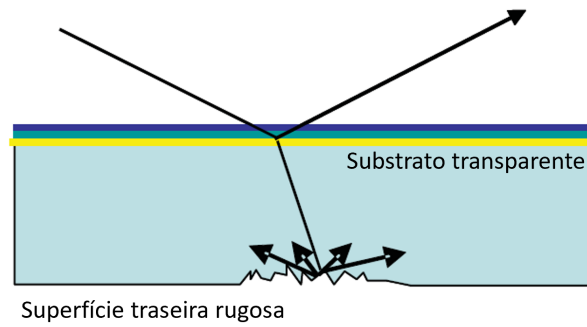


Figura 12 – Esquema da reflexão difusa na superfície rugosa. Adaptado de [48].

pelo detector [41, 47, 48]. A Figura 13 mostra um esquema desse processo. Este método foi aplicado na presente pesquisa com o objetivo de eliminar as reflexões incoerentes da superfície traseira do substrato inferior. Isso não resolveria totalmente nosso problema, mas poderia melhorar nossos resultados, permitindo uma análise mais acurada das medidas. Os materiais e métodos aplicados nessa parte do trabalho estão apresentados no Capítulo 8 e os resultados encontrados são discutidos no Capítulo 9.

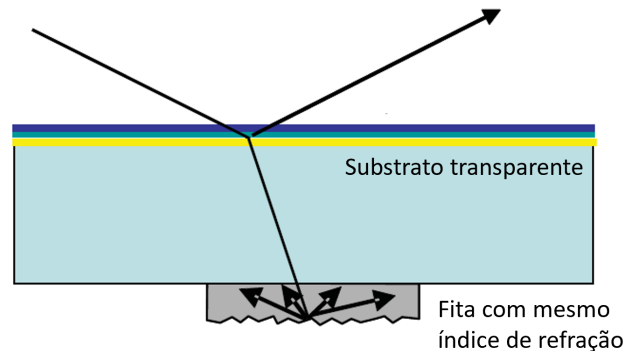


Figura 13 – Esquema da reflexão difusa na superfície da fita de mesmo índice de refração do substrato. Adaptado de [48].

Uma outra possibilidade que também foi adotada neste estudo não envolve utilizar nenhuma solução para evitar as reflexões incoerentes. Os ajustes do modelo óptico de uma amostra são feitos levando em conta que todas as reflexões existem. Nesse caso, é utilizada uma ferramenta de software que procede com correções matemáticas para atenuar a influência das reflexões incoerentes nos dados.

6 Aprendizado de Máquina

Aprendizado de máquina, ou *Machine Learning* (ML), é um amplo e muito importante ramo da inteligência artificial que possui diversas utilidades e está presente no cotidiano de muitas pessoas. Aprendizado de máquina tem sido usado em redes sociais, serviços de streaming, tradutores, corretores de texto, reconhecimento facial, equipamentos de diagnóstico médico, ferramentas de pesquisa, carros autônomos, entre muitas outras aplicações [13]. O uso de ML permite que computadores sejam capazes de aprender a partir de dados, sem necessidade de tudo ser explicitamente programado. Dessa forma, ele aprende a analisar os dados e buscar padrões para prever ou classificar novas informações ainda não conhecidas, de forma semelhante a como os humanos aprendem [13]. Usando uma notação matemática, de forma geral, podemos resumir o funcionamento do ML em um algoritmo que busca uma função alvo f que melhor descreve os dados de entrada X , resultando em uma saída Y , dessa forma, $Y = f(X)$ [49]. Existem diversos tipos de aprendizado usados em *machine learning*: supervisionado, não supervisionado, semi-supervisionado e por reforço. Aqui neste trabalho focaremos apenas no aprendizado supervisionado. O que, nesse caso, significa que os algoritmos serão treinados a partir de dados rotulados. Ou seja, a natureza dos dados será conhecida para os modelos de ML durante a sua etapa de treinamento. Após treinado, o modelo deve ser capaz de prever ou classificar um dado não rotulado. Essa abordagem permite que o aprendizado ocorra de forma progressiva e evolua sua eficiência com uma maior quantidade de dados [13, 50].

6.1 Algoritmos

Além de ter diferentes tipos de aprendizado, ML apresenta uma variedade de algoritmos, ou modelos, disponíveis para serem usados. Um algoritmo representa um conjunto de regras ou processos que são realizados com o objetivo de cumprir a tarefa de observar padrões nos dados e aprender a classificá-los. Cada modelo possui funções, cálculos, metodologias e finalidades específicas. A natureza do problema ou o objetivo para o qual ele será empregado é importante para a escolha de qual modelo pode ser usado em cada caso [51]. Neste projeto, foram utilizados seis algoritmos de classificação: *Decision Tree* (DT) [52], *eXtreme Gradient Boosting* (XGBoost) [53], *K-Nearest Neighbors* (KNN) [54], *Logistic Regression* (LR) [55], *Multilayer Perceptron* (MLP) [56] e *Support Vector Machine* (SVM) [57] (Tabela 1).

Tabela 1 – Descrição e vantagens de cada um dos modelos de aprendizado de máquina. Informações retiradas de [58].

Modelo	Descrição	Vantagem
KNN (<i>K-Nearest Neighbors</i>)	Cria um espaço n-dimensional com todas as instâncias dos dados de treino e define cada classe a partir da distância entre os pontos. A classe majoritária dos vizinhos mais próximos a um ponto definirá a classe do mesmo.	Simples e robusto a dados ruidosos.
LR (<i>Logistic Regression</i>)	Modelo estatístico que usa uma função logística (sigmoide) para mapear os dados e definir probabilidades de 0 a 1.	Funciona bem para bases de dados de alta dimensionalidade, especialmente para dados linearmente separáveis.
SVM (<i>Support Vector Machine</i>)	Cria um hiperplano ou um conjunto de hiperplanos em um espaço multidimensional que separa os pontos dos dados em diferentes classes.	Efetivo em espaços de alta dimensionalidade e pode ser ajustado para diferentes relações entre dados.
MLP (<i>Multi-Layer Perceptron</i>)	Rede neural totalmente conectada composta por camada de entrada, camadas ocultas e camada de saída. Aprende padrões nos dados por meio da técnica de ajuste de pesos chamada <i>backpropagation</i> .	Capacidade de processar grandes volumes de dados.
DT (<i>Decision Tree</i>)	Modelo representado por uma estrutura em árvore, em que cada nó (folha) checa um atributo do dado, levando a uma decisão final.	Simples e fácil de interpretar.
XGBoost (<i>eXtreme Gradient Boosting</i>)	Algoritmo de aprendizado por comitê (<i>ensemble</i>) que constrói um modelo final mais robusto a partir de um conjunto de modelos individuais mais simples de árvores de decisão.	Aumenta a eficiência em relação ao modelo individual e funciona bem com grandes bases de dados.

6.2 Otimização de hiperparâmetros

Somado ao fato de cada algoritmo funcionar de uma forma específica, na definição de cada um deles existem hiperparâmetros, que permitirão que o modelo consiga se ajustar para diferentes bases de dados. Os hiperparâmetros podem definir coeficientes, pesos, limites, arquitetura e complexidade do modelo, margens de erro, número de iterações, tamanho do passo de cada iteração, bem como estabelecer a natureza da relação entre os pontos de uma base de dados. Porém, não é possível saber, a priori, os melhores valores de hiperparâmetros para um determinado problema. Dessa forma, é necessário que exista um processo de otimização, que irá buscar, dentro de um certo intervalo, os valores para os quais cada modelo se ajusta melhor a cada caso específico. Nesse processo, os modelos são testados com diferentes valores de hiperparâmetros em uma série de iterações, visando selecionar aqueles que retornam os melhores resultados. Hiperparâmetros bem otimizados produzem modelos com erros minimizados e desempenho melhorado [59–61]. A técnica de otimização dos hiperparâmetros foi empregada aqui através da biblioteca Optuna [62].

6.3 Técnicas de tratamento de dados

Quando fornecemos dados brutos para os modelos de aprendizado de máquina, eles podem ser de diversos tipos, diferentes ordens de grandeza ou unidades de medida, podem possuir uma relação não linear ou complexa, assim como estão sujeitos a ruídos e erros, ou mesmo não serem significativamente importantes para o modelo. Dessa forma, é necessário que esses dados brutos sejam pré-processados para serem bem interpretados pelos algoritmos [63] (Tabela 2).

Tabela 2 – Descrição das diferentes categorias de pré-processamento de dados.

Categoria	Problema	Técnica	Descrição
<i>Scaling</i>	Mistura de escala e ordem de grandeza entre os dados	RobustScaler	Utiliza técnicas estatísticas para tornar os dados mais padronizados e robustos contra outliers [64].
<i>Oversampling</i>	Desbalanceamento entre as classes da base de dados	SMOTE	Promove o balanceamento dos dados, gerando novos dados sintéticos para a classe minoritária [65].
Seleção de features	Baixa importância de algumas features	SelectFromModel com LR	Utiliza o modelo LR para atribuir um valor de importância às features dos dados, selecionando as mais importantes e removendo as de menor relevância [66].
Detecção de <i>outliers</i>	Erros e anormalidades nos valores dos dados	KNN	Utiliza o modelo KNN para identificar e remover pontos que estão isolados, ou estão em uma região de menor densidade, em relação aos demais [67].
Imputação de dados	Ausência parcial de dados em amostras	IterativeImputer	Modela as features com dados faltantes com base nas outras features e imputa novos pontos para suprir os que estavam ausentes [68].

6.4 Métricas de desempenho

Nesta pesquisa, utilizamos ML para efetuar classificações binárias. Ou seja, somente existem 2 classes possíveis em que o algoritmo pode definir uma amostra, aqui denominadas de positiva (P) e negativa (N). A forma pela qual os modelos aplicados aqui realizam a classificação dos dados é, basicamente, por meio da definição de uma probabilidade, entre 0 e 1 para cada amostra. A partir disso, também será definida uma grandeza, chamada *threshold*, que será o limiar entre as duas classes. Assim, podemos entender que amostras classificadas com probabilidade acima do valor de *threshold* serão consideradas positivas; caso contrário, serão classificadas como negativas. Evidentemente, os modelos podem efetuar a tarefa de classificação de forma correta ou incorreta. Nesse caso, existem 4 possibilidades de resultado: verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN). A partir dos valores de cada uma dessas possibilidades, calculam-se métricas, que são usadas para avaliar o desempenho do classificador binário.

Desse modo, as métricas permitem monitorar a performance de um modelo, avaliando sua capacidade de retornar resultados corretos [69, 70] (Tabela 3).

Tabela 3 – Descrição, intervalo do resultado e equação das métricas de avaliação.

Métrica	Descrição	Intervalo	Equação [69]
Acc (Acurácia)	Retorna a fração de amostras classificadas corretamente [71].	[0,1]	$\text{Acc} = \frac{VP+VN}{VP+VN+FP+FN}$
Especificidade (Esp)	Retorna a capacidade do modelo de identificar corretamente a classe negativa [72].	[0,1]	$\text{Esp} = \frac{VN}{VN+FP}$
Sensibilidade (Sens)	Retorna a capacidade do modelo de identificar corretamente a classe positiva [72].	[0,1]	$\text{Sens} = \frac{VP}{VP+FN}$
F1	Retorna a média harmônica entre a sensibilidade e a precisão, sendo precisão a taxa de verdadeiros positivos na classe positiva [73].	[0,1]	$\text{F1} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}$
MCC (<i>Matthews Correlation Coefficient</i>)	Retorna um coeficiente de correlação que mede a qualidade das classificações [74].	[-1,1]	$\text{MCC} = \frac{VP \cdot VN - FN \cdot FP}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$
AUC (<i>Area Under the Curve</i>)	Retorna a área sob a curva ROC (<i>Receiver Operating Characteristic Curve</i>) [75].	[0,1]	-

A métrica AUC será tratada como a mais importante para este projeto, sendo ela utilizada para a otimização de parâmetros e hiperparâmetros. O diferencial dessa métrica é a sua não dependência direta da definição de um *threshold* ótimo. Sendo assim, o modelo é avaliado de uma forma mais geral, em vez de considerar apenas um caso específico. O AUC, também chamado de AUROC (*Area Under the Receiver Operating Characteristic Curve*), retorna a área sob a curva ROC. Essa curva traz uma relação entre as taxas de verdadeiro positivo e falso positivo para todos os valores de *threshold* testados durante o treino do modelo [69, 76] (Figura 14). Podemos observar que cada ponto da curva corresponde a uma certa taxa de verdadeiros e falsos positivos para cada valor de *threshold*. Nesse caso, a área sob a curva retornou um valor de 0,94. O valor máximo para a métrica seria igual a 1, o que significaria que a curva se aproximou o máximo possível do ponto (0,1), com taxa de verdadeiro positivo máxima e falso positivo mínima [76].

6.5 Métodos de validação

Para realizar a avaliação de um modelo de aprendizado, é necessário dividir as amostras em grupos, sendo eles de treino e de teste. No treino, o modelo vai aprender os padrões nos dados rotulados; no teste, ele irá tentar classificar os dados não rotulados com base no que foi aprendido [50]. Porém, a etapa de dividir a base de dados em grupos

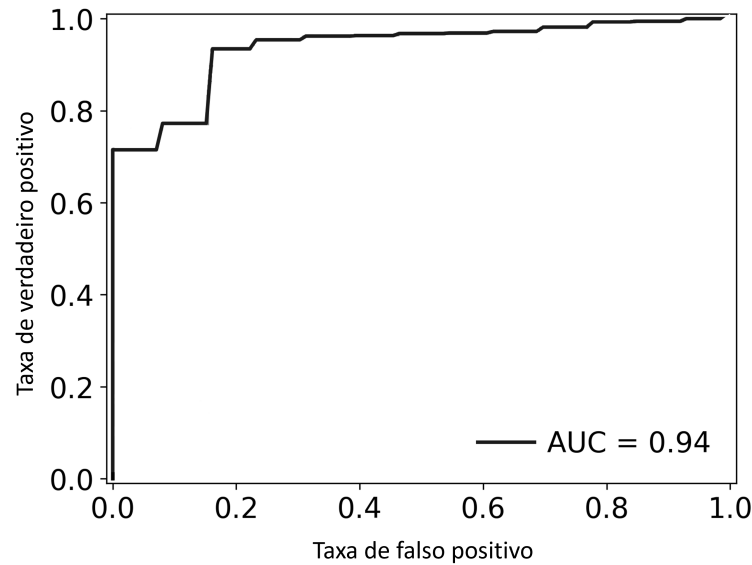


Figura 14 – Demonstração de uma curva ROC para um AUC de 0,94

de treino e teste é de suma importância. Uma divisão simples, com grupos fixos, pode não ser adequada. Nesse caso, poderíamos lidar com algum grupo com baixa quantidade de dados, bem como estaríamos deixando de aproveitar toda a base de dados de forma mais eficaz. Uma técnica, chamada de validação cruzada, visa melhorar a eficiência desse processo [77]. A seguir, será apresentado como funciona a validação cruzada, que recebe o nome K-fold [78], de forma geral e como ela foi empregada na presente pesquisa. Uma variação dessa abordagem, chamada de validação cruzada Nested [79], que foi utilizada aqui, também será demonstrada e a diferença entre ambas abordagens serão explicadas.

6.5.1 Validação cruzada K-fold

A validação cruzada K-fold consiste basicamente em dividir a base de dados em K partes, sendo um grupo destinado para teste e os demais para treino. O que é proposto na validação cruzada K-fold é a ideia de repetir o processo de treino e teste com diferentes subconjuntos de dados, escolhidos de forma aleatória e que não sejam sobrepostos. Dessa forma, o processo será realizado K vezes e haverá um rodízio dos dados presentes em cada grupo, mantendo sempre o grupo de teste diferente do anterior. Isso otimiza o aproveitamento dos dados disponíveis e diminui a incerteza dos resultados, ajudando a contornar o problema de uma base de dados pequena [77]. A Figura 15 ilustra como é efetuada a divisão dos grupos.

Além das repetições com a alternância dos grupos, pode-se ainda adicionar um loop, que fará com que a validação cruzada seja efetuada várias vezes, selecionando grupos aleatórios a cada ciclo. Isso permite aproveitar ainda mais da base de dados para realizar os treinamentos. Ademais, duas etapas de validação cruzada podem ser implementadas no processo. Uma que vai servir para otimização, utilizando os dados para treinar e encontrar



Figura 15 – Demonstração do processo de validação cruzada com 5 divisões.

os melhores parâmetros e hiperparâmetros dos modelos. A outra etapa funcionaria para validar os resultados, empregando os parâmetros e hiperparâmetros ótimos encontrados na etapa de otimização para obter a performance dos modelos treinados. Esse procedimento, envolvendo a validação cruzada aplicada a uma etapa de otimização e de validação foi adotada aqui. Um esquema desse processo é demonstrado na Figura 16. É importante ressaltar que a otimização é feita baseada na métrica AUC. Ou seja, os melhores parâmetros e hiperparâmetros são selecionados a partir dos modelos que retornam os valores mais elevados dessa métrica.

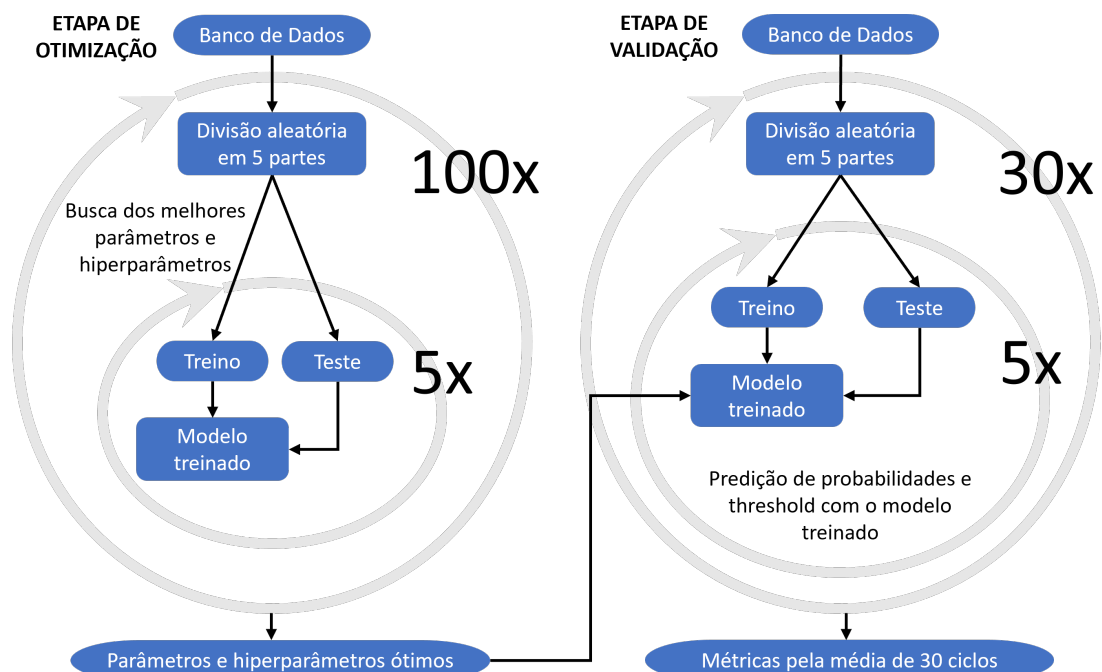


Figura 16 – Esquema do processo de otimização e validação na validação cruzada K-fold utilizado neste trabalho.

Apesar dessa técnica aprimorar a eficácia dos modelos de ML, existe um problema quando ela é usada para otimizar os hiperparâmetros e também validar o modelo. Pois,

nessa situação, dados que já foram usados na otimização dos hiperparâmetros poderão ser reutilizados na etapa de validação. Isso gera uma tendência de resultado enviesado e otimista para a performance do modelo [80]. A validação cruzada Nested é proposta justamente para solucionar esse problema.

6.5.2 Validação cruzada Nested

Um procedimento de validação cruzada K-fold também é aplicado na abordagem da validação cruzada Nested. Neste processo, a etapa de otimização é realizada em um loop interno com grupos diferentes de treino e teste para cada partição do loop externo. Além disso, esses grupos não se misturam com os dados do grupo de teste que são empregados na etapa de validação no loop externo. Ou seja, os procedimentos de otimização e validação são efetuados com grupos de teste diferentes em cada um deles, ao contrário do que foi apresentado na abordagem do K-fold. Esse processo reduz, ou elimina, o risco de um resultado enviesado. Portanto, esse método de validação cruzada pode ser visto como mais consistente do que o anterior. Entretanto, o seu ponto negativo é que o número de iterações aumenta consideravelmente, o que eleva o custo computacional [80]. A Figura 17 mostra um esquema da validação cruzada Nested da forma como foi aplicada nesta pesquisa.

6.6 Interpretabilidade

Em muitos casos, os modelos de aprendizado de máquina podem ser muito complexos, não sendo possível interpretá-los bem. A capacidade de interpretar um modelo é bastante importante. Pois permite entender como o processo está sendo aplicado, aumenta a confiabilidade dos resultados e também fornece informações que podem ser usadas para melhorar o desempenho do modelo. Nesse sentido, novos métodos, que fornecem ferramentas para melhorar a interpretabilidade dos modelos, têm sido propostos. Dentre eles, usamos neste trabalho uma técnica chamada *SHapley Additive exPlanations* (SHAP). O SHAP atribui a cada feature um valor de importância que indica qual foi a influência de cada uma delas na saída final do modelo. Nesse sentido, a soma dos valores de importância de cada feature para um determinado paciente vai definir em qual classe ele será classificado. Ou seja, o SHAP nos fornece uma maneira objetiva de obter informações sobre o funcionamento dos modelos de ML, bem como permite avaliar a contribuição de cada uma das propriedades analisadas [81].

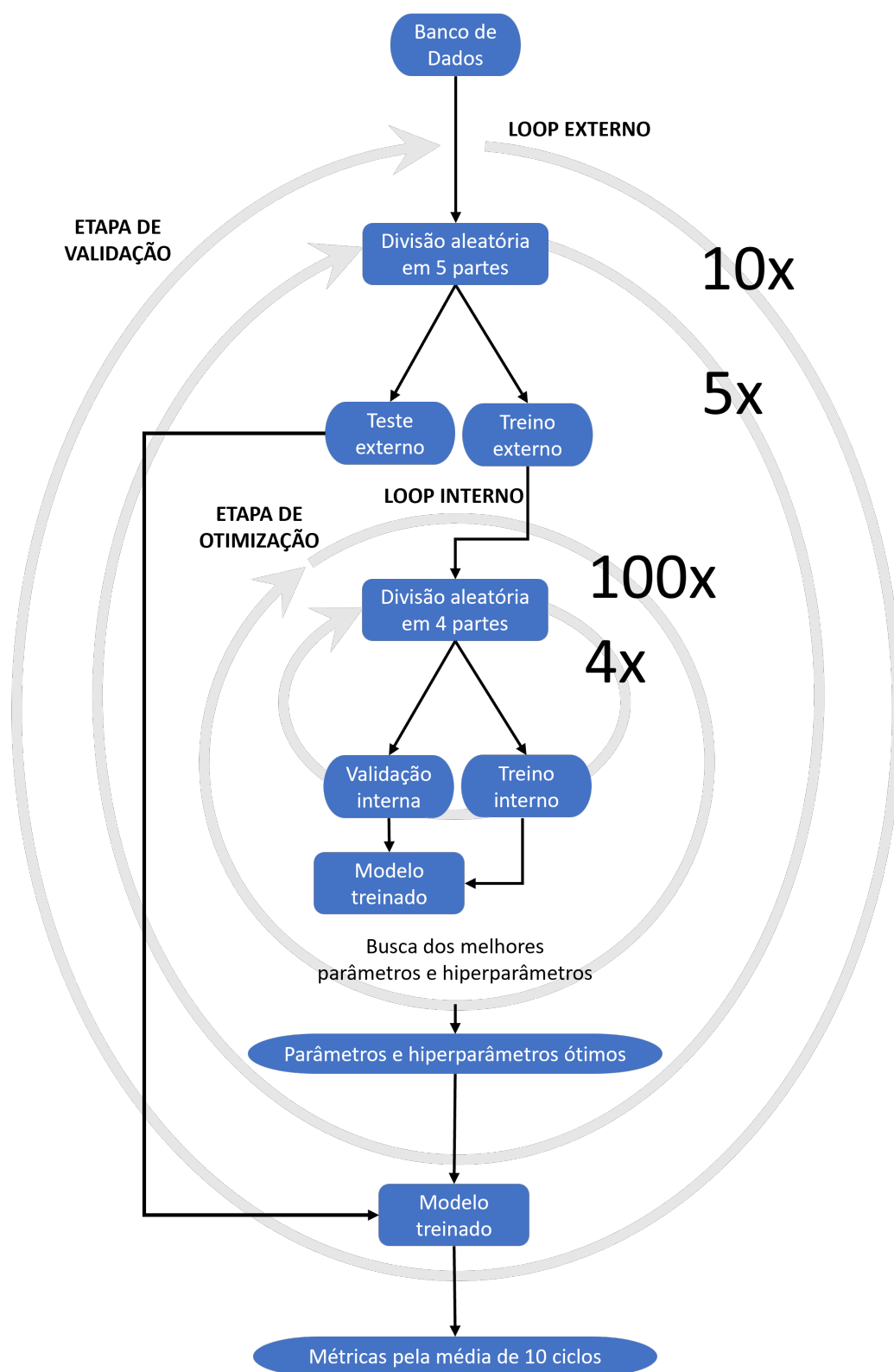


Figura 17 – Esquema do processo de otimização e validação na validação cruzada Nested utilizado neste trabalho.

7 Espectroscopia de Aprendizado de Máquina aplicada ao diagnóstico médico

Como vimos no Capítulo 4, a elipsometria, uma técnica de espectroscopia, retorna uma quantidade muito grande de dados sobre as amostras analisadas. Afinal, são medidas quatro grandezas (Ψ , Δ , dPol e uR), para cada ângulo de incidência, para cada uma das posições medidas em cada amostra e também para cada um dos 674 diferentes comprimentos de onda da luz [32]. Para lidar com isso, o aprendizado de máquina apresenta uma boa solução, pois permite detectar variações discretas e observar padrões pouco evidentes em uma grande base de dados [13]. Nos últimos anos, diversas pesquisas utilizaram a combinação dessas duas abordagens, chamada aqui de espectroscopia de aprendizado de máquina (do inglês Machine Learning Spectroscopy, MLS), para realizar detecções e diagnósticos médicos [17, 22, 82–86].

Diferentes estudos demonstraram a aplicabilidade do método de MLS, especificamente com a técnica de elipsometria. Os trabalhos apresentaram bom desempenho registrado pelas métricas de avaliação dos modelos de aprendizado de máquina. O primeiro a ser citado é o trabalho de Paiva (2025), que realizou a classificação de indivíduos saudáveis e reativos para malária utilizando espectroscopia de aprendizado de máquina em amostras de soro sanguíneo. Outro ponto analisado foi a classificação de amostras coletadas em áreas endêmicas e não endêmicas para a malária [82]. Outro trabalho que podemos citar é dos autores Amaral et al. (2025). A pesquisa analisou linhagens celulares humanas de carcinoma de células escamosas (A431), carcinoma de células escamosas da língua (SCC-4), carcinoma de faringe (FaDU) e células humanas de queratinócitos (HaCat). Foi empregado MLS para diferenciar culturas de células tumorais de células normais, avaliando para cada uma das linhagens citadas [83]. O terceiro estudo a ser citado, de Amaral et al. (2024), fez a classificação de amostras de sêmen obtidas de descendentes machos de camundongos infectados e não infectados pelo Zika Vírus [22]. Por último, citamos o trabalho de Ceccon et al. (2023), com bons resultados para o diagnóstico da infecção por SARS-CoV-2 em amostras de material nasofaríngeo e traqueal [17].

No mundo todo, pesquisas envolvendo aplicação de espectroscopia de aprendizado de máquina no diagnóstico médico estão sendo desenvolvidas [84–86]. Além da elipsometria, outras técnicas de medidas espectroscópicas também são aplicadas. Como exemplos podemos citar os trabalhos de Hano et al. (2024) e de Villazon et al. (2025). Ambas as pesquisas usaram a técnica de espectroscopia Raman associada a análise por aprendizado de máquina. A primeira demonstrou o potencial do método para diagnóstico de câncer de pulmão e a segunda para classificação de diferentes linhagens de células cancerígenas [84, 85].

Um outro trabalho, de Nascimento et al. (2022) evidenciou a aplicação do método com a técnica de espectroscopia de infravermelho para o diagnóstico de COVID-19 [86].

Portanto, por meio dessa pequena revisão bibliográfica, fica evidente que a combinação da elipsometria, ou outras técnicas de espectroscopia, com aplicação de aprendizado de máquina, tem se mostrado uma abordagem promissora para diagnósticos médicos. Diversas áreas de estudo podem ser avaliadas, como foi mostrado para malária, câncer e infecções virais. As pesquisas revisadas demonstram a aplicabilidade de um método simples, eficaz, de baixo custo e rápido para oferecer diagnósticos e estudos biológicos [17, 22, 82–86].

8 Metodologia

8.1 Coleta de amostras

Neste estudo, para o grupo de pacientes recuperados da COVID-19 foram recrutados 22 homens, com idades entre 18 e 50 anos, que foram atendidos no Hospital Mater Dei entre outubro de 2020 e outubro de 2021. Todos eles foram diagnosticados com COVID-19, a partir da detecção do vírus SARS-CoV-2 em swab orofaríngeo utilizando RT-qPCR (*Reverse Transcription Quantitative Polymerase Chain Reaction*). Todos eles também se recuperaram completamente da doença. Após a análise por RT-qPCR para o vírus SARS-CoV-2 em todas as amostras de sêmen, selecionamos aleatoriamente 29 amostras negativas para inclusão neste estudo. Então, nesse caso, essas amostras são de pacientes que foram infectados e posteriormente recuperados, não havendo presença do vírus no sêmen. Para o grupo controle, ou negativo, foram recrutados 13 homens, também com idades entre 18 e 50 anos, sem histórico de infecção por COVID-19. Esses indivíduos compareceram à clínica MF Fertilidade Masculina para avaliações de fertilidade entre maio de 2020 e outubro de 2020.

As amostras de sêmen foram coletadas de todos os participantes após a obtenção do consentimento de cada um. Este estudo foi autorizado pelo Comitê Nacional de Ética em Pesquisa (CONEP) e pelo Comitê de Ética em Pesquisa do Hospital Mater Dei, sob os números de protocolo CAAE: 36122420.5.0000.5128 e CAAE: 18456819.1.0000.5128, respectivamente. Todos os procedimentos foram conduzidos de acordo com os padrões éticos estabelecidos na Lei Nº 14.874, de 28 de maio de 2024 e das Resoluções 466/2012 CNS, 510/2016 CNS e Norma Operacional 001/2013 CNS..

8.2 Análise das amostras de sêmen

Para a análise das amostras seminais, o sêmen de cada paciente foi coletado por meio de masturbação após 2 a 3 dias de abstinência sexual e armazenado em um recipiente estéril. Logo após, seu volume foi medido. Os pacientes foram devidamente orientados em relação aos procedimentos adequados de higiene, incluindo a lavagem das mãos e dos genitais, a fim de prevenir a contaminação viral. As amostras foram então liquefeitas a 37°C por até 60 minutos em incubadora. O mesmo andrologista fez as medidas dos parâmetros espermáticos que serão analisados aqui. Após a análise do sêmen, o ejaculado restante foi imediatamente congelado e armazenado em nitrogênio líquido a -196°C, em um biorepositório estabelecido para este projeto localizado na Clínica MF Fertilidade Masculina, sob a supervisão do médico Marcelo H. Furtado.

As amostras de sêmen foram analisadas conforme as diretrizes da Organização Mundial da Saúde [87]. Os parâmetros seminais foram avaliados em ambos os grupos de pacientes, incluindo a concentração e a motilidade dos espermatozoides, todos medidos dentro de uma hora após a coleta. As amostras foram previamente homogeneizadas para a realização de análises em replicata.

A concentração de espermatozoides foi quantificada utilizando 10 μL da amostra em uma câmara de Neubauer (4x4; Lo-Laboroptik, Reino Unido), foi feita a contagem de pelo menos 200 espermatozoides. O número total de espermatozoides em cada amostra foi obtido multiplicando-se a concentração pelo volume de sêmen coletado.

A motilidade espermática foi avaliada com 10 μL da amostra em uma câmara de contagem Makler (Sefi-Medical Instrument, Haifa, Israel). Para esta análise, pelo menos 200 espermatozoides foram examinados em microscopia óptica com aumento de 400x. Cada espermatozoide foi classificado como progressivos (MP), não progressivos (NPM) e imóveis (IM). A concentração de células redondas também foi determinada utilizando a mesma câmara de contagem Makler.

Para a análise da morfologia dos espermatozoides, foi preparada uma lâmina por esfregaço, espalhando-se 10 μL da amostra sobre a sua superfície. A lâmina foi seca ao ar, fixada com etanol a 95% por 15 minutos e corada pelo método de Papanicolaou. Utilizaram-se os critérios rigorosos de Kruger para a avaliação morfológica. Nessa análise, 200 espermatozoides foram classificados como normais ou anormais quanto à sua cabeça, peça intermediária e cauda.

8.3 Aprendizado de Máquina aplicado aos parâmetros espermáticos

Para realizar a classificação dos pacientes com relação aos parâmetros espermáticos, um programa de classificação por aprendizado de máquina foi construído utilizando a linguagem de programação Python [17, 21, 22, 82]. Seis diferentes algoritmos de classificação foram treinados e avaliados para determinar a classificação das amostras: *Decision Tree* (DT) [52], *eXtreme Gradient Boosting* (XGBoost) [53], *K-Nearest Neighbors* (KNN) [54], *Logistic Regression* (LR) [55], *Multilayer Perceptron* (MLP) [56] e *Support Vector Machine* (SVM) [57].

Os algoritmos de classificação foram otimizados com o uso da biblioteca Optuna [62], utilizando 100 repetições para buscar os melhores parâmetros e hiperparâmetros com o objetivo de melhorar o desempenho dos modelos. As Tabelas 8 e 9 do APÊNDICE A mostram, respectivamente, o espaço de busca adotado para a otimização dos parâmetros do modelo e os hiperparâmetros de cada algoritmo. A otimização foi realizada maximizando a Área sob a Curva ROC (AUROC), ou simplesmente AUC (Área sob a Curva), utilizando validação cruzada K-fold com 5 divisões (*folds*). A probabilidade de classificação de cada

amostra nos 5 *folds* de teste foi determinada durante o processo de validação cruzada. Ao final dos ciclos, a AUROC foi calculada considerando as probabilidades de classificação de todas as amostras. Após a etapa de otimização, os algoritmos foram treinados e validados. Na fase de validação, foram conduzidos 30 experimentos de validação cruzada, cada um com cinco *folds*, gerando estimativas para seis métricas de desempenho diferentes: AUC, Acurácia (Acc), F1, Coeficiente de Correlação de Matthews (MCC), Especificidade (Esp) e Sensibilidade (Sens). Em cada uma das 30 validações cruzadas, as amostras foram distribuídas aleatoriamente (usando diferentes sementes aleatórias) entre os cinco *folds*. Um limiar de classificação móvel (threshold) foi implementado a partir das curvas ROC da base de treinamento, utilizando o critério da menor distância ao ponto (0,1) [88], para a separação entre as classes. Critério esse que permite obter o maior valor possível para a acurácia do modelo. O valor médio do limiar de classificação foi utilizado para calcular as métricas F1, MCC, Acc, Spec e Sens. Um loop externo foi implementado na etapa de validação para avaliar a variância das métricas de desempenho (Figura 16).

Com o objetivo de melhorar o desempenho dos modelos, foi realizada uma seleção manual de features. Para isso, utilizou-se de uma matriz de correlação entre as características para avaliar as que são correlatas e foram feitos testes buscando os melhores resultados. Para compreender melhor a importância das diferentes características no processo de classificação do melhor modelo de ML, foi utilizado o método SHapley Additive exPlanations (SHAP) [81]. O SHAP foi utilizado aqui para oferecer uma maneira consistente e objetiva de fomentar ideias sobre como um modelo de aprendizado de máquina realiza suas previsões e quais características têm maior influência no processo de classificação.

8.4 Medidas de espectroscopia

O sêmen de pacientes do grupo controle e de pacientes recuperados da COVID-19 foi coletado, mantendo uma concentração de aproximadamente 10^7 células por mL, e fixado em etanol a 70%. As amostras foram cuidadosamente preparadas por deposição em gota sobre lamínulas Corning No. 1½, com um volume total de 20 μ L. O material biológico foi então coberto com uma segunda lamínula Corning No. 1½, formando o sistema de camadas da Figura 11.

As medições espectroscópicas seguiram os mesmos protocolos que já foram implementados e testados para o diagnóstico de COVID-19 [17], diagnóstico de malária [82] e para avaliação das alterações espermáticas em camundongos nascidos de mães infectadas com o vírus Zika [21]. As amostras foram montadas em um elipsômetro J.A. Woollam M-2000 e medidas na configuração padrão. Nessa configuração, os espectros dos ângulos elipsométricos Ψ e Δ , o grau de despolarização (dPol) e a reflexão não polarizada (uR), foram adquiridos na faixa de 240 nm a 1700 nm em seis diferentes ângulos de incidência

(45°, 50°, 55°, 60°, 65° e 70°). As medições de elipsometria foram realizadas em 9 diferentes regiões, cobrindo aproximadamente toda a superfície das lamínulas de vidro. Os espectros adquiridos foram então armazenados para posteriormente serem utilizados na análise por ML.

8.5 O problema das reflexões incoerentes

Como foi apresentado no Capítulo 5, de todas as soluções estudadas para resolver o problema das reflexões incoerentes, somente uma foi considerada adequada e aplicada neste trabalho. Foram fixados diferentes tipos de fita adesiva na parte traseira de uma lamínula de vidro transparente. Sendo a mesma lamínula usada na montagem das amostras. As fitas utilizadas aqui foram: fita Durex branca, fita crepe, fita dupla face com proteção de papel, massa adesiva Blu-Tack e fita mágica fosca Scotch. Cada uma das lamínulas montadas foi medida em elipsômetro da mesma forma que foi feita para as amostras. Também medimos uma lamínula limpa, sem nenhuma fita anexada. Além das medidas de elipsometria, também medimos a espessura média das lamínulas utilizando um perfilômetro. Dessa forma, usando como referência a montagem da Figura 11, o objetivo é obter informações da lamínula da primeira camada e do substrato para que possamos tentar obter as constantes ópticas da amostra. Para isso, os dados coletados das lamínulas foram utilizados para criar um modelo óptico das mesmas a partir do programa CompleteEASE. Cada ajuste foi realizado por meio de uma modelagem Cauchy dos dados. Foi adotado como referência o índice de refração tabelado da lamínula de $n = 1,523$ para um comprimento de onda de 589 nm. Após o ajuste de cada lamínula em separado, medimos uma amostra completa, como da Figura 11, mas usando água deionizada (DI), por ter um índice de refração bem conhecido.

8.6 Aprendizado de Máquina aplicado aos dados de espectroscopia

Uma metodologia semelhante de validação cruzada com 5 partições foi utilizada para treinar, otimizar e validar os algoritmos de classificação com os dados de espectroscopia. No entanto, neste caso, o conjunto de dados apresenta alta dimensionalidade, considerando 674 diferentes comprimentos de onda para cada um dos seis ângulos de incidência e para cada uma das propriedades medidas (Ψ , Δ , dPol e uR) em 9 posições diferentes de cada amostra. O que totaliza mais de 145 mil dados por amostra. Após a seleção dos melhores modelos com base na validação cruzada k-fold, foi aplicada uma validação cruzada Nested para avaliar o desempenho dos modelos de forma mais robusta e livre de viés (Figura 17).

8.7 Análise estatística dos dados

Os dados da análise dos parâmetros espermáticos foram apresentados por meio de estatísticas descritivas em cada caso. Foram aplicadas média e desvio padrão (DP) para variáveis contínuas com distribuição normal, mediana e intervalo interquartil (IIQ) para variáveis contínuas sem distribuição normal, e proporções e frequências para variáveis categóricas. Quando necessário, normalização foi aplicada usando a transformação apropriada. A normalidade dos dados foi avaliada por meio do teste de Shapiro–Wilk [89]. Os dados seminais com distribuição normal foram analisados com o teste t de Student [90], enquanto os dados seminais não paramétricos foram submetidos ao teste de Mann-Whitney [91] (teste U) para avaliar possíveis variações nas características seminais. As análises estatísticas foram realizadas utilizando o software GraphPad Prism 8 (GraphPad Software, Inc.), considerando-se um nível de significância de $P < 0,05$.

9 Resultados e Discussão

9.1 Análise das amostras de sêmen

Considerando as diretrizes da Organização Mundial da Saúde [87], os parâmetros seminais analisados, tanto para o grupo controle, como para os recuperados da COVID-19, mostraram-se semelhantes e acima do quinto percentil dos valores de referência da OMS (Tabela 4). Ademais, não houve diferenças significativas entre os valores de todos os parâmetros para cada grupo. Portanto, os resultados estavam dentro de uma faixa normal e consistente com os valores de referência da OMS. A Tabela 4 mostra os resultados obtidos para cada grupo de pacientes, assim como os valores de referência da OMS. Os resultados são apresentados a partir da sua mediana por intervalo interquartil (IIQ). Entre parênteses estão anotados os valores do primeiro quartil e terceiro quartil de cada caso. A Figura 18 mostra como as amostras estão distribuídas entre os valores medidos para cada parâmetro espermático. Ambas as representações dos resultados, Tabela 4 e Figura 18, demonstram como a diferença entre as medições não é expressiva o suficiente para observar uma distinção entre elas. Fazendo uma revisão de trabalhos anteriores que realizaram a mesma análise, observamos que os trabalhos de Guo et al. (2021) [92], Donders et al. (2022) [93] e Paoli et al. (2022) [30] também não obtiveram alterações significativas nos parâmetros espermáticos após a infecção por COVID-19. Portanto, concluímos que uma análise tradicional dos parâmetros seminais não é suficiente para diferenciar indivíduos sem histórico de COVID-19 e aqueles que se recuperaram da infecção. Sendo necessário, assim, a aplicação de uma ferramenta que possibilite uma análise profunda dos dados, buscando padrões que não conseguimos observar. Justificando, nesse caso, a proposta da adoção de aprendizado de máquina para essa tarefa [13].

Tabela 4 – Valores medianos (intervalo interquartil - IIQ) dos parâmetros seminais dos pacientes controle e dos pacientes recuperados de COVID-19, comparado com os valores de referência da OMS. [87].

Características seminais	Pacientes Controle	Pacientes COVID-19	OMS (2021)
Volume de sêmen (mL)	3,0 (2,7 - 3,8)	2,6 (2 - 4,3)	> 1,4
Concentração de espermatozoides (10^6 /mL)	24,5 (14,1 - 54,1)	40,8 (19,3 - 72)	> 16
Número total de espermatozoides (10^6)	83 (48,2 - 135,4)	123,2 (71,2 - 210,6)	> 39
Morfologia normal (%)	4 (3,5 - 4,5)	4 (3 - 5,75)	> 4
Motilidade progressiva (%)	47 (27,5 - 51,5)	50 (42,4 - 64)	> 30
Motilidade não progressiva (%)	9 (8 - 13)	6,4 (5 - 11)	> 1
Espermatozoide imóvel (%)	43 (37 - 60)	41 (26 - 50)	> 20
Células redondas (10^6 /mL)	0,9 (0,8 - 1,4)	2,4 (1,4 - 3,7)	-

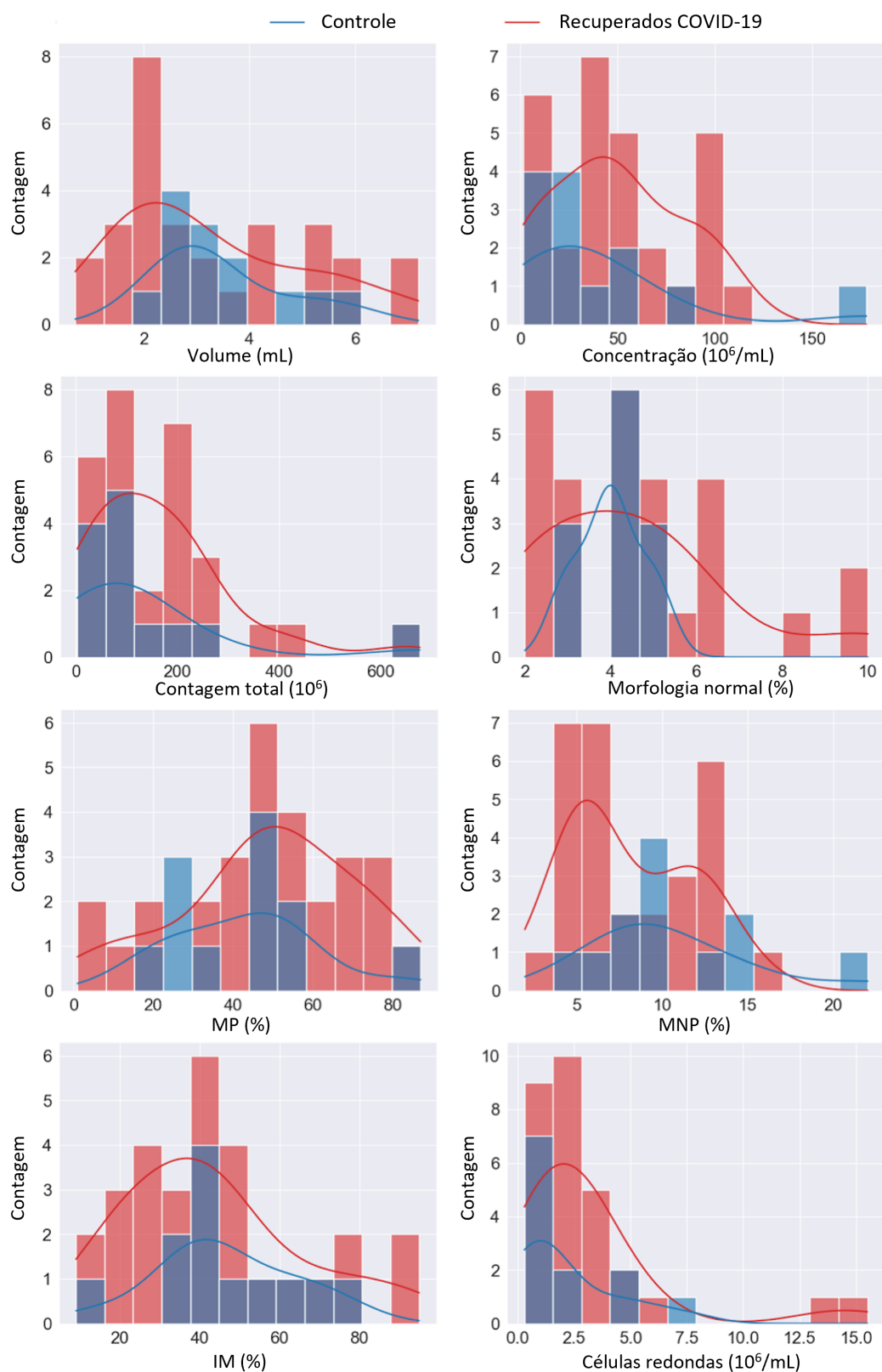


Figura 18 – Comparação entre a distribuição dos parâmetros espermáticos de pacientes-controle (azul) e pacientes recuperados de COVID-19 (vermelho). Parte das distribuições estão sobrepostas, nesse caso a barra fica azul escuro, indicando que a distribuição vermelha está atrás da azul.

9.2 Aprendizado de Máquina aplicado aos parâmetros espermáticos

A Figura 19 mostra a matriz de correlação para os parâmetros espermáticos criada para avaliar e realizar a seleção manual das features para melhorar a performance dos modelos de *Machine Learning*. Como exemplo, podemos notar a forte correlação entre Contagem total e Concentração, que vem do fato de que elas estão relacionadas de forma que: $\text{Contagem total} = \text{Concentração} \times \text{Volume}$. Outra forte correlação observada foi entre MP e IM, sendo que $\text{MP} + \text{IM} + \text{MNP} = 100\%$. Ademais, percebemos que MNP apresenta valores baixos em todos os casos. Entretanto, avaliando o desempenho dos modelos de ML, somente a eliminação da feature Concentração produziu uma melhora nos resultados. Então todas as demais foram mantidas para o processo de classificação das amostras, sendo elas: Volume, Contagem total, Morfologia normal, MP, MNP, IM e Células redondas.

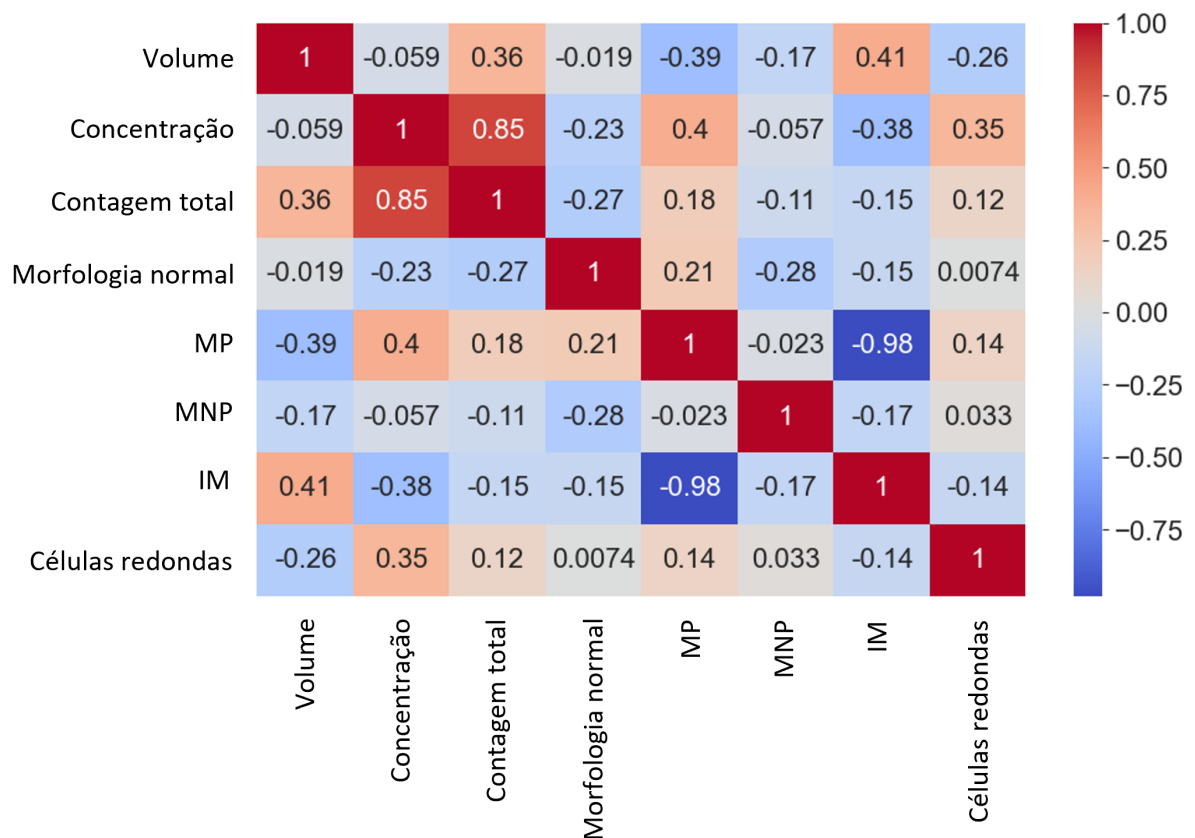


Figura 19 – Matriz de correlação para todos os parâmetros espermáticos.

A Tabela 5 apresenta o desempenho de cada um dos 6 algoritmos de aprendizado de máquina utilizados na classificação das amostras de sêmen a partir dos seus parâmetros espermáticos. Na tabela, a métrica “AUC treino” é aquela obtida na base de dados de treino, já “AUC teste” é obtida na base de teste, que possui um caráter menos enviesado. Th corresponde ao threshold ótimo avaliado para cada modelo.

Avaliando os resultados, percebemos que os modelos SVM e MLP obtiveram os melhores desempenhos na classificação. Podemos destacar o resultado para o SVM, que

Tabela 5 – Métricas de desempenho e valor ótimo de threshold obtidos para os modelos de classificação dos parâmetros espermáticos.

Modelo	AUC treino	AUC teste	F1	MCC	Acc (%)	Esp (%)	Sens (%)	Th (%)
KNN	1,00 ± 0,00	0,72 ± 0,05	0,74 ± 0,09	0,44 ± 0,09	70 ± 8	80 ± 12	66 ± 14	47 ± 18
LR	0,75 ± 0,01	0,56 ± 0,04	0,65 ± 0,08	0,22 ± 0,05	60 ± 6	67 ± 11	56 ± 12	98 ± 6
DT	1,00 ± 0,00	0,71 ± 0,05	0,81 ± 0,04	0,42 ± 0,11	75 ± 5	61 ± 8	81 ± 7	100 ± 5
SVM	0,99 ± 0,01	0,81 ± 0,05	0,82 ± 0,05	0,53 ± 0,08	77 ± 5	77 ± 12	78 ± 11	57 ± 9
XGBoost	0,95 ± 0,01	0,75 ± 0,04	0,86 ± 0,04	0,56 ± 0,07	81 ± 4	65 ± 11	88 ± 9	52 ± 14
MLP	0,99 ± 0,01	0,76 ± 0,04	0,81 ± 0,06	0,52 ± 0,08	77 ± 5	81 ± 6	75 ± 8	96 ± 7

apresentou um AUC teste = $0,81 \pm 0,05$, F1 = $0,82 \pm 0,05$, MCC = $0,53 \pm 0,08$ e um ótimo equilíbrio entre Acc, Esp e Sens.

Para a interpretabilidade, a Figura 20 apresenta o gráfico da análise do SHAP para o modelo SVM. Valores SHAP positivos contribuem para classificar um paciente como pertencente ao grupo recuperado da COVID-19, e valores SHAP negativos contribuem para classificá-lo como pertencente ao grupo controle.

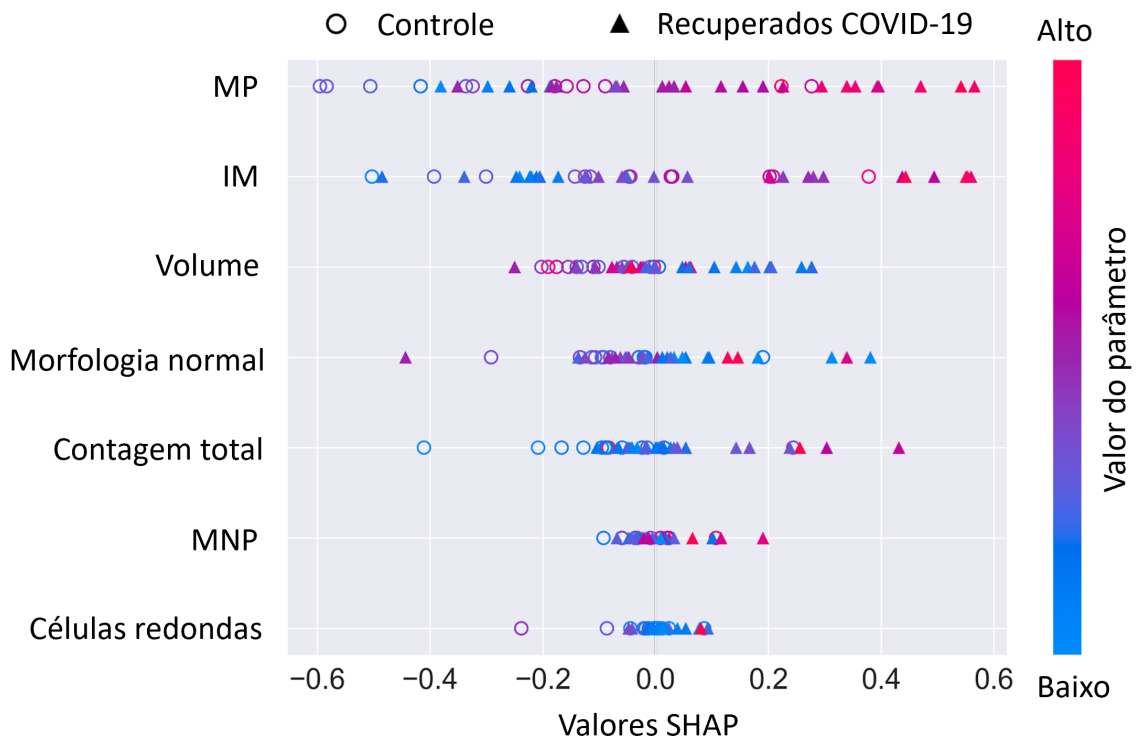


Figura 20 – Gráfico SHAP estratificado por classe para explicar a importância das diferentes características no modelo para pacientes-controle (círculos) e pacientes recuperados de COVID-19 (triângulos). O eixo vertical lista os parâmetros espermáticos e o eixo horizontal mostra os valores SHAP, indicando a contribuição de cada característica para a previsão do modelo. A barra de cores no lado direito varia de azul (valor baixo) a vermelho (valor alto), refletindo a magnitude dos valores das características. Padrões de contribuição semelhantes ou divergentes entre os dois grupos podem ser observados através da posição e cor dos pontos.

Portanto, as features mais importantes para a classificação foram MP e IM. Observamos também que o valor SHAP aumenta junto com o valor do parâmetro para os casos, por exemplo, de MP, IM e Contagem total. Mas o oposto acontece para o Volume. Podemos concluir então, de forma geral, que uma amostra com valor alto de MP, IM e Contagem total, mas baixo para Volume, tende a ser classificada como sendo da classe 1, ou seja, de um paciente recuperado da COVID-19.

É interessante observar no resultado do SHAP que MP (motilidade progressiva) e IM (imotilidade) foram os dois fatores mais importantes para classificar as amostras entre controle e pacientes recuperados da COVID-19. A análise estatística tradicional com resultados apresentados anteriormente na Tabela 4 não demonstra nenhuma diferença significativa entre as classes. Enquanto que, por meio da análise SHAP, conclui-se que altos valores de MP e IM estão associados ao grupo recuperado da COVID-19. Seria realmente esperado que os efeitos negativos da doença no sistema reprodutor masculino possam levar a um aumento dos valores de IM [94,95].

É perceptível o cenário complexo que temos e que permite muito avanço em novas pesquisas. Em pacientes recuperados da COVID-19, a azoospermia (ausência total de espermatozoides no ejaculado) foi observada em 14,3% a 18,6% dos casos, com forte correlação com a gravidade da doença [94,96]. A oligozoospermia (contagem de espermatozoides abaixo do normal) foi observada em 7% dos pacientes [94]. Reduções significativas na concentração espermática e no número total de espermatozoides foram relatadas em vários estudos [9,92,93,96,97], enquanto outros não encontraram alterações significativas antes e depois da infecção [95,98]. A motilidade progressiva também apresentou queda significativa em alguns estudos [92,93,99], embora outros estudos, semelhantes ao nosso, não tenham encontrado diferença significativa [98,100]. Além disso, foi observada uma redução na morfologia normal dos espermatozoides tanto em pacientes com COVID-19 quanto em recuperados [9,100], enquanto alguns estudos não encontraram diferenças significativas [92]. O volume seminal geralmente não apresentou diferença significativa em alguns desses trabalhos [92], embora uma redução tenha sido relatada em outros [98,99,101]. Um aumento nas células redondas foi observado em pacientes recuperados da COVID-19 [9], porém não houve diferença significativa nos leucócitos entre os diferentes grupos de pacientes [94]. Interessantemente, alguns estudos encontraram melhores resultados para a concentração espermática e número total de espermatozoides em pacientes recuperados, em comparação com o grupo controle de inférteis, com maior motilidade progressiva [30]. Essa variabilidade de resultados destaca a complexidade e individualidade do impacto da COVID-19 na qualidade do sêmen, indicando que, embora a COVID-19 afete negativamente o sêmen, os efeitos podem variar amplamente entre os indivíduos. Portanto, a pesquisa realizada aqui fornece avanços valiosos para a área, mas também ressalta a necessidade e a importância de investigações adicionais na área.

9.3 Medidas de espectroscopia

Dentre os dados medidos a partir da técnica de elipsometria para cada amostra, a Figura 21 traz um exemplo dos valores médios que foram obtidos para cada propriedade considerando o ângulo de incidência de 55° . É perceptível que os espectros são bastante afetados na região do comprimento de onda menor que 350 nm. Isso ocorre porque a lamínula superior da amostra possui uma forte absorção óptica nessa região. Observando o gráfico de Δ , notamos que, para comprimentos de onda maiores que 350 nm, o espectro permanece aproximadamente estável próximo a um valor de 180° . Essa característica do gráfico indica que, em toda essa região, a amostra se comporta de forma muito similar a um material transparente. Outra observação pode ser feita na região próxima de 1000 nm no gráfico de uR. A grande variação sofrida se deve ao fato do espectrômetro utilizado para medir essa parte dos dados da região NIR é diferente do usado para valores de comprimento de onda menores. Já para valores maiores, oscilações de espessura da amostra produzem as variações no espectro de uR, bem como minimiza a diferença entre os espectros médios de cada classe para as outras grandezas (Ψ , Δ e dPol). No geral, as maiores diferenças entre uma classe e outra são observadas para a região do visível.

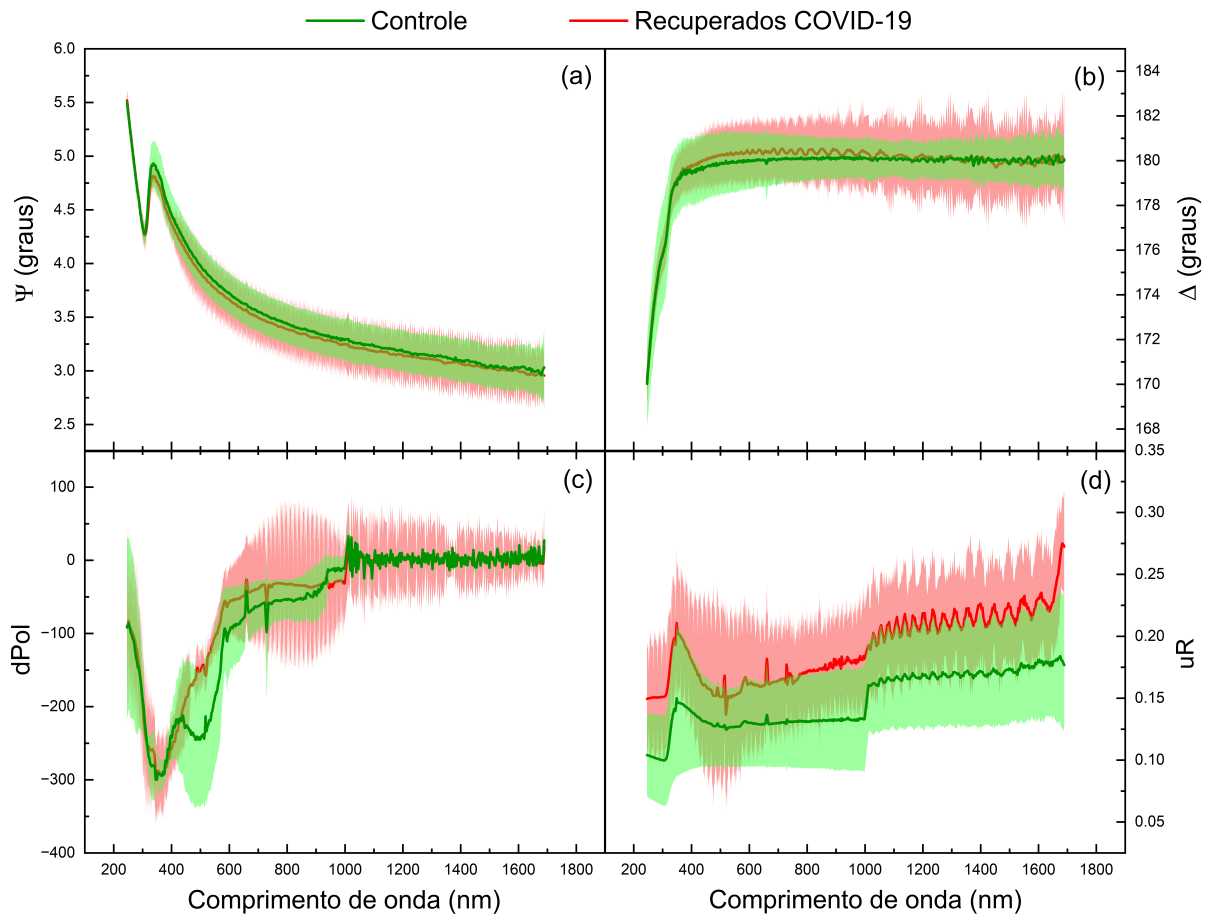


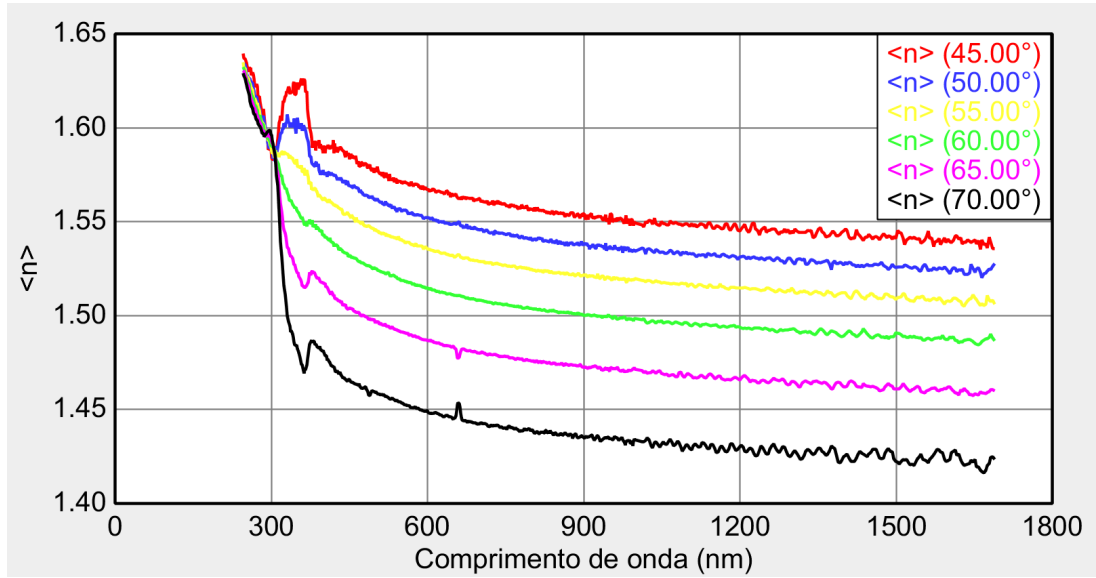
Figura 21 – Espectro médio (linha sólida) e desvio padrão (região sombreada) obtidos de Ψ , Δ , dPol e uR para ambas as classes de amostras.

É importante enfatizar que os espectros (linhas sólidas) apresentados nos gráficos da Figura 21 correspondem aos valores médios das medidas. Isso significa que cada classe possui uma variedade de amostras com diferentes medidas ao longo dos comprimentos de onda, distribuídas em torno desse valor médio. Essa variação é representada pela região sombreada no gráfico, que indica o desvio padrão das medidas. Observa-se que as regiões sombreadas dos grupos controle (verde) e de pacientes recuperados (vermelho) se sobrepõem em praticamente todo o espectro das quatro grandezas. Portanto, a discreta diferença entre as classes observada entre as linhas dos gráficos só é visível por se tratar da média. Porém, se consideramos uma amostra individual qualquer em meio a todos os dados, tradicionalmente seria impossível identificar algum padrão que torne possível distingui-la das demais e classificá-la como de uma classe ou outra. Nesse sentido, fica evidente a necessidade da aplicação de uma abordagem de aprendizado de máquina para buscar padrões entre os dados e realizar a classificação das amostras da melhor forma possível.

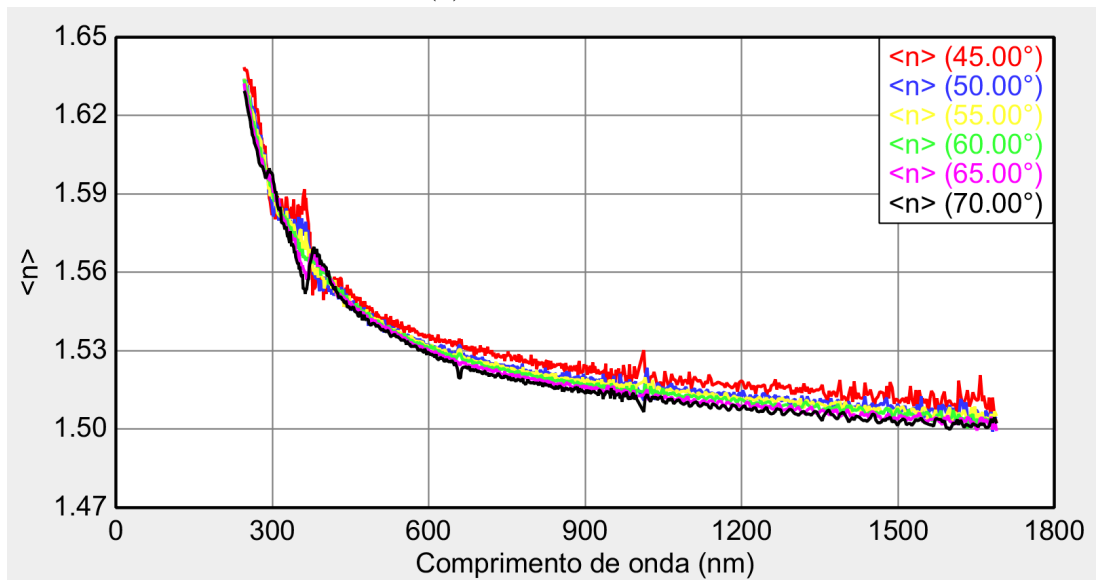
9.4 O problema das reflexões incoerentes

A partir das várias medidas que fizemos utilizando diferentes fitas adesivas anexadas às lamínulas, selecionamos a que apresentou os melhores resultados, sendo ela a da fita mágica Scotch fosca (Figura 22). Para a lamínula limpa (sem fita), nota-se que o valor do seu índice de refração tem uma dependência muito forte com o ângulo de incidência da luz. É justamente a presença de reflexões incoerentes advindas da superfície traseira do vidro que tornam os resultados inconsistentes (Figura 22a). Já para o gráfico (b), é perceptível como as linhas se uniram ao utilizar a fita mágica Scotch na superfície traseira da lamínula, convergindo para um valor de índice de refração mais bem definido (Figura 22b). Assim como abordado no Capítulo 5, a fita promove o espalhamento da luz que chega à sua superfície, desse modo, a maior parte dela não é refletida até o detector. Nesse caso, as reflexões incoerentes causam menor impacto nas medidas.

Um outro ponto que podemos analisar para avaliar o resultado é a comparação com o índice de refração esperado para a lamínula ($n = 1,523$ para um comprimento de onda de 589 nm). Após modelar opticamente cada lamínula por meio da equação de Cauchy, obtivemos os modelos que melhor as descrevem. Para a lamínula limpa, o índice de refração encontrado para 589 nm foi de $n = 1,53439$. Enquanto que para a lamínula com a fita Scotch chegamos em $n = 1,52484$. Para o primeiro caso, foi utilizada uma opção do programa que inclui uma correção matemática para considerar as reflexões incoerentes da superfície traseira do substrato no cálculo. Apesar de ter sido possível chegar, com este método, a um resultado razoavelmente próximo, a aplicação da fita na superfície se demonstrou uma técnica superior. Afinal, a diferença entre o índice de refração obtido para a lamínula com a fita e o seu valor tabelado ($n = 1,523$) está na terceira casa decimal.



(a) Lamínula limpa



(b) Lamínula com fita mágica Scotch

Figura 22 – Gráfico do índice de refração em função do comprimento de onda para diferentes ângulos para as lamínulas (a) limpa e (b) com fita mágica Scotch.

Embora tenhamos chegado a resultados satisfatórios para essa solução, é importante lembrar que esses dados obtidos são apenas para uma lamínula individual. O sistema de multicamadas que montamos nossas amostras é muito mais complexo para modelar. Ademais, infelizmente, a solução da fita adesiva é insuficiente, pois resolve o problema apenas para a superfície traseira do substrato. Para avaliar o nosso caso, foram feitas medidas com uma amostra de água deionizada (DI), que possui um valor de índice de refração bem conhecido de $n = 1,33336$ para um comprimento de onda de 589 nm [102]. Nesse caso, a amostra montada era então composta por vidro/água DI/vidro/fita adesiva. Usando as medidas anteriores para as lamínulas individuais com fita e sem fita, modelamos as medidas de água deionizada. O resultado obtido para o índice de refração, a um comprimento de onda de 589 nm, foi de $n = 1,52962$. A diferença entre o resultado obtido e o valor esperado é muito grande e demonstra quão distante esta solução está de produzir resultados satisfatórios para uma amostra montada em multicamadas. Usando a medida com um ângulo de 70° como exemplo, observamos que a leitura do índice de refração dessa medida é bastante ruidosa e apresenta variações significativas (Figura 23). As oscilações, manifestadas principalmente na região do NIR, são padrões de interferência decorrentes da variação de espessura da amostra. Variações essas que não deveriam estar presentes em uma medida fidedigna.

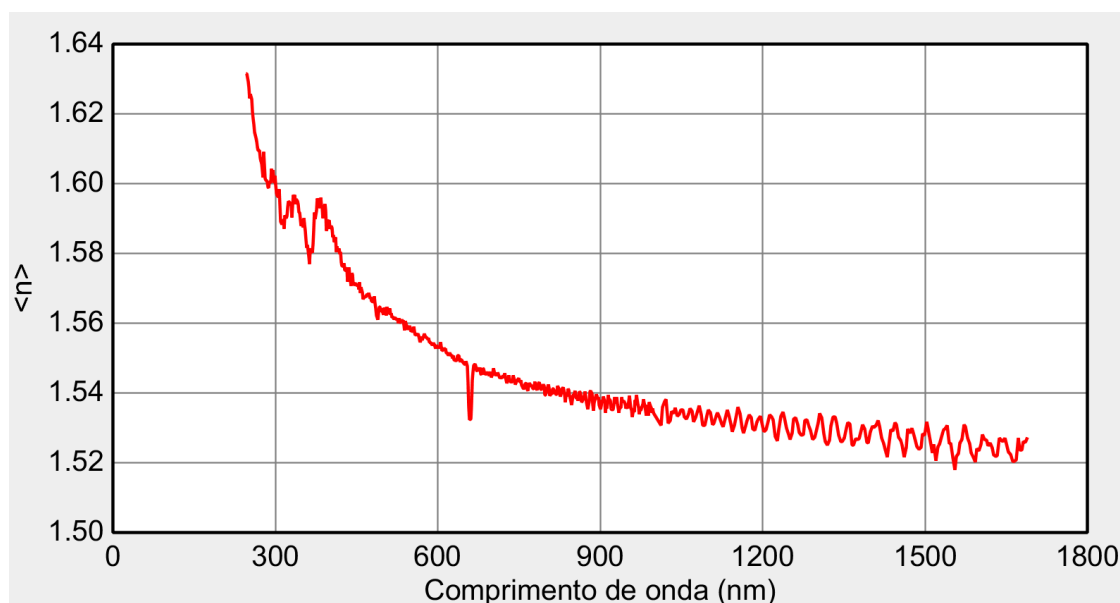


Figura 23 – Gráfico do índice de refração em função do comprimento de onda para o ângulo de 70° medido com água deionizada em uma montagem vidro/água DI/vidro/fita.

Portanto, não foi possível realizar uma modelagem precisa das medidas de elipsometria das nossas amostras para obter as constantes ópticas das mesmas. Entretanto, apresentamos uma metodologia alternativa que, em vez de envolver a classificação das amostras a partir das suas propriedades ópticas, fazemos a classificação utilizando as medições brutas da elipsometria. Afinal, como vimos, as medidas espectroscópicas são

influenciadas pelas propriedades físicas da amostra, que por sua vez, são influenciadas pelas alterações provocadas no material biológico por conta da infecção viral. Nesse sentido, a nossa técnica é adequada e viável. Ademais, a abordagem por aprendizado de máquina é bastante adequada para lidar com os dados de elipsometria, por envolver análise de uma base de dados muito grande e com diferenças sutis entre as classes.

9.5 Aprendizado de Máquina aplicado aos dados de espectroscopia

A Tabela 6 apresenta o desempenho da análise por aprendizado de máquina, utilizando a abordagem com validação cruzada K-fold, para os dados obtidos por elipsometria. As métricas são mostradas juntas ao valor do threshold para cada algoritmo e propriedade da elipsometria, bem como com o valor do ângulo de incidência que produziu o melhor resultado no processo de otimização. Para cada resultado apresentado aqui, um conjunto de parâmetros e hiperparâmetros foi otimizado de forma a buscar o melhor desempenho. No APÊNDICE B, a Tabela 10 mostra o resultado dessa otimização para os parâmetros, e a Tabela 11 para os hiperparâmetros dos modelos.

Tabela 6 – Métricas de desempenho e valor ótimo de threshold obtidos com a validação cruzada K-fold para todos os modelos de MLS.

Modelo	Prop	Ângulo	AUC treino	AUC teste	F1	MCC	Acc (%)	Esp (%)	Sens (%)	Th (%)
KNN	Ψ	45	$0,90 \pm 0,01$	$0,75 \pm 0,06$	$0,20 \pm 0,10$	$0,10 \pm 0,10$	68 ± 3	90 ± 4	18 ± 8	50 ± 1
	Δ	65	$0,83 \pm 0,01$	$0,74 \pm 0,05$	$0,30 \pm 0,10$	$0,20 \pm 0,10$	69 ± 4	90 ± 5	20 ± 10	55 ± 2
	dPol	60	$0,89 \pm 0,01$	$0,85 \pm 0,05$	$0,62 \pm 0,09$	$0,50 \pm 0,10$	77 ± 5	83 ± 4	60 ± 10	56 ± 2
	uR	70	$0,92 \pm 0,01$	$0,91 \pm 0,03$	$0,81 \pm 0,04$	$0,73 \pm 0,06$	89 ± 2	94 ± 3	78 ± 5	56 ± 2
LR	Ψ	45	$0,81 \pm 0,01$	$0,80 \pm 0,04$	$0,61 \pm 0,05$	$0,40 \pm 0,09$	70 ± 5	67 ± 7	76 ± 8	21 ± 2
	Δ	45	$0,89 \pm 0,01$	$0,73 \pm 0,07$	$0,02 \pm 0,05$	$0,04 \pm 0,09$	69 ± 1	100 ± 1	1 ± 3	71 ± 2
	dPol	50	$0,92 \pm 0,01$	$0,86 \pm 0,05$	$0,66 \pm 0,05$	$0,50 \pm 0,08$	78 ± 4	81 ± 5	70 ± 7	41 ± 3
	uR	60	$0,93 \pm 0,01$	$0,92 \pm 0,06$	$0,85 \pm 0,03$	$0,79 \pm 0,05$	91 ± 2	94 ± 2	84 ± 5	67 ± 2
SVM	Ψ	55	$0,75 \pm 0,02$	$0,64 \pm 0,07$	$0,71 \pm 0,05$	$0,20 \pm 0,10$	63 ± 6	50 ± 10	67 ± 7	51 ± 2
	Δ	65	$0,82 \pm 0,01$	$0,67 \pm 0,08$	$0,82 \pm 0,03$	$0,30 \pm 0,20$	73 ± 5	30 ± 10	91 ± 6	52 ± 2
	dPol	50	$0,91 \pm 0,01$	$0,93 \pm 0,04$	$0,86 \pm 0,03$	$0,67 \pm 0,08$	82 ± 4	95 ± 8	77 ± 3	59 ± 2
	uR	50	$0,95 \pm 0,01$	$0,95 \pm 0,04$	$0,94 \pm 0,02$	$0,80 \pm 0,07$	92 ± 3	81 ± 9	96 ± 1	27 ± 4
XGBoost	Ψ	60	$0,91 \pm 0,01$	$0,80 \pm 0,04$	$0,81 \pm 0,01$	$0,02 \pm 0,10$	69 ± 2	2 ± 3	99 ± 2	32 ± 4
	Δ	60	$0,82 \pm 0,01$	$0,74 \pm 0,05$	$0,80 \pm 0,01$	$-0,04 \pm 0,10$	67 ± 2	3 ± 5	96 ± 2	21 ± 3
	dPol	50	$0,93 \pm 0,01$	$0,91 \pm 0,03$	$0,88 \pm 0,03$	$0,60 \pm 0,10$	82 ± 4	60 ± 20	91 ± 5	34 ± 6
	uR	45	$0,95 \pm 0,01$	$0,94 \pm 0,04$	$0,90 \pm 0,03$	$0,60 \pm 0,10$	85 ± 5	60 ± 20	96 ± 1	30 ± 6
MLP	Ψ	45	$0,88 \pm 0,01$	$0,84 \pm 0,04$	$0,86 \pm 0,03$	$0,50 \pm 0,10$	79 ± 5	40 ± 20	96 ± 3	47 ± 4
	Δ	65	$0,90 \pm 0,01$	$0,72 \pm 0,07$	$0,82 \pm 0,01$	$0,01 \pm 0,04$	69 ± 1	$0,3 \pm 1$	100 ± 1	11 ± 4
	dPol	50	$0,93 \pm 0,01$	$0,93 \pm 0,03$	$0,87 \pm 0,04$	$0,67 \pm 0,08$	84 ± 5	90 ± 10	81 ± 8	52 ± 9
	uR	50	$0,96 \pm 0,01$	$0,95 \pm 0,04$	$0,93 \pm 0,02$	$0,80 \pm 0,07$	91 ± 3	89 ± 9	92 ± 3	42 ± 7
DT	Ψ	45	$0,77 \pm 0,01$	$0,77 \pm 0,06$	$0,52 \pm 0,08$	$0,40 \pm 0,10$	73 ± 5	85 ± 6	47 ± 9	50 ± 1
	Δ	55	$0,71 \pm 0,01$	$0,63 \pm 0,08$	$0,30 \pm 0,20$	$0,06 \pm 0,20$	63 ± 6	80 ± 8	30 ± 10	55 ± 2
	dPol	50	$0,85 \pm 0,01$	$0,84 \pm 0,04$	$0,61 \pm 0,09$	$0,40 \pm 0,10$	75 ± 5	81 ± 5	60 ± 10	50 ± 1
	uR	55	$0,83 \pm 0,01$	$0,88 \pm 0,06$	$0,74 \pm 0,08$	$0,60 \pm 0,10$	83 ± 5	85 ± 5	78 ± 9	50 ± 1

Observamos que, de forma geral, todos os modelos, independentemente do ângulo ou da propriedade, apresentaram resultados bons ou ótimos (Tabela 6). Mas podemos dar destaque para três modelos que alcançaram um ótimo AUC teste de aproximadamente 0.95, sendo eles: SVM + uR, XGBoost + uR e MLP + uR. Nota-se ainda que, em todos os casos, a propriedade uR produziu o melhor resultado.

Uma outra forma de observar nossos resultados é através do gráfico de distribuição de probabilidades. Como exemplo, é mostrado na Figura 24 o histograma normalizado da

distribuição de probabilidades obtido para o modelo SVM + uR com ângulo de incidência de 50° . Podemos ver nesse histograma como foi realizada a definição da probabilidade de cada amostra e como o threshold médio (linha vertical pontilhada) divide as classes. As linhas sólidas representam as funções de densidade de probabilidade. Elas são obtidas por meio da *Kernel Density Estimation* (KDE), uma técnica estatística capaz de estimar a densidade de probabilidade de uma distribuição a partir de um conjunto de dados calculados ou medidos.

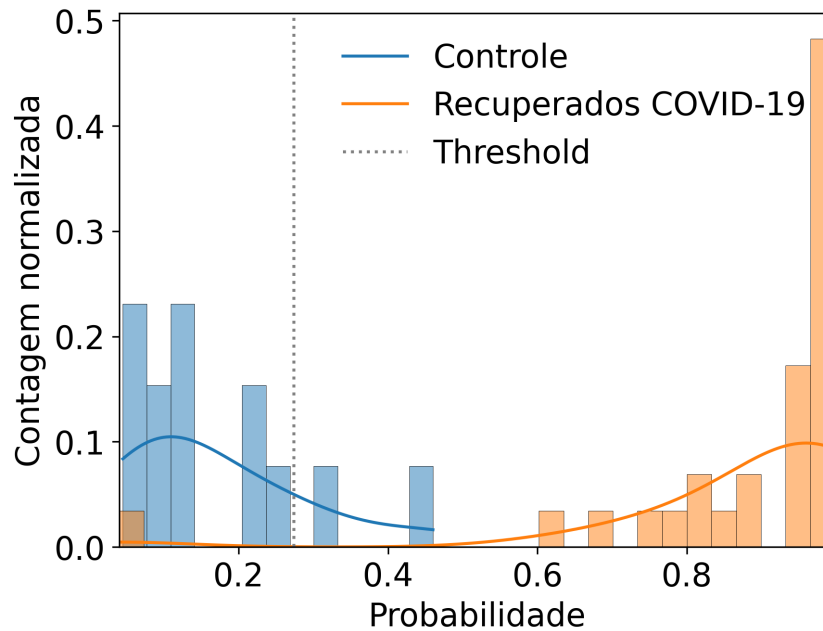


Figura 24 – Histograma normalizado das probabilidades de classificação das amostras de controle (barras azuis) e das amostras de pacientes recuperados de COVID-19 (barras laranja), obtidas com SVM, uR e com ângulo de incidência de 50° . As linhas sólidas azuis e laranjas representam as funções de densidade de probabilidade. A linha pontilhada vertical representa o valor do threshold que divide as duas classes.

Como vimos no Capítulo 6, os classificadores podem ser representados e avaliados por uma relação entre a taxa de verdadeiro positivo e taxa de falso positivo dada pela curva ROC. Sendo que é a área sob essa curva que nos fornece o valor da métrica AUC. Nesse caso, é apresentada na Figura 25 o exemplo da curva ROC para o modelo SVM + uR com ângulo de incidência de 50° .

Os melhores modelos selecionados pela abordagem de validação cruzada K-fold foram avaliados também através da validação cruzada Nested para obter um resultado menos enviesado. Os resultados das métricas e valor do threshold para cada modelo são exibidos na Tabela 7. Todos os modelos apresentaram resultados excelentes para a classificação das amostras. Todos eles obtiveram o mesmo AUC teste, mas avaliando as outras métricas, notamos que o modelo SVM + uR se destaca como o melhor e mais equilibrado desempenho. A redução geral dos valores das métricas observada ao comparar

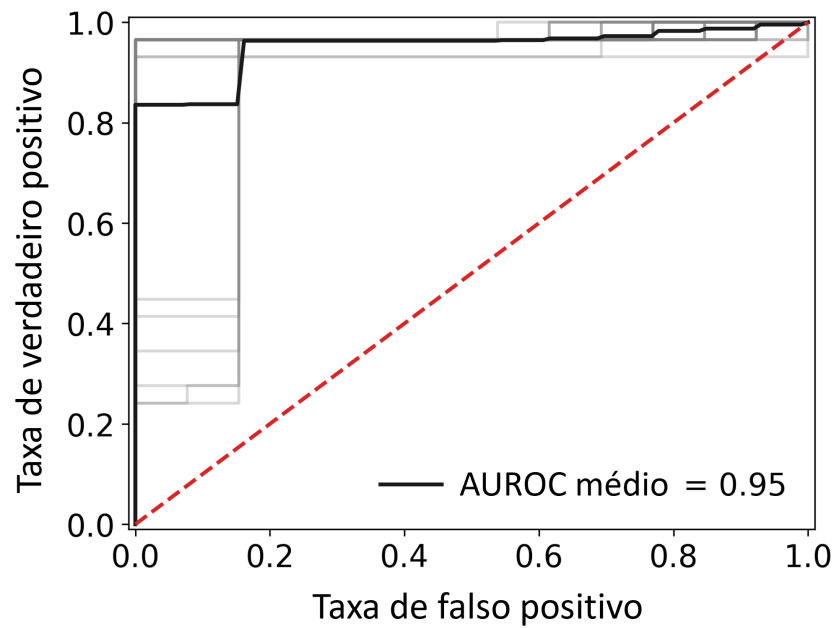


Figura 25 – Curva ROC obtida para o modelo SVM + uR com ângulo de incidência de 50° . A linha em preto mostra o valor médio da curva ROC em relação a todas repetições efetuadas no processo de otimização, que são mostradas pelas linhas em cinza claro. A linha pontilhada em vermelho simboliza um classificador aleatório, ou seja, um modelo com $AUC = 0.50$.

os resultados obtidos pelas validações cruzada K-fold e Nested (Tabelas 6 e 7) era esperada. Essa variação negativa se deve ao fato da validação cruzada Nested produzir resultados com um menor viés, visto que as amostras utilizadas na base de teste da etapa de validação são sempre diferentes dos dados aplicados na etapa de otimização. Essa separação mais bem definida das bases de dados torna os resultados menos otimistas, porém, mais confiáveis.

Tabela 7 – Métricas de desempenho obtidas com a validação cruzada nested para os modelos de melhor desempenho na validação cruzada Nested.

Modelo	AUC treino	AUC teste	F1	MCC	Acc (%)	Esp (%)	Sens (%)
MLP + uR	$0,98 \pm 0,03$	$0,9 \pm 0,2$	$0,9 \pm 0,2$	$0,8 \pm 0,4$	90 ± 20	80 ± 30	90 ± 20
XGBoost + uR	$0,98 \pm 0,03$	$0,9 \pm 0,1$	$0,85 \pm 0,09$	$0,8 \pm 0,3$	90 ± 10	80 ± 40	90 ± 10
SVM + uR	$0,96 \pm 0,03$	$0,9 \pm 0,1$	$0,9 \pm 0,2$	$0,6 \pm 0,2$	90 ± 10	90 ± 20	90 ± 20

Dessa forma, além do modelo SVM ter apresentado o melhor desempenho na análise dos parâmetros espermáticos, ele também foi o que melhor desempenhou para os dados de espectroscopia. Além disso, é bastante claro que a propriedade uR da elipsometria é aquela que permitiu a melhor distinção entre as classes, produzindo resultados de desempenho mais elevados.

Como vimos no Capítulo 4, a propriedade uR obtida da elipsometria, sendo ela a intensidade refletida não polarizada, possui uma dependência com o índice de refração do material medido, assim como as demais propriedades. Sabemos que alterações nas propriedades físicas de um meio produzem variações na interação da luz com o mesmo.

Dessa forma, fica evidente que as medidas que fazemos são capazes de distinguir uma amostra da outra porque elas trazem uma informação relacionada com as propriedades físicas da amostra, que por sua vez, são influenciadas pelas alterações morfológicas e químicas que são induzidas em resposta à infecção viral [22, 36–38]. Mas vale ressaltar que todas as amostras de sêmen analisadas foram testadas negativas para a presença do vírus SARS-CoV-2 por meio do exame RT-PCR. Então, as alterações provocadas no material biológico não podem ser atribuídas de forma direta à infecção das células espermáticas pelo vírus.

10 Conclusões

Os resultados alcançados aqui comprovaram que nosso método pode realizar uma classificação acurada na distinção do sêmen de pacientes que nunca foram diagnosticados com COVID-19 e o sêmen de pacientes recuperados da COVID-19. Tanto a metodologia que envolveu medidas com parâmetros espermáticos quanto a das medidas de elipsometria, demonstraram ótima eficiência no processo de classificação avaliado pelas métricas elevadas de ML.

O trabalho aqui realizado foi cuidadosamente elaborado, de modo a garantir a maior confiabilidade possível nos resultados obtidos. Isso pode ser observado, por exemplo, ao considerar os diferentes algoritmos, técnicas de processamento de dados, processos de otimização, métricas, métodos de validação cruzada e a avaliação por interpretabilidade que foram empregados na análise de ML. Além disso, os métodos aplicados aqui vieram evoluindo ao longo dos anos com a experiência adquirida de trabalhos anteriores do grupo de pesquisa, o que tornou este estudo mais sólido, rigoroso e confiável em seus resultados.

A presente pesquisa possui principalmente duas limitações. Uma delas é a quantidade limitada de amostras disponíveis para o estudo. A outra é devido ao fato de não conhecermos a natureza biológica das alterações observadas nas amostras de sêmen medidas por elipsometria. Ainda assim, a abordagem MLS representa uma grande evolução na área da saúde reprodutiva. Os resultados apresentados demonstram a eficácia e aplicabilidade do método, cumprindo o objetivo proposto. A solução de diagnóstico apresentada aqui possui baixo custo, é rápida e não necessita de biomarcadores (*label-free*). Este estudo é o primeiro a apresentar uma metodologia com uso de inteligência artificial capaz de distinguir de forma eficaz entre amostras de sêmen de indivíduos que nunca foram diagnosticados com COVID-19 e daqueles indivíduos recuperados da doença. Dessa forma, produzimos um conhecimento valioso acerca dos problemas de infertilidade e sua relação com a doença, bem como sobre os impactos no sistema reprodutor masculino. Portanto, este campo de pesquisa, ainda pouco explorado, ganha, com esta pesquisa, um recurso promissor para novos estudos na área.

Um potencial ponto de avanço nessa área seria estudar amostras que estão infectadas pelo vírus. Focando a atenção em como a presença do vírus no sistema reprodutor masculino pode interferir e causar variações nas propriedades físicas das amostras. Nesse caso, o objetivo seria não apenas realizar o diagnóstico, mas também criar um modelo que descreva como as propriedades físicas são influenciadas pelo vírus. Para isso, seria necessário buscar ideias discutidas e demonstradas na literatura para solucionar o problema das reflexões incoerentes. Com essa questão resolvida, em vez de usar dados brutos de elipsometria, seria

possível usar os dados para modelar opticamente as amostras e obter a função dielétrica das mesmas. A partir disso, seria viável classificar as amostras por meio de uma propriedade física. O que tornaria possível avaliar como a infecção pelo vírus pode influenciar na função dielétrica do material biológico. Outro ponto de avanço importante que pode ser explorado está relacionado com o ótimo desempenho dos modelos de ML na análise dos dados de uR das medidas de elipsometria. Por se tratar de uma medida simples, um reflectômetro portátil poderia ser desenvolvido com o objetivo de medir apenas essa propriedade. Nesse caso, não seria mais necessário um elipsômetro, então o custo do equipamento de medida seria reduzido drasticamente. Além disso, a tarefa de medir as amostras se tornaria mais simples e rápida.

Referências

- [1] World Health Organization: *Coronavirus disease (COVID-19)*, 2025. https://www.who.int/health-topics/coronavirus#tab=tab_1, Acesso em: 22 mar. 2025. Citado na página 13.
- [2] Centers for Disease Control and Prevention: *Coronavirus Disease 2019 (COVID-19)*, 2025. Acesso em: 9 jul. 2025. Citado na página 13.
- [3] Amaral, P. H. R., L. M. Andrade, F. G. Fonseca e J. C. G. Perez: *Impact of COVID-19 in Minas Gerais, Brazil: Excess deaths, sub-notified cases, geographic and ethnic distribution*. Transboundary and Emerging Diseases, 68(4):2521–2530, 2021. <https://doi.org/10.1111/tbed.13922>. Citado na página 13.
- [4] Gupta, A., M. V. Madhavan, K. Sehgal *et al.*: *Extrapulmonary manifestations of COVID-19*. Nature Medicine, 26:1017–1032, 2020. <https://doi.org/10.1038/s41591-020-0968-3>. Citado na página 13.
- [5] Vasquez-Bonilla, W. O., R. Orozco, V. Argueta, M. Sierra, L. I. Zambrano, F. Muñoz-Lara e A. J. Rodríguez-Morales: *A review of the main histopathological findings in coronavirus disease 2019*. Human Pathology, 105:74–83, 2020. <https://doi.org/10.1016/j.humpath.2020.07.023>. Citado na página 13.
- [6] Peiris, S., H. Mesa, A. Aysola, J. Manivel, J. Toledo, M. Borges-Sa, S. Aldighieri e L. Reveiz: *Pathological findings in organs and tissues of patients with COVID-19: A systematic review*. PLoS ONE, 16(4):e0250708, 2021. <https://doi.org/10.1371/journal.pone.0250708>. Citado na página 13.
- [7] Li, H., X. Xiao, J. Zhang, M. I. Zafar, C. Wu, Y. Long e C. Xiong: *Impaired spermatogenesis in COVID-19 patients*. EClinicalMedicine, 28:100604, 2020. <https://doi.org/10.1016/j.eclinm.2020.100604>. Citado 3 vezes nas páginas 13, 16 e 17.
- [8] Yang, M., S. Chen, B. Huang, J. M. Zhong, H. Su, Y. J. Chen e X. Nie: *Pathological findings in the testes of COVID-19 patients: clinical implications*. European Urology Focus, 6(5):1124–1129, 2020. <https://doi.org/10.1016/j.euf.2020.05.009>. Citado 3 vezes nas páginas 13, 16 e 17.
- [9] Hamarat, M. B., M. S. Ozkent, B. Yilmaz, S. Y. Aksanyar e K. Karabacak: *Effect of SARS-CoV-2 infection on semen parameters*. Canadian Urological Association Journal, 16(3):E173, 2022. <http://dx.doi.org/10.5489/cuaj.7292>. Citado 3 vezes nas páginas 13, 17 e 53.

- [10] Irene, E. A.: *A Brief History and State of the Art of Ellipsometry*. Em *Ellipsometry at the Nanoscale*, capítulo 1, página 1. Springer, 2013. <http://dx.doi.org/10.1007/978-3-642-33956-1>. Citado na página 14.
- [11] Tompkins, H. G. e J. N. Hilfiker: *Spectroscopic Ellipsometry - Practical Application to Thin Film Characterization*. Momentum Press, 2016. Citado 5 vezes nas páginas 14, 18, 19, 21 e 22.
- [12] Jung, J., J. Bork, T. Holmgaard e N. A. Kortbek: *Ellipsometry*. Relatório Técnico, Aalborg University, Institute of Physics and Nanotechnology, 2004. Citado na página 14.
- [13] Brown, S.: *Machine Learning, Explained*. MIT Management Sloan School, 2021. <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>. Citado 4 vezes nas páginas 14, 34, 42 e 49.
- [14] Choudhury, A.: *The Role of Machine Learning Algorithms in Materials Science: A State of Art Review on Industry 4.0*. Archives of Computational Methods in Engineering, 28:3361–3381, 2021. <https://doi.org/10.1007/s11831-020-09503-4>. Citado na página 14.
- [15] Liu, J., D. Zhang, D. Yu *et al.*: *Machine learning powered ellipsometry*. Light: Science and Applications, 10(55), 2021. <https://doi.org/10.1038/s41377-021-00482-0>. Citado na página 14.
- [16] Lourenço, A. A., P. H. R. Amaral, A. A. O. Paim, G. F. Marques, L. Gomes-de Pontes, C. P. S. M. Mata, F. G. Fonseca, J. C. G. Pérez e J. G. A. Coelho-dos Reis: *Algorithms for predicting COVID outcome using ready-to-use laboratorial and clinical data*. Frontiers in Public Health, 12:1347334, 2024. <https://doi.org/10.3389/fpubh.2024.1347334>. Citado na página 14.
- [17] Ceccon, D. M., P. H. R. Amaral, L. M. Andrade, M. I. N. Silva, L. A. F. Andrade, T. F. S. Moraes e J. C. Gonzalez: *New, fast, and precise method of COVID-19 detection in nasopharyngeal and tracheal aspirate samples combining optical spectroscopy and machine learning*. Brazilian Journal of Microbiology, 54(2):769–777, 2023. <https://doi.org/10.1007/s42770-023-00923-5>. Citado 5 vezes nas páginas 14, 42, 43, 45 e 46.
- [18] Suleman, M., M. Ilyas, M. I. U. Lali, H. T. Rauf e S. Kadry: *A review of different deep learning techniques for sperm fertility prediction*. MATH, 8(7):16360–16416, 2023, ISSN 2473-6988. <https://doi.org/10.3934/math.2023838>. Citado na página 14.
- [19] Hanassab, S., A. Abbata, A. C. Yeung, M. Voliotis, K. Tsaneva-Atanasova, T. W. Kelsey e W. S. Dhillon: *The prospect of artificial intelligence to personalize*

- assisted reproductive technology*. npj Digital Medicine, 7(55):1–19, 2024. <https://doi.org/10.1038/s41746-024-01006-x>. Citado na página 14.
- [20] Cherouveim, P., C. Velmahos e C. L. Bormann: *Artificial intelligence for sperm selection—a systematic review*. Fertil. Steril., 120(1):24–31, 2023, ISSN 0015-0282. <https://doi.org/10.1016/j.fertnstert.2023.05.157>. Citado na página 14.
- [21] Wnuk, N. T., A. F. A. Figueiredo, T. O. Farias, M. R. G. Brener, S. M. S. N. Lacerda, V. N. Camargos e G. M. J. Costa: *Congenital Zika virus infection impacts on male mouse offspring’s reproductive biology*. Reproduction, 167(5):e230343, 2024. <https://doi.org/10.1530/REP-23-0343>. Citado 4 vezes nas páginas 14, 16, 45 e 46.
- [22] Amaral, P. H., N. T. Wnuk, V. N. Camargos, L. M. Andrade, M. da Silva, M. M. Teixeira, D. da Glória Souza, V. V. Costa, S. M. dos Santos Nassif Lacerda, G. M. J. Costa e J. González: *Machine learning ellipsometry as a sensitive diagnostic tool to study reproductive biology in zika virus infected murine models*. Microchemical Journal, 207:111973, 2024. <https://doi.org/10.1016/j.microc.2024.111973>. Citado 6 vezes nas páginas 14, 30, 42, 43, 45 e 61.
- [23] World Health Organization: *1 in 6 people globally affected by infertility: WHO*, 2025. <https://www.who.int/news/item/04-04-2023-1-in-6-people-globally-affected-by-infertility>, Acesso em: 27 mar. 2025. Citado na página 16.
- [24] Le Tortorec, A., G. Matusali, D. Mahé, F. Aubry, S. Mazaud-Guittot, L. Houzet e N. Dejucq-Rainsford: *From ancient to emerging infections: the odyssey of viruses in the male genital tract*. Physiological Reviews, 100(3):1349–1414, 2020. <https://doi.org/10.1152/physrev.00021.2019>. Citado na página 16.
- [25] Liu, W., R. Han, H. Wu e D. Han: *Viral threat to male fertility*. Andrologia, 50(11):e13140, 2018. <https://doi.org/10.1111/and.13140>. Citado na página 16.
- [26] Costa, G. M. J., S. M. S. N. Lacerda, A. F. A. Figueiredo, N. T. Wnuk, M. R. G. Brener, L. M. Andrade e M. H. Furtado: *High SARS-CoV-2 tropism and activation of immune cells in the testes of non-vaccinated deceased COVID-19 patients*. BMC Biology, 21(36):1–24, 2023. <https://doi.org/10.1186/s12915-022-01497-8>. Citado 2 vezes nas páginas 16 e 17.
- [27] Gonzalez, D. C., K. Khodamoradi, R. Pai, K. Guarch, Z. M. Connelly, E. Ibrahim, H. Arora e R. Ramasamy: *A systematic review on the investigation of SARS-CoV-2 in semen*. Research and Reports in Urology, 12:615–621, 2020. <https://doi.org/10.2147/RRU.S277679>. Citado na página 16.

- [28] Duarte-Neto, A. N., T. A. Teixeira, E. G. Caldini, C. T. Kanamura, M. S. Gomes-Gouvêa, A. B. Santos e J. Hallak: *Testicular pathology in fatal COVID-19: A descriptive autopsy study*. *Andrology*, 10(1):13–23, 2021. <https://doi.org/10.1111/andr.13073>. Citado 2 vezes nas páginas 16 e 17.
- [29] Wang, Z. e X. Xu: *scRNA-seq profiling of human testes reveals the presence of the ACE2 receptor, a target for SARS-CoV-2 infection in spermatogonia, Leydig and Sertoli cells*. *Cells*, 9(4):E920, 2020. <https://doi.org/10.3390/cells9040920>. Citado na página 16.
- [30] Paoli, D., F. Pallotti, A. Anzuini, S. Bianchini, L. Caponecchia, A. Carraro e F. Lombardo: *Male reproductive health after 3 months from SARS-CoV-2 infection: a multicentric study*. *Journal of Endocrinological Investigation*, 46(1):89–101, 2022. <https://doi.org/10.1007/s40618-022-01887-3>. Citado 3 vezes nas páginas 17, 49 e 53.
- [31] Eckstein, V., K. Glaß, M. E. Leßmann, J. Schaar, A. Klimova, P. Wimberger e M. Goeckenjan: *Assisted reproduction after SARS-CoV-2-infection: results of a single-center cohort-study*. *Archives of Gynecology and Obstetrics*, 309(1):305–313, 2024. <https://doi.org/10.1007/s00404-023-07228-w>. Citado na página 17.
- [32] Fujiwara, H.: *Spectroscopic Ellipsometry: Principles and Applications*. John Wiley and Sons Ltd, Chichester, West Sussex, England, 2007. Citado 14 vezes nas páginas 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31 e 42.
- [33] Griffiths, David J.: *Introduction to Electrodynamics*. Pearson, Boston, 4ª edição, 2013, ISBN 9780321856562. Citado na página 19.
- [34] J. A. Woollam: *Ellipsometry FAQ - J.A. Woollam*. 2018. <https://www.jawoollam.com/resources/ellipsometry-faq#toggle-id-14>, Acesso em: 15 abr. 2025. Citado na página 28.
- [35] Liu, P. Y., L. K. Chin, W. Ser, H. F. Chen, C. M. Hsieh, C. H. Lee, K. B. Sung, T. C. Ayi, P. H. Yap, B. Liedberg, K. Wang, T. Bourouina e Y. Leprince-Wang: *Cell refractive index for cell biology and disease diagnosis: past, present and future*. *Lab Chip*, 16(4):634–644, 2016, ISSN 1473-0197. <https://doi.org/10.1039/C5LC01445J>. Citado na página 30.
- [36] Möckel, C., T. Beck, S. Kaliman, S. Abuhattum, K. Kim, J. Kolb, D. Wehner, V. Zaburdaev e J. Guck: *Estimation of the mass density of biological matter from refractive index measurements*. *Biophysical Reports*, 4(2):100156, 2024, ISSN 2667-0747. <https://doi.org/10.1016/j.bpr.2024.100156>. Citado 2 vezes nas páginas 30 e 61.

- [37] Zhang, Q., L. Zhong, P. Tang, Y. Yuan e X. Lu: *Quantitative refractive index distribution of single cell by combining phase-shifting interferometry and AFM imaging*. Sci. Rep., 7(1), 2017, ISSN 2045-2322. <https://doi.org/10.1038/s41598-017-02797-8>. Citado 2 vezes nas páginas 30 e 61.
- [38] Lai, J. C., Y. Y. Zhang, Z. H. Li, H. J. Jiang e A. Z. He: *Complex refractive index measurement of biological tissues by attenuated total reflection ellipsometry*. Appl. Opt., 49(16):3235–3238, 2010, ISSN 2155-3165. <https://doi.org/10.1364/AO.49.003235>. Citado 2 vezes nas páginas 30 e 61.
- [39] Ossikovski, R., M. Kildemo, M. Stchakovsky e M. Mooney: *Anisotropic incoherent reflection model for spectroscopic ellipsometry of a thick semitransparent anisotropic substrate*. Appl. Opt., 39(13):2071–2077, 2000, ISSN 2155-3165. <https://doi.org/10.1364/AO.39.002071>. Citado na página 31.
- [40] Joerger, R., K. Forcht, A. Gombert, M. Köhl e W. Graf: *Influence of incoherent superposition of light on ellipsometric coefficients*. Appl. Opt., 36(1):319–327, 1997, ISSN 1559-128X. <https://doi.org/10.1364/ao.36.000319>. Citado na página 31.
- [41] Peng, L., J. Wang, J. Wang, F. Gao, J. Zhang, W. Zhai, L. Zhou, L. Zhou e X. Jiang: *Incoherent partial superposition modeling for single-shot angle-resolved ellipsometry measurement of thin films on transparent substrates*. Opt. Express, 32(9):15774–15787, 2024, ISSN 1094-4087. <https://doi.org/10.1364/OE.517216>. Citado 3 vezes nas páginas 31, 32 e 33.
- [42] Chen, J., J. Li, D. Sainju, K. D. Wells, N. J. Podraza e R. W. Collins: *Multilayer Analysis of the CdTe Solar Cell Structure by Spectroscopic Ellipsometry*. Em *2006 IEEE 4th World Conference on Photovoltaic Energy Conference*, páginas 07–12. IEEE, 2006. <https://doi.org/10.1109/WCPEC.2006.279494>. Citado na página 32.
- [43] Chen, J., J. Li, C. Thornberry, M. N. Sestak, R. W. Collins e J. D. Walker: *Through-the-glass spectroscopic ellipsometry of CdTe solar cells*. Em *2009 34th IEEE Photovoltaic Specialists Conference (PVSC)*, páginas 07–12. IEEE, 2009. <https://doi.org/10.1109/PVSC.2009.5411452>. Citado na página 32.
- [44] Koirala, P., J. Li, H. P. Yoon, P. A., S. Marsillac, A. A. Rockett, N. J. Podraza e R. W. Collins: *Through-the-glass spectroscopic ellipsometry for analysis of CdTe thin-film solar cells in the superstrate configuration*. Prog. Photovoltaics Res. Appl., 24(8):1055–1067, 2016, ISSN 1062-7995. <https://doi.org/10.1002/pip.2759>. Citado na página 32.

- [45] Hilfiker, J. N., C. M. Herzinger, T. Wagner, Antigone Marino, Guiseppe Delgais e Giancarlo Abbate: *Mueller-matrix characterization of liquid crystals*. Thin Solid Films, 455-456:591–595, 2004, ISSN 0040-6090. <https://doi.org/10.1016/j.tsf.2004.02.011>. Citado na página 32.
- [46] Hayton, D. J. e T. E. Jenkins: *On the frustration of back-surface reflection from transparent substrates in ellipsometry*. Meas. Sci. Technol., 15(2):N17, 2003, ISSN 0957-0233. Citado na página 32.
- [47] Synowicki, R. A.: *Suppression of backside reflections from transparent substrates*. Phys. Status Solidi C, 5(5):1085–1088, 2008, ISSN 1862-6351. Citado 2 vezes nas páginas 32 e 33.
- [48] Synowicki, R. A.: *Transparent Substrates: How To Suppress Reflections from the Back Surface*. J.A. Woollam Newsletter Issue 11, páginas 10–11, 2010. Citado 2 vezes nas páginas 32 e 33.
- [49] Brownlee, J.: *How Machine Learning Algorithms Work (they learn a mapping of input to output)*. MachineLearningMastery, 2019. <https://machinelearningmastery.com/how-machine-learning-algorithms-work>, Acesso em: 03 abr. 2025. Citado na página 34.
- [50] IBM: *What is machine learning?* 2025. <https://www.ibm.com/think/topics/machine-learning>, Acesso em: 03 abr. 2025. Citado 2 vezes nas páginas 34 e 37.
- [51] IBM: *What is a machine learning algorithm?* 2024. <https://www.ibm.com/think/topics/machine-learning-algorithms>, Acesso em: 04 abr. 2025. Citado na página 34.
- [52] Breiman, L., J. Friedman, R. Olshen e C. Stone: *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984. Citado 2 vezes nas páginas 34 e 45.
- [53] Chen, T. e C. Guestrin: *XGBoost: A Scalable Tree Boosting System*. Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, páginas 785–794, 2016. <https://doi.org/10.1145/2939672.2939785>. Citado 2 vezes nas páginas 34 e 45.
- [54] Fix, E. e J. L. Hodges: *Discriminatory analysis. Nonparametric discrimination: Consistency properties*. International Statistical Review/Revue Internationale de Statistique, 57(3):238–247, 1989. Citado 2 vezes nas páginas 34 e 45.
- [55] Cox, D. R.: *The regression analysis of binary sequences*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2):215–232, 1958. Citado 2 vezes nas páginas 34 e 45.

- [56] Haykin, S.: *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994. Citado 2 vezes nas páginas 34 e 45.
- [57] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*: *Scikit-learn: Machine Learning in Python*. The Journal of Machine Learning Research, 12:2825–2830, 2011. Citado 2 vezes nas páginas 34 e 45.
- [58] Sarker, I. H.: *Machine Learning: Algorithms, Real-World Applications and Research Directions*. SN Comput. Sci., 2(3):1–21, 2021, ISSN 2661-8907. <https://doi.org/10.1007/s42979-021-00592-x>. Citado na página 35.
- [59] Brownlee, J.: *How to Manually Optimize Machine Learning Model Hyperparameters*. MachineLearningMastery, 2021. <https://machinelearningmastery.com/manually-optimize-hyperparameters>, Acesso em: 04 abr. 2025. Citado na página 35.
- [60] Brownlee, J.: *What is the Difference Between a Parameter and a Hyperparameter?* MachineLearningMastery, 2019. <https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter>, Acesso em: 04 abr. 2025. Citado na página 35.
- [61] IBM: *What is hyperparameter tuning?* 2025. <https://www.ibm.com/think/topics/hyperparameter-tuning>, Acesso em: 04 abr. 2025. Citado na página 35.
- [62] Akiba, T., S. Sano, T. Yanase, T. Ohta e M. Koyama: *Optuna: A Next-generation Hyperparameter Optimization Framework*. Em *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, páginas 2623–2631. ACM, 2019. <https://doi.org/10.1145/3292500.3330701>. Citado 2 vezes nas páginas 35 e 45.
- [63] Brownlee, J.: *Data Preparation for Machine Learning*. MachineLearningMastery, 2023. <https://machinelearningmastery.com/data-preparation-for-machine-learning>, Acesso em: 03 abr. 2025. Citado na página 36.
- [64] Scikit-learn: *RobustScaler*. 2025. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>, Acesso em: 03 abr. 2025. Citado na página 36.
- [65] Brownlee, J.: *SMOTE for Imbalanced Classification with Python*. MachineLearningMastery, 2021. <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification>, Acesso em: 03 abr. 2025. Citado na página 36.

- [66] Scikit-learn: *1.13. Feature selection*. 2025. https://scikit-learn.org/stable/modules/feature_selection.html#select-from-model, Acesso em: 03 abr. 2025. Citado na página 36.
- [67] Scikit-learn: *Outlier detection with Local Outlier Factor (LOF)*. 2025. https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html, Acesso em: 03 abr. 2025. Citado na página 36.
- [68] Scikit-learn: *IterativeImputer*. 2025. <https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html>, Acesso em: 03 abr. 2025. Citado na página 36.
- [69] Rainio, O., J. Teuvo e R. Klén: *Evaluation Metrics and Statistical Tests for Machine Learning*. Scientific Reports, 14:6086, 2024. <https://doi.org/10.1038/s41598-024-56706-x>. Citado na página 37.
- [70] IBM: *Evaluation metrics*. 2025. <https://www.ibm.com/docs/en/watsonx/saas?topic=models-evaluation-metrics>, Acesso em: 06 abr. 2025. Citado na página 37.
- [71] Scikit-learn: *accuracy_score*. 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score, Acesso em: 04 abr. 2025. Citado na página 37.
- [72] Scikit-learn: *classification_report*. 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html, Acesso em: 04 abr. 2025. Citado na página 37.
- [73] Scikit-learn: *f1_score*. 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score, Acesso em: 04 abr. 2025. Citado na página 37.
- [74] Scikit-learn: *matthews_corrcoef*. 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html#sklearn.metrics.matthews_corrcoef, Acesso em: 04 abr. 2025. Citado na página 37.
- [75] Scikit-learn: *roc_auc_score*. 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html, Acesso em: 04 abr. 2025. Citado na página 37.
- [76] Sontag, D. e P. Szolovits: *Machine Learning for Healthcare*. MIT OpenCourseWare, 2025. <https://ocw.mit.edu/courses/6-s897-machine-learning-for-healthcare-spring-2019/resources/mit6-s897s19 lec5>, Acesso em: 06 abr. 2025. Citado na página 37.

- [77] Goodfellow, I., Y. Bengio e A. Courville: *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>. Citado na página 38.
- [78] Stone, M.: *Cross-validatory choice and assessment of statistical predictions*. Journal of the Royal Statistical Society: Series B (Methodological), 36(2):111–147, 1974. Citado na página 38.
- [79] Varma, S. e R. Simon: *Bias in error estimation when using cross-validation for model selection*. BMC Bioinformatics, 7(1):91, 2006. Citado na página 38.
- [80] Brownlee, J.: *Nested Cross-Validation for Machine Learning with Python*. MachineLearningMastery, 2021. <https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/>, Acesso em: 07 abr. 2025. Citado na página 40.
- [81] Lundberg, S. M. e S. I. Lee: *A unified approach to interpreting model predictions*. Em *Advances in Neural Information Processing Systems (NeurIPS)*, páginas 4768–4777. Curran Associates, Inc., 2017. <https://dl.acm.org/doi/10.5555/3295222.3295230>. Citado 2 vezes nas páginas 40 e 46.
- [82] Paiva, V. H. S. de: *Aprendizado de máquina aplicado à elipsometria espectroscópica para o diagnóstico sorológico de malária*. Tese de Mestrado, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil, 2025. Orientador: Juan Carlos González Pérez. Citado 4 vezes nas páginas 42, 43, 45 e 46.
- [83] Amaral, P. H. R., L. M. Andrade, M. I. N. da Silva e J. C. González: *Label-free spectroscopical differentiation between tumoral and non-tumoral epithelial cell lines through machine learning*. 2025. Artigo submetido. Citado 2 vezes nas páginas 42 e 43.
- [84] Hano, H., C. H. Lawrie, B. Suarez, A. P. Lario, I. E. Echeverría, J. G. Mediavilla, M. I. C. Cruz, E. Lopez e A. Seifert: *Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer*. ACS Omega, 9(12):14084–14091, 2024. <https://doi.org/10.1021/acsomega.3c09537>. Citado 2 vezes nas páginas 42 e 43.
- [85] Villazon, J., N. D. Cruz e L. Shi: *Cancer Cell Line Classification Using Raman Spectroscopy of Cancer-Derived Exosomes and Machine Learning*. Anal. Chem., 97(13):7289–7298, abril 2025, ISSN 0003-2700. <https://doi.org/10.1021/acs.analchem.4c06966>. Citado 2 vezes nas páginas 42 e 43.
- [86] Nascimento, M. H. C., W. D. Marcarini, G. S. Folli, W. G. da Silva Filho, L. L. Barbosa, E. H. Paulo, P. F. Vassallo, J. G. Mill, V. G. Barauna, F. L. Martin, E. V.

- R. de Castro, W. Romão e P. R. Filgueiras: *Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis*. Anal. Chem., 94(5):2425–2433, 2022, ISSN 0003-2700. <https://doi.org/10.1021/acs.analchem.1c04162>. Citado 2 vezes nas páginas 42 e 43.
- [87] World Health Organization: *WHO Laboratory Manual for the Examination and Processing of Human Semen*. World Health Organization, Geneva, sixth edition edição, 2021. <https://www.who.int/publications/i/item/9789240030787>. Citado 2 vezes nas páginas 45 e 49.
- [88] Unal, I.: *Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach*. Comput. Math. Methods Med., 2017:3762651.(;):Comput, 2017, ISSN 1748-6718. <https://doi.org/10.1155/2017/3762651>. Citado na página 46.
- [89] Shapiro, S. S. e M. B. Wilk: *An analysis of variance test for normality (complete samples)*. Biometrika, 52(3/4):591–611, 1965. Citado na página 48.
- [90] Student: *The probable error of a mean*. Biometrika, 6(1):1–25, 1908. Citado na página 48.
- [91] Mann, H. B. e D. R. Whitney: *On a test of whether one of two random variables is stochastically larger than the other*. The Annals of Mathematical Statistics, 18(1):50–60, 1947. Citado na página 48.
- [92] Guo, T. H., M. Y. Sang, S. Bai, H. Ma, Y. Y. Wan, X. H. Jiang, Y. W. Zhang, B. Xu, H. Chen, X. Y. Zheng, S. H. Luo, X. F. Xie, C. J. Gong, J. P. Weng e Q. H. Shi: *Semen parameters in men recovered from COVID-19*. Asian J. Androl., 23(5):479, 2021, ISSN 1008-682X. https://doi.org/10.4103/aja.aja_31_21. Citado 2 vezes nas páginas 49 e 53.
- [93] Donders, G. G. G., E. Bosmans, J. Reumers, F. Donders, J. Jonckheere, G. Salembier, N. Stern, Y. Jacquemyn, W. Ombelet e C. E. Depuydt: *Sperm quality and absence of SARS-CoV-2 RNA in semen after COVID-19 infection: a prospective, observational study and validation of the SpermCOVID test*. Fertil. Steril., 117(2):287–296, 2022, ISSN 0015-0282. <https://doi.org/10.1016/j.fertnstert.2021.10.022>. Citado 2 vezes nas páginas 49 e 53.
- [94] Gacci, M., M. Coppi, E. Baldi, A. Sebastianelli, C. Zaccaro, S. Morselli, A. Pecoraro, A. Manera, R. Nicoletti, A. Liaci, C. Bisegna, L. Gemma, S. Giancane, S. Pollini, A. Antonelli, F. Lagi, S. Marchiani, S. Dabizzi, S. Degl’Innocenti, F. Annunziato, M. Maggi, L. Vignozzi, A. Bartoloni, G. M. Rossolini e S. Serni:

- Semen impairment and occurrence of SARS-CoV-2 virus in semen after recovery from COVID-19.* Hum. Reprod., 36(6):1520–1529, 2021, ISSN 0268-1161. <https://doi.org/10.1093/humrep/deab026>. Citado na página 53.
- [95] Koç, E. e B. B. Keseroğlu: *Does COVID-19 Worsen the Semen Parameters? Early Results of a Tertiary Healthcare Center.* Urol. Int., 105(9-10):743–748, 2021, ISSN 0042-1138. <https://doi.org/10.1159/000517276>. Citado na página 53.
- [96] Aksak, T., D. A. Satar, R. Bağci, E. O. Gultekin, A. Coşkun e U. Demirdelen: *Investigation of the effect of COVID-19 on sperm count, motility, and morphology.* J. Med. Virol., 94(11):5201–5205, 2022, ISSN 0146-6615. <https://doi.org/10.1002/jmv.27971>. Citado na página 53.
- [97] Best, J. C., M. Kuchakulla, K. Khodamoradi, T. F. N. Lima, F. S. Frech, J. Achua, O. Rosete, B. Mora, H. Arora, E. Ibrahim e R. Ramasamy: *Evaluation of SARS-CoV-2 in Human Semen and Effect on Total Sperm Number: A Prospective Observational Study.* World Journal of Men's Health, 39(3):489–495, 2021, ISSN 2287-4208. <https://doi.org/10.5534/wjmh.200192>. Citado na página 53.
- [98] Cakir, C., G. Kuspinar, G. Kurt, M. Berber, K. Aslan, I. Kasapoglu, G. Uncu e B. Avci: *Comparison of semen parameters in the same patients before and after diagnosis of COVID-19.* J. Med. Virol., 95(9):e29094, 2023, ISSN 0146-6615. <https://doi.org/10.1002/jmv.29094>. Citado na página 53.
- [99] Erbay, G., A. Sanli, H. Turel, U. Yavuz, A. Erdogan, M. Karabakan, M. Yaris e M. H. Gultekin: *Short-term effects of COVID-19 on semen parameters: A multicenter study of 69 cases.* Andrology, 9(4):1060–1065, 2021, ISSN 2047-2919. <https://doi.org/10.1111/andr.13019>. Citado na página 53.
- [100] Temiz, M. Z., M. M. Dincer, I. Hacibey, R. O. Yazar, C. Celik, S. H. Kucuk, G. Alkurt, L. Doganay, E. Yuruk e A. Y. Muslumanoglu: *Investigation of SARS-CoV-2 in semen samples and the effects of COVID-19 on male sexual health by using semen analysis and serum male hormone profile: A cross-sectional, pilot study.* Andrologia, 53(2):e13912, 2021, ISSN 0303-4569. <https://doi.org/10.1111/and.13912>. Citado na página 53.
- [101] Scropo, F. I., E. Costantini, A. Zucchi, E. Illiano, F. Trama, S. Brancorsini, F. Crocetto, M. R. Gismondo, F. Dehò, A. Mercuriali, R. Bartoletti e F. Gaeta: *COVID-19 disease in clinical setting: impact on gonadal function, transmission risk, and sperm quality in young males.* J. Basic Clin. Physiol. Pharmacol., 33(1):97–102, 2022, ISSN 2191-0286. <https://doi.org/10.1515/jbcpp-2021-0227>. Citado na página 53.

- [102] Harvey, A. H.: *Index of Refraction of Liquid Water*. NIST, 2024. <https://www.nist.gov/publications/index-refraction-liquid-water>. Citado na página 57.

Apêndices

APÊNDICE A – Tabelas de espaços de busca para os parâmetros e hiperparâmetros

Tabela 8 – Parâmetros e espaços de busca para os modelos de aprendizado de máquina.

Parâmetros	Espaço de busca
Ângulo	45, 50, 55, 60, 65, 70
Comprimento de onda mínimo (wl min)	245 - 1660
Comprimento de onda máximo (wl max)	(wl min + 20) - 1700
Número de features	1 - (comprimentos de onda entre wl max e wl min)

Tabela 9 – Hiperparâmetros e espaços de busca para os modelos de aprendizado de máquina.

Modelos	Hiperparâmetros	Espaço de busca
KNN	n neighbors weights p	2 e 3 uniform, distance 1, 2 e 3
LR	C penalty class weight [†] solver max iter	0,01 - 10,00 l1, l2 0,05 - 0,95 saga 100000
SVM	C gamma class weight [†]	0,01 - 100000 0,0001 - 0,0100 0,1 - 0,9
XGBoost	n estimators learning rate max depth subsample colsample bytree colsample bylevel colsample bynode scale pos weight	1 - 1000 0,001 - 0,900 1 - 10 0,1 - 1,0 0,1 - 1,0 0,1 - 1,0 0,1 - 1,0 0,1 - 2,0
MLP	hln1* hln2** solver activation momentum max iter	1 até 2× número de features 1 até 2× número de features sgd, adam logistic, tanh, relu 0,01 - 1,00 100000
DT	criterion max depth max features	entropy, log loss, gini 1 - 50 auto, sqrt, log2, None

[†] O valor aqui se refere à classe 0. Para a classe 1, o valor é o complemento: 1 - class weight.

* Número de neurônios para a primeira camada oculta.

** Número de neurônios para a segunda camada oculta.

APÊNDICE B – Tabelas de parâmetros e hiperparâmetros otimizados

Tabela 10 – Parâmetros otimizados para cada um dos modelos.

Modelo	Parâmetros	Ψ	Δ	dPol	uR
KNN	ângulo	45	65	60	70
	faixa espectral (nm)	270 - 654	338 - 414	1438 - 1525	1255 - 1587
	número de features	164	27	13	27
LR	ângulo	45	45	50	60
	faixa espectral (nm)	265 - 361	701 - 1408	1166 - 1577	567 - 1660
	número de features	55	281	98	368
SVM	ângulo	55	65	50	50
	faixa espectral (nm)	506 - 1138	1131 - 1277	1026 - 1360	957 - 1666
	número de features	9	24	8	168
XGBoost	ângulo	60	60	50	45
	faixa espectral (nm)	300 - 455	334 - 408	1059 - 1321	834 - 1674
	número de features	77	23	25	222
MLP	ângulo	45	65	50	50
	faixa espectral (nm)	268 - 344	936 - 1321	1161 - 1388	1510 - 1692
	número de features	44	92	29	15
DT	ângulo	45	55	50	55
	faixa espectral (nm)	328 - 429	257 - 580	1307 - 1466	867 - 1683
	número de features	26	90	25	267

Tabela 11 – Hiperparâmetros otimizados para cada um dos modelos.

Modelo	Hiperparâmetros	Ψ	Δ	dPol	uR
KNN	n neighbors	2	2	2	2
	weights	uniform	distance	distance	distance
	p	3	2	1	2
LR	C	2,27	6,05	8,98	7,31
	penalty	l2	l1	l1	l1
	class weight	0: 0,85, 1: 0,15	0: 0,57, 1: 0,43	0: 0,60, 1: 0,40	0: 0,37, 1: 0,63
	solver	saga	saga	saga	saga
	max iter	100000	10000	10000	10000
SVM	C	3561	92885	1567	64787
	gamma	0,00063	0,0081	0,0092	0,0095
	weight0	0,49	0,37	0,87	0,17
XGBoost	n estimators	999	545	683	477
	learning rate	0,079	0,34	0,89	0,53
	max depth	9	9	3	2
	subsample	0,94	0,53	1,00	0,58
	colsample bytree	0,66	0,73	0,56	0,88
	colsample bylevel	0,97	0,98	0,37	0,96
	colsample bynode	0,63	0,35	0,55	0,36
	scale pos weight	1,13	0,11	0,34	1,01
MLP	hln1	31	136	35	27
	hln2	37	112	35	14
	solver	sgd	sgd	adam	adam
	activation	tanh	tanh	tanh	relu
	momentum	0,87	0,99	0,57	0,58
	max iter	100000	100000	100000	100000
DT	criterion	log loss	gini	gini	log loss
	max depth	36	8	16	48
	max features	None	sqrt	None	auto