



Universidade Estadual de Feira de Santana  
Programa de Pós-Graduação em Ciência da Computação

# Classificação de Amiloidose em Imagens Digitais de Biópsias Renais Utilizando Corantes não Específicos

Gledson de Oliveira

Feira de Santana

2023



Universidade Estadual de Feira de Santana  
Programa de Pós-Graduação em Ciência da Computação

Gledson de Oliveira

## **Classificação de Amiloidose em Imagens Digitais de Biópsias Renais Utilizando Corantes não Específicos**

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Angelo Amâncio Duarte

Coorientador: Washington Luis Conrado dos Santos

Feira de Santana

2023

Ficha Catalográfica - Biblioteca Central Julieta Carteado - UEFS

O47c

Oliveira, Gledson de

Classificação de amiloidose em imagens digitais em biópsias renais utilizando corantes não específicos / Gledson de Oliveira.– 2023.

74 f.: il.

Orientador: Angelo Amâncio Duarte.

Coorientador: Washington Luis Conrado dos Santos.

Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-graduação em Ciência da Computação, Feira de Santana, 2023.

1. Patologia computacional. 2. Rede neural. 3. Amiloidose renal. I. Título. II. Duarte, Angelo Amâncio, orient. III. Santos, Washington Luis Conrado dos, coorient. IV. Universidade Estadual de Feira de Santana.

CDU 616:004.32.26

Gledson de Oliveira


## **Classificação de amiloidose em imagens digitais de biópsias renais utilizando corantes não específicos**

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Feira de Santana, 18 de julho de 2023

### **BANCA EXAMINADORA**


---

Documento assinado digitalmente  
 ANGELO AMANCIO DUARTE  
Data: 05/09/2023 08:13:46-0300  
Verifique em <https://validar.iti.gov.br>

---

Angelo Amâncio Duarte (Orientador)  
Universidade Estadual de Feira de Santana


---

Documento assinado digitalmente  
 RODRIGO DE MELO SOUZA VERAS  
Data: 19/07/2023 08:46:27-0300  
Verifique em <https://validar.iti.gov.br>

---

Rodrigo de Melo Souza Veras  
Universidade Federal do Piauí

---

Documento assinado digitalmente  
 MATHEUS GIOVANNI PIRES  
Data: 20/07/2023 14:24:25-0300  
Verifique em <https://validar.iti.gov.br>

---

Matheus Giovanni Pires  
Universidade Estadual de Feira de Santana



# Abstract

Computational Pathology is a field of study that utilizes computational methods to assist medical pathologists in the analysis of pathological images, employing machine learning algorithms to contribute to faster and more precise diagnoses. Nevertheless, there is still much to be researched, as exemplified by the case of amyloidosis, a rare condition that poses a challenge to the development of efficient classifiers for the condition due to a limited amount of available images for training automatic classifiers. Additionally, amyloidosis presents an additional complication arising from the necessity of using specific dyes for lesion detection by physicians, further reducing the pool of available images. Based on this research problem, this study employed an approach utilizing classical convolutional neural network models to construct an automatic amyloidosis classifier, trained using colored images with non-specific dyes for the lesion. Initially, these models were trained using an imbalanced dataset with fewer images for amyloidosis to establish a research baseline. Subsequently, data balancing techniques such as Random Undersampling and Random Oversampling, along with Ensemble-Based algorithms, were applied to address class imbalance. As a result of this work, models capable of identifying the lesion with a false negative rate of up to 4.5% were obtained, with better performance observed for the Inception model when trained with the RUS dataset and for the Ensemble-RUS model.

**Keywords:** Computational Pathology, Amyloidosis, Class Imbalance

# Resumo

A Patologia Computacional é um campo de estudo que utiliza métodos computacionais para auxiliar médicos patologistas na análise de imagens patológicas, fazendo uso de algoritmos de aprendizado de máquina que contribuem para diagnósticos mais rápidos e precisos. No entanto, ainda existe muito a ser pesquisado, como por exemplo no caso da amiloidose, uma lesão pouco frequente que impõe um desafio à construção de classificadores eficientes para a lesão, devido a uma baixa quantidade de imagens disponíveis para treinamento de classificadores automáticos. Além disso, a amiloidose também apresenta uma complicação adicional oriunda da necessidade de uso de corantes específicos para a detecção da lesão pelos médicos, o que reduz ainda mais o número de imagens disponíveis. Com base nesse problema de pesquisa, esse trabalho aplicou uma abordagem utilizando modelos de redes neurais convolucionais clássicas para a construção de um classificador automático de amiloidose, treinado utilizando imagens coloridas com corantes não específicos para a lesão. Inicialmente, tais modelos foram treinados utilizando a base de dados desbalanceada com menos imagens para a amiloidose, com o objetivo de estabelecer uma referência de pesquisa. Em seguida, foram aplicadas técnicas de balanceamento de dados, como *Random Undersample* e *Random Oversample*, e algoritmos do tipo *Ensemble-Based*, com o objetivo de lidar com o desbalanceamento entre as classes. Como resultados deste trabalho foram obtidos modelos capazes de identificar a lesão com uma taxa de falsos negativos de até 4,5%, com um melhor desempenho para o modelo Inception, quando treinado com o Dataset RUS e para o modelo o Ensemble-RUS.

**Palavras-chave:** Patologia Computacional, Amiloidose, Desbalanceamento de Classe.

# Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

A dissertação foi desenvolvida no Programa de Pós-Graduação em Ciência da Computação (PGCC), tendo como orientador o Prof. Dr. **Angelo Amâncio Duarte**. O Prof. Dr. **Washington Luis Conrado dos Santos** foi coorientador(a) deste trabalho.

Esta pesquisa foi financiada pela FAPESB (ou CAPES). *Obrigatório colocar esse texto, caso o(a) estudante tenha recebido bolsa da FAPESB (ou da CAPES).*

# Agradecimentos

Agradeço imensamente à minha família, especialmente à minha Mãe, meu Pai e minha irmã, por todo o encorajamento e suporte ao longo da minha vida. Expresso meu agradecimento às minhas madrinhas, Silvanete e Paulina, e ao meu padrinho, Antônio Simeão, por estarem sempre presentes nesta jornada. Agradeço também a todos os meus amigos pelo apoio constante e pelos momentos de descontração e relaxamento tão valiosos.

Quero manifestar meu profundo agradecimento ao Prof. Dr. Angelo Amâncio Duarte e ao Prof. Dr. Washington L. C. dos Santos pelas orientações excepcionais que me proporcionaram ao longo do meu percurso de mestrado. Essas orientações foram fundamentais para o meu desenvolvimento acadêmico e profissional.

Estendo meus agradecimentos aos professores e colegas do Programa de Pós-Graduação em Ciência da Computação (PGCC) da Universidade Estadual de Feira de Santana (UEFS), bem como aos colegas do Laboratório de Computação de Alto Desempenho (LaCAD). A presença e o apoio de todos vocês foram inestimáveis em minha jornada acadêmica.

# Sumário

<b>Abstract</b>	<b>i</b>
<b>Resumo</b>	<b>ii</b>
<b>Prefácio</b>	<b>iii</b>
<b>Agradecimentos</b>	<b>iv</b>
<b>Sumário</b>	<b>vi</b>
<b>Lista de Tabelas</b>	<b>viii</b>
<b>Lista de Figuras</b>	<b>ix</b>
<b>Lista de Abreviações</b>	<b>x</b>
<b>1 Introdução</b>	<b>1</b>
1.0.1 Justificativa . . . . .	3
1.0.2 Objetivos . . . . .	3
<b>2 Referencial Teórico</b>	<b>5</b>
2.0.1 Conceitos de Nefropatologia . . . . .	5
2.0.2 Conceitos de Computação . . . . .	8
<b>3 Revisão Bibliográfica</b>	<b>16</b>
<b>4 Metodologia</b>	<b>20</b>
4.0.1 Ferramentas Utilizadas . . . . .	20
4.0.2 <i>Dataset</i> . . . . .	20
4.0.3 Modelos Utilizados . . . . .	25
4.0.4 Estratégias para lidar com o Desbalanceamento de Classes . .	25
4.0.5 Avaliação de Desempenho . . . . .	26
4.0.6 Treinamento, Validação e Teste . . . . .	27

<b>5</b>	<b>Resultados</b>	<b>29</b>
5.0.1	Treinamento dos Modelos utilizando o <i>Dataset Baseline</i> . . . . .	29
5.0.2	Testes dos Modelos utilizando o <i>Dataset Baseline</i> . . . . .	31
5.0.3	Avaliação dos Modelos após Realização de Balanceamento de Dados por meio de <i>Random Undersampling</i> . . . . .	36
5.0.4	Testes dos Modelos após Realização de Balanceamento de Dados por meio de <i>Random Undersampling</i> . . . . .	36
5.0.5	Avaliação dos Modelos após Realização de Balanceamento de Dados por meio de <i>Random Oversampling</i> . . . . .	42
5.0.6	Testes dos Modelos após Realização de Balanceamento de Dados por meio de <i>Random Oversampling</i> . . . . .	42
5.0.7	Resultados Modelos <i>Ensemble-Based</i> . . . . .	48
<b>6</b>	<b>Discussão</b>	<b>52</b>
<b>7</b>	<b>Conclusões</b>	<b>55</b>
	<b>Referências</b>	<b>57</b>

# Lista de Tabelas

4.1	Relação da Quantidade de Imagens por Corante para o Dataset RUS.	24
4.2	Quantidade de Imagens por Lesão e Corante presentes no <i>Dataset Baseline</i> .	24
4.3	Relação da Quantidade de Imagens por Corante para o Dataset RUS.	26
4.4	Relação da Quantidade de Imagens por Corante para o Dataset ROS.	26
4.5	Quantidade de Imagens por Lesão e Corante presentes na base de dados de teste.	28
5.1	Erros de Classificação por para os Melhores Modelos Treinados com o <i>Dataset Baseline</i> .	33
5.2	Erros de classificação para o modelo VGG16	33
5.3	Erros de classificação para o modelo VGG19.	33
5.4	Erros de classificação para o modelo <i>Xception</i> .	33
5.5	Erros de classificação para o modelo <i>Inception</i> .	33
5.6	Erros de classificação para o modelo <i>Inception-Resnet</i> .	33
5.7	Média das Métricas de desempenho para o <i>Dataset Baseline</i> .	34
5.8	Intervalo de Confiança utilizando o <i>Dataset Baseline</i> .	34
5.9	Erros de Classificação por para os Melhores Modelos Treinados com o <i>Dataset RUS</i> .	40
5.10	Erros de classificação para o modelo VGG16.	40
5.11	Erros de classificação para o modelo VGG19.	40
5.12	Erros de classificação para o modelo <i>Xception</i> .	40
5.13	Erros de classificação para o modelo <i>Inception</i> .	40
5.14	Erros de classificação para o modelo <i>Inception-Resnet</i> .	40
5.15	Média das Métricas de desempenho Utilizando o <i>Dataset RUS</i> .	41
5.16	Intervalo de confiança utilizando o <i>Dataset RUS</i> .	41
5.17	Erros de Classificação por para os Melhores Modelos Treinados com o <i>Dataset ROS</i> .	46
5.18	Erros de classificação para o modelo VGG16.	46
5.19	Erros de classificação para o modelo VGG19.	46
5.20	Erros de classificação para o modelo <i>Xception</i> .	46
5.21	Erros de classificação para o modelo <i>Inception</i> .	46
5.22	Erros de classificação para o modelo <i>Inception-Resnet</i> .	46
5.23	Média das Métricas de desempenho Utilizando o <i>dataset ROS</i> .	47

5.24	Intervalo de confiança utilizando o <i>Dataset ROS</i> . . . . .	47
5.25	Erros de classificação por para o melhor modelos <i>Ensemble</i> . . . . .	50
5.26	Erros de classificação para o modelo <i>Ensemble-Baseline</i> . . . . .	50
5.27	Erros de classificação para o modelo <i>Ensemble-RUS</i> . . . . .	50
5.28	Erros de classificação para o modelo <i>Ensemble-ROS</i> . . . . .	50
5.29	Média das Métricas de desempenho para o modelo <i>Ensemble-Based</i> . . . . .	51
5.30	Intervalo de confiança para o modelo <i>Ensemble-Based</i> . . . . .	51



# Lista de Figuras

2.1	Exemplo de uma imagem histológica de um glomérulo. . . . .	6
2.2	Exemplo de um glomérulo com e sem Amiloidose. . . . .	6
2.3	Imagens histológicas após o processo de coloração. . . . .	7
2.4	Matriz de Confusão. . . . .	13
4.1	Exemplo de imagen com amiloidose na coloração vermelho congo. . .	21
4.2	Exemplos de Imagens fora do escopo presentes na base de dados original.	22
4.3	Exemplos de Imagens presentes no <i>Dataset Baseline</i> . . . . .	23
4.4	Arquitetura do Modelo <i>Ensemble-Based</i> Proposto. . . . .	27
5.1	Acurácia de Treinamento e Validação para o <i>Fold</i> com melhor Sensi- bilidade. . . . .	30
5.2	Matriz de Confusão para os melhores modelos treinados com o <i>Dataset</i> <i>Baseline</i> . . . . .	32
5.3	Acurácia de Treinamento e Validação para os melhores Modelos. . . .	37
5.4	Matriz de Confusão para os melhores modelos treinados com o <i>Dataset</i> <i>RUS</i> . . . . .	39
5.5	Acurácia de Treinamento e Validação para o <i>Fold</i> com melhor Sensi- bilidade . . . . .	43
5.6	Matriz de Confusão para os melhores modelos treinados com o <i>Dataset</i> <i>ROS</i> . . . . .	45
5.7	Matriz de Confusão para os melhores modelos. . . . .	49

# Lista de Abreviações

<b>Abreviação</b>	<b>Descrição</b>
CNN	<i>convolutional Neural Network</i>
DL	<i>Deep Learning</i>
SVM	<i>Support Vector Machine</i>
VP	<i>Verdadeiro Positivo</i>
VN	<i>Verdadeiro Negativo</i>
FP	<i>Falso Positivo</i>
FN	<i>Falso Negativo</i>
AUC	<i>Area Under Curve</i>

# Capítulo 1

## Introdução

A aplicação de métodos computacionais no auxílio ao diagnóstico por meio de imagens patológicas provocou o surgimento da Patologia Computacional. Nela, o trabalho de médicos patologistas é auxiliado por programas de computadores, construídos com algoritmos de aprendizado de máquina que realizam análise e classificação de imagens médicas, ajudando assim na realização de diagnósticos mais rápidos e precisos (Agibetov et al., 2021)(Abels et al., 2019)(Calumby et al., 2023).

A despeito da expansão da aplicação de métodos de aprendizagem de máquina para auxiliar o diagnóstico de lesões histológicas em imagens digitais de biópsias, muito ainda há por fazer na área de patologia renal ou nefropatologia. A complexidade das imagens obtidas impõe um grande desafio para a construção de classificadores de lesões nas estruturas renais, principalmente devido à demanda por um número expressivo de imagens para que estes métodos de aprendizado de máquina alcancem um desempenho aceitável ao nível de poderem operar como ferramenta de auxílio ao diagnóstico. Isso se torna crítico para as lesões pouco frequentes ou raras, pois não é possível coletar um grande número de amostras para treinar os classificadores. No caso da amiloidose, uma lesão rara causada por acúmulos de proteínas deformadas nos rins, existe o agravante de que a geração das imagens demanda a utilização de corantes específicos, o que reduz ainda mais o número de imagens disponíveis para esta lesão, criando, do ponto de vista computacional, um cenário de forte desbalanceamento de dados para treinamento de um classificador (Monteiro e Diz, 2015).

Diversas soluções são propostas para lidar com situações de desbalanceamento de classes, podendo envolver estratégias relacionadas tanto ao tratamento de dados quanto ao desenvolvimento de algoritmos. Estratégias de tratamento de dados visam criar uma nova base de dados, na qual as quantidades de amostras por classe sejam equivalentes. Para tal, podem ser aplicadas técnicas de rebalanceamento de classes antes do treinamento do classificador, tais como *oversampling* e *undersampling*. É possível também tratar o problema nos próprios algoritmos de treinamento dos modelos, utilizando técnicas como o ajustes do pesos por classe no processo

de treinamento do classificador e a aplicação de paradigmas de aprendizado de máquina como o *Ensemble-Based*, que de modo geral, combina um conjunto de modelos que podem aprender a lidar com as classes majoritária e minoritária de maneiras diferentes (Wang et al., 2020).

No cenário da patologia computacional, o projeto *PathoSpotter* (<https://pathospotter.bahia.fiocruz.br>) propõe um conjunto de ferramentas computacionais para auxiliar médicos patologistas em trabalhos como anotação, busca e classificação de imagens digitais de biópsias renais. Nesse contexto, uma vez que o PathoSpotter já conta com classificadores para as lesões renais mais comuns, as pesquisas avançaram para a classificação de lesões raras, escolhendo-se nesse momento a amiloidose. Considerando que essa lesão requer que a preparação das lâminas de biópsia com um corante específico, o que encarece e dificulta a sua análise, além de diminuir a quantidade de lâminas disponíveis para que o patologista possa realizar o diagnóstico, a equipe médica do projeto sugeriu a construção de um classificador para análise de amiloidose em lâminas disponíveis com outros corantes não específicos para esta lesão. Esse recurso facilitaria enormemente o diagnóstico, uma vez que a quantidade de lâminas coradas aumentaria substancialmente, aumentando a robustez dos diagnósticos de amiloidose, além de abrir margem para uma simplificação no processo de obtenção das lâminas.

Assim, este trabalho enfrentou o desafio de criar, a partir de modelos de redes neurais convolucionais públicas, um classificador que possa detectar a presença de amiloidose em imagens não coradas com o corante específico usado para esta lesão. Devido a raridade da amiloidose, surge um problema adicional de desbalanceamento de classe, que também precisa ser enfrentado para que se possa resolver o desafio principal. Essa solução tem potencial para simplificar o protocolo de diagnóstico de amiloidose renal, já que poderá evitar a necessidade da utilização de corantes especiais para a avaliação da lesão por médicos especialistas, com ganhos operacionais importantes do ponto de vista da nefropatologia.

Os modelos de classificação apresentados nesse trabalho foram construídos utilizando modelos públicos de redes neurais convolucionais, inicialmente treinados com a base de dados original desbalanceada, para que fosse estabelecida uma linha de base de comparação da pesquisa. Em uma segunda etapa foram aplicadas técnicas para lidar com desbalanceamento de classes, tanto no sentido de aplicar estratégias relacionadas ao tratamento de dados, como *Random Undersample* e *Random Oversample*, quanto ao desenvolvimento de algoritmo, como o métodos *Ensemble-Based*.

Até onde pode-se avaliar, esse é o primeiro trabalho que propõe um método automático de classificação de amiloidose renal, ainda mais usando imagens com corantes não específicos. Entre os modelos de classificadores desenvolvidos, o melhor resultado foi obtido pelo modelo baseado na rede *Inception*, com uma taxa de falso negativo de 4,5%. Essa medida foi tomada como referência neste trabalho, já que entende-se que uma falha na classificação da amiloidose (falso negativo) tem pior consequência para os pacientes que uma classificação equivocada (falso positivo), já

que outras variáveis clínicas são analisadas além do diagnóstico da imagem, o que reduz o risco de um provável tratamento desnecessário para a doença.

### 1.0.1 Justificativa

Até o presente momento não foram encontradas disponíveis na literatura pesquisas ou trabalhos na área de classificação automática de Amiloidose Renal. Nesse sentido, os resultados deste trabalho contribuem como um estudo inicial sobre a classificação automática de amiloidose renal, servindo como referência para estudos posteriores e trabalhos mais aprofundados. Além disso, o classificador desenvolvido será incorporado ao *PathoSpotter*, agregando assim mais uma funcionalidade importante para o projeto.

O principal desafio imposto para se lidar com a amiloidose é o fato de que para que ela seja identificada em imagens histológicas é necessária a utilização de corantes específicos. Este requerimento limita enormemente o número de imagens disponíveis da lesão, gerando uma maior complexidade no momento da realização de diagnósticos. Nesse sentido, dado que este trabalho se propôs a desenvolver e avaliar um classificador que possa detectar a presença de amiloidose em imagens não coradas com o corante padrão usado para esta lesão, a solução proposta aqui tem potencial para simplificar o protocolo usado na análise desta lesão, considerando que pode dispensar a necessidade de corantes especiais para o processamento da biópsica, o que resultará em benefícios operacionais significativos do ponto de vista da nefropatologia.

Além disso, dado que a classificação de amiloidose traz inerente o problema de desbalanceamento de classes, foram realizadas também análises comparativas entre diferentes métodos para lidar com esta situação, utilizando como referencial teórico abordagens disponíveis na literatura. Com base nos resultados coletados, foi realizada uma avaliação criteriosa dos impactos, vantagens e desvantagens de cada uma das abordagens e se uma se sobressai sobre as demais.

Este trabalho utilizou como base de dados, imagens histológicas coradas utilizando os corantes AZAN, HE, PAMS e PAS, desta forma será possível avaliar se e como estes corantes influenciam o resultado de classificação. Assim, uma vez que se tenha identificado se um corante influencia positivamente ou negativamente na classificação, este trabalho servirá também como referencial para orientar a seleção dos melhores corantes não específicos para análise automática de amiloidose.

### 1.0.2 Objetivos

#### Objetivo Geral

Realizar classificação automática de amiloidose renal em imagens digitais de biópsias renais utilizando corantes não específicos para esta lesão.

**Objetivos Específicos**

- Construção de um dataset de imagens digitais de amiloidose usando corantes não específicos;
- Comparar quais técnicas de balanceamento de classe são mais adequadas para o problema;
- Comparar a influência dos corantes não específicos sobre o desempenho da classificação;
- Desenvolver um classificador de amiloidose utilizando redes neurais convolucionais clássicas;

# Capítulo 2

## Referencial Teórico

Neste capítulo são abordados os principais conceitos que são utilizados neste trabalho, servindo como base para o entendimento dos métodos, experimentos e resultados obtidos na pesquisa.

### 2.0.1 Conceitos de Nefropatologia

Nesta sessão são introduzidos os principais conceitos de nefropatologia utilizados neste trabalho, como rins, glomérulo e amiloidose, servindo como referencial teórico para o entendimento dos mesmo.

#### Rins e Glomérulo

O sistema renal é formado por dois rins, que são responsáveis por uma parte das funções de filtração e excreção de substâncias que são removidas do sangue durante o processo de filtração do mesmo. A urina é o produto final do processo de filtração do sangue e, através dela, são eliminados água, sais minerais, resíduos metabólicos, dentre outras substâncias. Na Figura 1 pode-se observar uma amostra de um rim (Moraes e Colicigno, 2007).

O glomérulo é uma das estruturas que compõem o rim. Ele é formado por uma rede de capilares vasculares que tem como principal função a filtração do plasma sanguíneo (Moraes e Colicigno, 2007). Um exemplo de glomérulo pode ser observado dentro do círculo na Figura 2.1. A compreensão dessa estrutura é importante para que se possa ter um entendimento fundamental de como a amiloidose afeta os rins, visto que o glomérulo é uma das estruturas do órgão afetada pela lesão.

#### Amiloidose

A amiloidose é uma doença causada por acúmulos de proteínas deformadas em tecidos do corpo humano e afeta órgãos como coração e rins. Essas proteínas são chamadas de proteínas amilóides e quando acumuladas acarretam a formação de

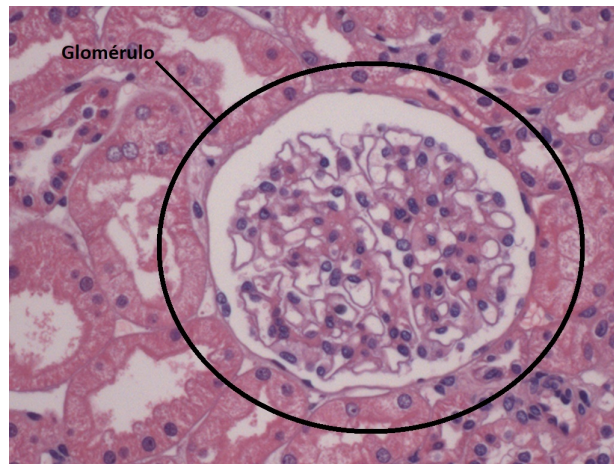
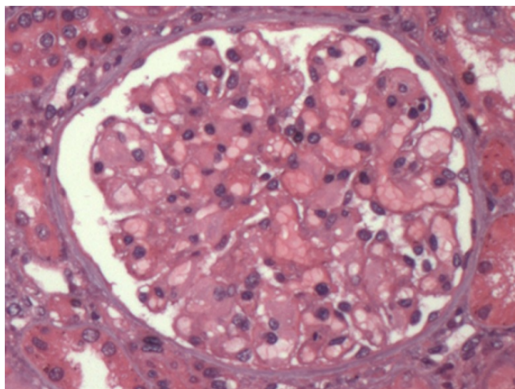


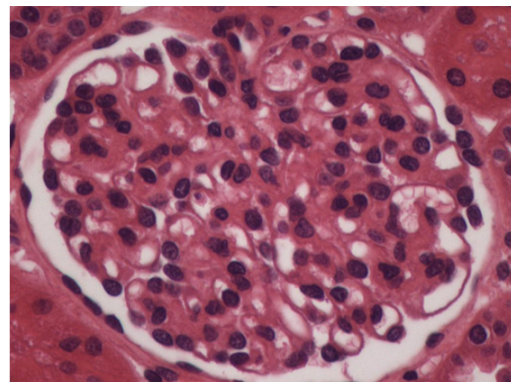
Figura 2.1: Exemplo de uma imagem histológica de um glomérulo.

fibrilas insolúveis, que por sua vez, podem causar perda ou disfunção do órgão e em um último estágio, a morte do paciente (Agibetov et al., 2021)(Palladini et al., 2003). No caso da amiloidose renal, o acúmulo de proteínas amiloides ocorrem no rim como um todo e quando atingem os glomérulos acarretam uma deformação dos mesmos, tendo como consequência um comprometimento de sua função.

A Figura 2.2 mostra um exemplo de um tecido renal com amiloidose e sem amiloidose, coloridos utilizando o corante HE.



(a) Exemplo de Imagem com Amiloidose



(b) Exemplo de Imagem sem Amiloidose

Figura 2.2: Exemplo de um glomérulo com e sem Amiloidose.

Entretanto, apesar de ser uma doença conhecida, o número de casos de amiloidose identificados ainda são baixos, fazendo com que o número de registros da doença disponíveis para realização de pesquisas seja escasso (Sipe et al., 2016).



## Biopsia

Uma das formas com a qual se faz diagnóstico de amiloidose é por meio da realização de uma biópsia renal. Uma biópsia é um procedimento médico no qual uma pequena amostra do tecido renal de interesse é coletada e posta em uma lâmina de microscópio para ser estudada (Monteiro e Diz, 2015). A partir desta análise, é realizada a digitalização da biópsia no formato de imagem, que então é salva para análises futuras. Uma outra etapa importante que é feita para que se possa realizar um estudo de tecido renal no microscópio é a coloração histológica, que será discutida com mais detalhes na sessão a seguir.

## Coloração Histológica

A coloração de tecidos é uma etapa importante para analisar as estruturas e componentes presentes no mesmo. Ela consiste no processo de adição de uma substância química, o corante, junto ao tecido estudado, sendo que o método escolhido para tal depende do objetivo da análise e das estruturas que se deseja realçar (Abels et al., 2019).

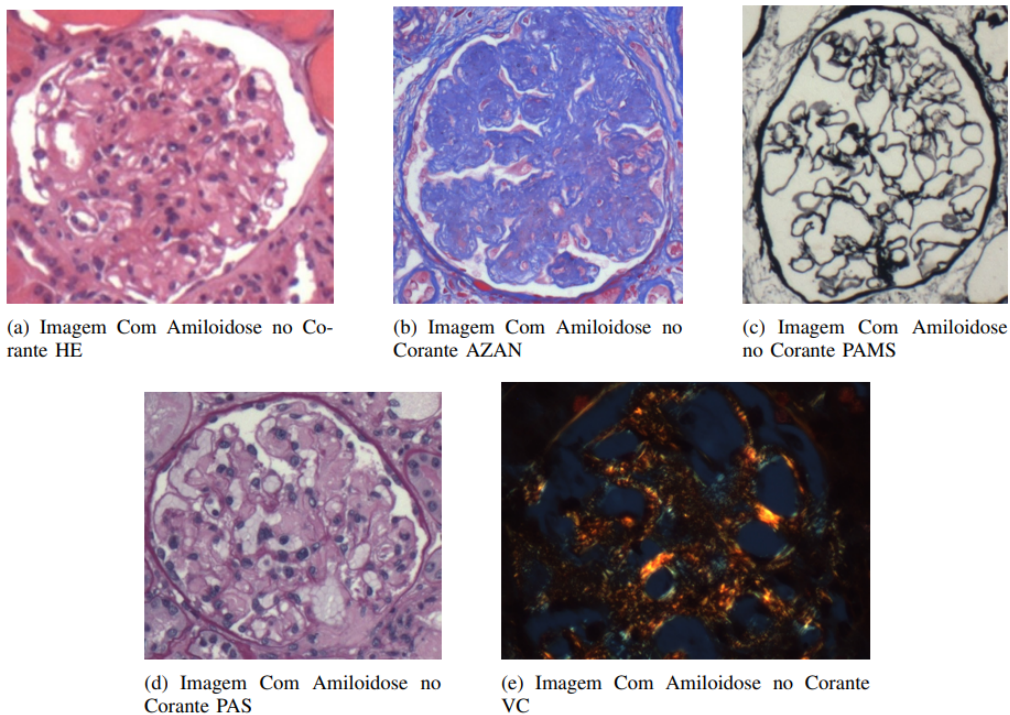


Figura 2.3: Imagens histológicas após o processo de coloração.

Os corantes mais comuns e mais utilizados para observar estruturas gerais dos tecidos são os hematoxilina e eosina (HE) (Timm, 2005). Ainda assim, diversos outros métodos são utilizados no processo de coloração, como por exemplo os ácido periódico de Schiff (PAS), o método de PAMS (PAMS) e o Tricromo de Azan (AZAN).

No caso da Amiloidose, o corante padrão utilizado é o vermelho congo, já que ele permite ressaltar a lesão de forma que possa ser identificada por um médico patologista (Monteiro e Diz, 2015). A Figura 2.3 mostra exemplos de imagens com amiloidose após coloração com diferentes tipos de corantes.

## 2.0.2 Conceitos de Computação

Nesta sessão são introduzidos os principais conceitos de Computação utilizados neste trabalho, como aprendizado de máquina e aprendizado profundo (Do inglês: *deep learning*). O entendimento desses conceitos permitirá ao leitor ter uma maior compreensão do que foi realizado neste estudo, do ponto de vista computacional.

### Patologia Computacional

A Patologia Computacional é um campo de estudo no qual técnicas de aprendizado de máquina e reconhecimento de padrões são utilizadas para auxiliar e aprimorar a realização de diagnósticos médicos via análise e processamento de imagens digitais de biopsias. Alguns dos principais objetivos deste campo de estudo são melhorar a precisão de diagnósticos e viabilizar análises epidemiológicas em larga escala (Agibetov et al., 2021)(Abels et al., 2019).

A digitalização de imagens de biopsias, através de *scanners* de imagens ou via captura de imagens utilizando uma câmera convencional acoplada a um microscópio(Chagas et al., 2020), permitiu a ampliação do uso da computação tanto no processamento das imagens quanto no auxílio à análise de imagens na patologia. Na última década, a aplicação de algoritmos de *Deep Learning* na área médica proporcionou uma melhoria expressiva no desempenho das tarefas de aprendizagem de máquina relacionadas a análise de imagens médicas. Este advento não somente facilitou a realização de análises de imagens pelos médicos, como tornou possível a construção de programas mais avançados para dar auxílio ao diagnóstico médico feito por patologistas (Chagas et al., 2020)(Abels et al., 2019).

### Aprendizado de Máquina

Aprendizado de máquina é uma área da computação que se dedica ao estudo e construção de sistemas capazes de obter conhecimento de forma automática a partir de um conjunto de dados. Estes algoritmos podem ser utilizados nos mais diversos campos do conhecimento, como processamento de linguagem natural, previsões diversas e visão computacional (Shinde e Shah, 2018).

Algoritmos de aprendizado de máquina são divididos em dois grupos: aprendizado supervisionado e não supervisionado. Para o aprendizado não supervisionado, o algoritmo utilizado é aplicado a um conjunto de dados não rotulados, com o objetivo de identificar padrões ou agrupamentos (Shinde e Shah, 2018). Dois exemplos de algoritmos de aprendizado supervisionado são o *K-means* e o *K-Apriori* (Mahesh, 2020).

Para o aprendizado supervisionado, os algoritmos são treinados utilizando uma base de dados com amostras rotuladas. Dentro do aprendizado supervisionado, existem os algoritmos chamados de algoritmos de classificação que são responsáveis por determinar em qual classe uma amostra deve ser enquadrada e pode, ser agrupados em três tipos:

- Classificação Binária: Quando a amostra pode ser classificada em somente uma classes, de um conjunto de 2 classes.
- Classificação Multiclasse: Quando a amostra pode ser classificada em somente uma classes, de um conjunto de N classes.
- Classificação Multilabel: Quando a amostra pode ser classificada em múltiplas classes, dado um conjunto de N classes.

### ***Deep Learning* e Redes Neurais Convolucionais**

*Deep Learning* é um tipo de algoritmo usado em aprendizado de máquina, constituído por meio de modelos de redes neurais complexos, em termos de número de camadas e parâmetros de treinamento. Este tipo de algoritmo se disseminou nos últimos anos, devido a um aumento do poder computacional de hardwares, bem como na quantidade de dados disponíveis, que permitiram o treinamento destes modelos de forma mais rápida e com mais qualidade (Shinde e Shah, 2018).

Algoritmos de aprendizado de máquina possuem limitações em termos de processamento de dados naturais em sua forma bruta, como textos e imagens. Tipicamente nesses cenários, era necessário a aplicação de uma série de transformações e tratamentos nos dados, com o objetivo de extrair informações ou características dos mesmos, que então seria utilizadas pelos algoritmos para obtenção dos modelos de representação dos dados. Este trabalho exigia não somente um maior cuidado em termos de análise e desenvolvimento de código, como também a necessidade de profissionais com um maior expertise no domínio do problema (Shinde e Shah, 2018).

Esta dificuldade foi minimizada ou até mesmo superada com o desenvolvimento de algoritmos de *deep learning*, devido a sua capacidade de extrair informações, características e padrões de forma automática. No caso de imagens, por exemplo, são aplicadas uma série de operações matemáticas que automatizam a obtenção de um vetor que representa a imagem por meio de suas características extraídas (Shinde e Shah, 2018).

Redes neurais convolucionais (do inglês *Convolutional Neural Networks* ou CNN) são as arquiteturas de *Deep Learning* que permitem obter alta eficiência na análise de imagens (Arafa et al., 2022). Existem registros de aplicações de CNN em imagens médicas desde os anos de 1990, mas somente com desenvolvimento de hardware com mais poder de cômputo e um aumento na quantidade de dados disponíveis, foi possível a disseminação destes algoritmos (Tajbakhsh et al., 2016) (Medela et al., 2019a).

Uma série de modelos diferentes de CNN são utilizados em trabalhos de classificação de imagens médicas. Dentre estes modelos, pode-se citar o *VGG* (Simonyan e Zisserman, 2014), o *Inception* (Szegedy et al., 2016) e o *Resnet* (He et al., 2016). Tais modelos são comumente aplicados através do uso de estratégias de *transfer learning*, que transportam o conhecimento do modelo em um determinado domínio para outro (Medela et al., 2019b).

Para que seja possível, com o uso de CNN, alcançar desempenho semelhantes ao de um especialista, é necessário a utilização de base de dados com um grande volume de informações previamente rotulado (Medela et al., 2019b). Porém, existem situações na medicina, como é o caso do problema tratado neste trabalho, que a construção de uma base de dados com grande quantidades de amostras para ambas as classes não é uma tarefa possível. Para este tipo de caso, faz-se necessário a aplicação de métodos específicos que consigam lidar com o problema de desbalanceamento.

### **Problemas de Desbalanceamento de Classe**

Problemas de desbalanceamento de classe ocorrem quando a quantidade de amostras disponíveis por classes é assimétrica e a diferença é grande ao ponto de uma das classes se tornar preponderante em relação às demais. Este desbalanceamento pode ser devido a uma característica própria do problema ou a uma falta de meios para se obter dados em quantidade satisfatória para ambas as classes (Abd Elrahman e Abraham, 2013).

A medição de desempenho de algoritmos de classificação no cenário de balanceamento deve ser feita de forma bastante criteriosa, já que o uso de uma métrica inadequada pode gerar um indicador enganoso de bom desempenho, uma vez que os erros oriundos de uma classificação enganosa para a classe minoritária são mascarados pela classe majoritária (Johnson e Khoshgoftaar, 2019).

Os métodos propostos na literatura para lidar com situações de desbalanceamento de classe podem ser agrupados em três grandes grupos: Estratégia de balanceamento de dados, estratégia de desenvolvimento de algoritmo para lidar com desbalanceamento de dados e uma estratégia híbrida, em que ambas as estratégias anteriores são combinadas.

### **Estratégias de Balanceamento de Dados**

Estratégias de balanceamento de dados englobam um conjunto de métodos que atuam na base de dados utilizada no processo de treinamento dos algoritmos. Dentre estes métodos, pode-se citar técnicas como *undersampling* e *oversampling*. Nas seções a seguir serão discutidas com mais detalhes o funcionamento das técnicas citadas.

## Métodos de Amostragem

Métodos de amostragem buscam reduzir o nível de desbalanceamento de uma base de dados por meio da realização de ajustes na distribuição dos dados da classe minoritária ou majoritária. *undersampling* e *oversampling* são duas técnicas classificadas como métodos de amostragem (Johnson e Khoshgoftaar, 2019).

*Undersampling* realiza balanceamento por meio da coleta de amostras da classe majoritária, reduzindo assim a quantidade de dados disponíveis para o treinamento de um algoritmo de aprendizado de máquina. O processo de coleta de amostras pode ser feito de diversas maneiras, como uma amostragem aleatória, *random undersampling* (RUS), ou por meio da aplicação de algoritmos mais complexos como o *edited nearest neighbours* (ENN) ou o *k-nearest neighbours* (KNN). A aplicação de estratégias mais sofisticadas de amostragem são importantes, para alguns tipos de dados como dados tabulares, para que seja possível manter preservadas informações relevantes presentes na base de dados, como amostras significativas e padrões, que poderiam ser perdidos caso fosse feito uso de uma amostragem aleatória. Ao mesmo tempo, estes tipos de amostragem permitem também remover informações redundantes ou ruídos nos dados que possam causar um viés de treinamento (Johnson e Khoshgoftaar, 2019).

Já o *oversampling* é uma outra estratégia que realiza o balanceamento dos dados por meio do aumento do tamanho da quantidade de amostras da classe minoritária. Este aumento de dados pode ser feito por meio da criação de cópias das amostras da classe minoritária selecionadas aleatoriamente, usando técnicas como *random oversampling* (ROS), *SMOTE* (Técnica de sobreamostragem minoritária sintética, do inglês - *Synthetic minority over-sampling technique*) e suas variações, como o *Borderline-SMOTE* e o *Safe-Level-SMOTE*. Um dos problemas associados ao uso de ROS é que ele pode causar *overfitting* para a classe minoritária, na qual a técnica foi utilizada. O SMOTE e suas variações são algoritmos utilizados em problemas de desbalanceamento de classe que geram novos dados sintéticos para a classe minoritária, e por conta disso, conseguem reduzir os problemas associados ao uso do ROS, como o *overfitting* (Johnson e Khoshgoftaar, 2019).

Quando se trata de imagens, podem ser aplicadas tanto estratégias de *undersampling* quanto estratégias de *oversampling*. Dentre as estratégias de *undersampling* associadas a imagens, pode-se identificar a *random undersampling* (RUS) ou amostragem baseada em grupos. Dentre as estratégias de *oversampling* associadas a imagens, pode-se citar o *oversampling* por meio da replicação de imagens e o uso da técnica SMOTE.

## Algoritmo para lidar com Desbalanceamento de Dados

Nessa sessão é apresentado os conceitos associados a estratégia estudada neste trabalho para lidar com desbalanceamento de classe por meio da aplicação de desenvolvimento de algoritmo.

### Métodos *Ensemble-Based*

O método *Ensemble-Based* funciona por meio da combinação da predição de diversos classificadores para determinar o resultado final da classificação. A aplicação desta técnica gera uma melhoria na habilidade de generalização dos algoritmos de aprendizado de máquina clássicos, como o SVM e árvores de decisão (Taherkhani et al., 2020a), bem como um ganho de desempenho. Esta melhoria nas capacidades de generalização e taxa de acerto obtidas pelo uso da técnica citada ocorre devido ao fato de que a combinação de diversos classificadores resulta em um modelo final com desempenho mais elevado do que o de cada um dos classificadores individualmente (Johnson e Khoshgoftaar, 2019).

As duas abordagens do método *Ensemble-Based* mais comuns são a *boosting* e a *bagging*. Na abordagem de *bagging*, cada um dos classificadores individuais são treinados de forma independente, podendo utilizar diferentes porções do *dataset* original para cada classificador, ou um conjunto de modelos diferentes, obtendo-se assim uma maior diversidade nos modelos. No final, quando uma amostra desconhecida é apresentada, o resultado da classe é definido por meio da votação de cada classificador individualmente, escolhendo-se a classe mais votada (Galar et al., 2011).

Na abordagem de *boosting* o treinamento se dá com toda a base de dados e cada classificador é dependente do classificador anterior, focando em reduzir os erros de seu antecessor. Para isso, após cada etapa o classificador seguinte dedica mais tempo de treinamento nas instâncias mais difíceis (Entende-se instâncias mais difíceis as que apresentaram mais erros nos classificadores anteriores), gerando assim um resultado final com maior desempenho. O *AdaBoost* é um dos algoritmos mais conhecidos que usa o método de *boosting*, sendo citado como um dos melhores algoritmos de aprendizado de máquina (Galar et al., 2011).

A utilização do método *ensemble-based* encontra aplicabilidade em contextos caracterizados pela presença de desbalanceamento de classes, devido à sua capacidade de atenuar os efeitos prejudiciais decorrentes do viés de classificação para a classe majoritária., decorrente da disparidade na distribuição de dados entre as classes positiva e negativa. Nesse contexto, a assembleia de modelos diferentes proporcionada pelo método *ensemble-based* viabiliza a construção de um modelo final mais robusto que demonstra a habilidade de apreender características inerentes tanto a classe majoritária quanto a minoritária, as quais poderiam ter sido negligenciadas devido à predominância evidenciada pela classe com mais dados (Galar et al., 2011).

### Métricas de Avaliação de Desempenho

Diversas métricas são utilizadas para avaliar o desempenho de modelos de machine learning, como por exemplo acurácia ou precisão. Duas métricas típicas utilizadas na avaliação de desempenho de um modelo de aprendizado de máquina são acurácia e taxa de erro, entretanto, elas não são apropriadas para lidar com situações de desbalanceamento de classe. Ambas são altamente enviesadas pela classe majoritária,

tendo seu resultado fortemente determinado por ela, enquanto que a classe minoritária exerce pouca influência no valor final obtido pela medida (Abd Elrahman e Abraham, 2013).

Quando se trata de problemas de classificação binária, as métricas de avaliação utilizadas são derivadas da matriz de confusão, mostrada na Figura 2.4.

		Valor Predito	
		Sim	Não
Valor Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 2.4: Matriz de Confusão.

A matriz de confusão é composta por medidas utilizadas no processo de avaliação de algoritmos de aprendizado de máquina. São elas:

- Verdadeiros Positivos (VP): Expressa a quantidade de amostras corretamente classificadas como positivas.
- Verdadeiros Negativos (VN): Expressa a quantidade de amostras corretamente classificadas como Negativas.
- Falso Positivos (FP): Expressa a quantidade de amostras que foram classificadas como Positivas incorretamente, uma vez que elas pertencem ao grupo Negativo.
- Falso Negativo (FN): Expressa a quantidade de amostras que foram classificadas como Negativas incorretamente, uma vez que elas pertencem ao grupo Positivo.

A partir destas medidas, são desenvolvidas as seguintes métricas utilizadas no processo de avaliação de algoritmos de aprendizado de máquina:

- Acurácia: A acurácia é uma medida que fornece a proporção entre o número total de classificações corretas sobre o número de amostras a serem classificadas.

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \quad (2.1)$$

- **Precisão:** Expressa a proporção entre os verdadeiros positivos e a soma dos verdadeiros positivos, e os falsos positivos. Esta métrica identifica qual a proporção de identificações positivas que realmente estavam corretas.

$$Precisao = \frac{VP}{VP + FP} \quad (2.2)$$

- **Sensibilidade (*Recall*):** Também chamada de Taxa de Verdadeiros Positivos, expressa a proporção entre os verdadeiros positivos e a soma dos verdadeiros positivos com os falsos negativos. Nesse sentido, o *recall* expressa uma medida de quão bem o modelo identificou as amostras positivas.

$$Sensibilidade = \frac{VP}{VP + FN} \quad (2.3)$$

- **Especificidade:** Pode ser chamado também de Taxa dos Verdadeiros Negativos, ela expressa a taxa de amostras negativas corretamente classificadas. Ela é calculada por meio da proporção entre os verdadeiros negativos e a soma dos verdadeiros negativos com os falsos positivos.

$$Especificidade = \frac{VN}{VN + FP} \quad (2.4)$$

- **F1-Score:** O F1-Score é a média harmônica entre a precisão e a sensibilidade, representando um balanço entre essas duas métricas.

$$F1\text{-Score} = 2 * \frac{Precisao * Sensibilidade}{Precisao + Sensibilidade} \quad (2.5)$$

- **G-Score:** G-Score, ou G-means, é uma métrica que avalia o balanço de desempenho em ambas as classes, majoritária e minoritária. Esta métrica é útil em problemas de desbalanceamento de classe pois leva em consideração o desempenho de ambas as classes. Um G-Score baixo é um indicativo de que uma das classes está com desempenho ruim, mesmo que a outra tenha um desempenho de classificação elevada.

$$G\text{-Score} = \sqrt{Especificidade * Sensibilidade} \quad (2.6)$$

- **Area Under the ROC Curve (AUC):** A curva ROC pode ser entendida como uma método gráfico que expressa um balanço entre a sensibilidade e a especificidade. Nesse sentido, a AUC é entendida como a área sob a curva ROC e pode ser obtida calculando a média entre a sensibilidade e a especificidade.



$$AUC = \frac{\textit{Sensibilidade} + \textit{Especificidade}}{2} \quad (2.7)$$

Em uma situação de desbalanceamento de classes, as métricas mais comumente usadas são especificidade, sensibilidade (recall), precisão e F-Score, por permitirem uma análise de resultado levando em conta tanto a classe majoritária quanto a classe minoritária. Especificidade e sensibilidade são utilizadas quando se busca medir o desempenho de uma das classes individualmente. Precisão, por sua vez, é utilizada em problemas que demandam um alto desempenho em uma única classe. Já a F-Score e G-Score são utilizadas quando é importante obter um alto desempenho em ambas as classes, tanto a minoritária quanto a majoritária.

## Capítulo 3

# Revisão Bibliográfica

Neste capítulo são apresentados trabalho de outros autores que estão relacionados com este projeto e servem de embasamento para a pesquisa. A leitura desde capítulo permitirá ao leitor ter uma maior compreensão do que já foi pesquisado por outros autores em áreas de similares a desde trabalho.

Em (Litjens et al., 2017) os autores realizaram uma revisão de mais de 300 artigos com diversas aplicações de *Deep Learning* (DL) na medicina, apresentando de forma resumida e centralizada em um único trabalho uma vasta gama de conceitos, métodos e tarefas associadas ao uso de *Deep Learning* nesta área. Dentre os exemplos mostrados no trabalho pode-se citar o uso de estratégias de segmentação e classificação, associadas ao uso de *Deep Learning* em imagens médicas e a apresentação de diferentes arquiteturas de Redes Neurais Convolucionais (CNN) que são aplicadas ao tema.

(Chagas et al., 2020) realizam classificação automática de hiperplasticidade em imagens de glomérulos utilizando redes neurais convolucionais. Os autores propõem uma arquitetura própria de rede neural convolucional para extração de características da imagem acoplado a um classificador do tipo SVM. A arquitetura proposta foi treinada utilizando uma base de dados com 811 imagens, sendo que dessas 511 são de hiperplasticidade e as demais de imagens sem nenhum tipo de doença. Os modelos treinados foram avaliados em termos de acurácia, precisão, sensibilidade e f1 score. Para efeito de estudo, os autores compararam o desempenho da arquitetura proposta com modelos clássicos de redes neurais convolucionais vistos na literatura, como o Xception, ResNet50 e o InceptionV3.

(da Silva et al., 2021) avaliam um conjunto de modelos de redes neurais convolucionais bem como condições para a aplicação de *deep learning* na tarefa de classificação de imagens histológicas com crescente glomerulares. No trabalho, os autores avaliam os desempenhos das redes *Xception*, *InceptionV3*, *MobileNet*, *VGG16* e *ResNet50* para classificação binária, onde as classes foram imagens normais (sem nenhum tipo de lesão) e imagens de glomérulos com crescente. Os autores também aplicaram

uma outra abordagem metodológica, mantendo a classificação binária, mas alterando a classe de imagens normais adicionando a ela imagem com outros tipos de lesão, chamando-a de classe não-crescente. O objetivo desta etapa foi treinar o classificador em um cenário mais próximo ao visto no mundo ao real. Os modelos foram avaliados utilizando as métricas acurácia, precisão, sensibilidade e f1-score e os autores obtiveram resultados onde a rede ResNet50 apresentou desempenho significativamente superior, se comparada com as demais.

(Chagas et al., 2021) avaliaram três diferentes tipos de modelos clássicos de redes neurais convolucionais na classificação nefropatia membranosa, sendo eles o ResNet-18, DenseNet e o Wide-ResNet. Os autores também utilizam uma abordagem de estimativa de incerteza combinando *Monte-Carlo Dropout* e *Test-Time Data Augmentation*, visando construir um modelo com resultados mais realísticos e resilientes a ruídos desconhecidos. Os modelos foram avaliados utilizando as métricas acurácia, precisão, sensibilidade e F1-score e treinados em uma base de dados com 4,682 imagens de glomérulos humanos.

Em (Devi et al., 2020) os autores discutem que estratégias de balanceamento de classe que utilizam técnicas de *undersampling* são efetivas em situações na qual o desbalanceamento de classes não é muito alto, enquanto que estratégias de *oversampling* conseguem lidar melhor com esta situação de extremo desbalanceamento. No trabalho os autores realizam um apanhado sobre uma série de métodos diferentes de *undersampling*. Tais métodos são subdivididos em dois grandes grupos: Abordagem de *Undersampling Pura* e Abordagem de *Undersampling Híbrida*. Métodos de amostragem tidos como puros funcional por meio da seleção de instâncias das classes de interesse. No caso de problemas de desbalanceamento de classe, tais métodos são comumente aplicados na classe majoritária com o objetivo de coletar um número de amostras semelhantes, ou igual, ao da classe minoritária. Dentre os métodos de *undersampling* puros, são citados no artigo o *Random Undersampling* (RUR) e o *Condensed Nearest Neighbour*. Já os métodos de *undersampling* híbridos são técnicas que fundem métodos de amostragem com técnicas de *clustering*, *ensemble learning*, ou ainda algoritmos evolutivos.

(Zhang et al., 2010) apresentam um método de amostragem da classe majoritária baseado em *clusters*. O objetivo principal da técnica é reduzir a perda de informações oriundas do processo de amostragem permitindo realizar uma seleção de grupos representativos da base de dados. Os autores utilizaram 10 bases de dados diferentes para a avaliação do trabalho, todas obtidas do UCI *Repository*. A performance dos métodos foram obtidas por meio do uso das métricas Precisão, Sensibilidade, *F-measure*, *G-means* e BACC.

(Alex et al., 2022) constroem um modelo para classificação de diabetes utilizando o *Pima Indian Dataset* que possuía um total de 768 amostras, sendo que apenas 105 eram de pacientes com diabetes, o que correspondia a aproximadamente 13,6% da base de dados. O artigo realiza investigação sobre o desempenho de diferentes topologias de redes neurais como CNN, LSTM, CNN-LSTM, ConvLSTM e DCNN,

comparando-as com a técnica proposta pelos autores, que consiste em um modelo do tipo LSTM combinado com o SMOTE. Os resultados do trabalho foram avaliados em termos de acurácia, Precisão, Sensibilidade e AUC, no qual o modelo proposto pelos autores foi superior que as demais técnicas apresentadas no trabalho.

(Dablain et al., 2022) apresentam um método de *Oversampling* chamado de DeepSMOTE, que toma como base o algoritmo SMOTE. A proposta do algoritmo é utilizar o método de *Oversampling* construído pelo SMOTE, ampliando-o para dados em dimensões mais altas, como imagens. Após o treinamento do *DeepSMOTE*, ele consegue gerar novas imagens artificiais por meio de uma técnica de *encoder-decoder*. Em um primeiro passo, *encoder* reduz as imagens originais a mapas de características nos quais são aplicados o SMOTE para a geração de novas amostras. Feito isso, o algoritmo de *decoder* gera novas imagens artificiais, recebendo como entrada os mapas de características previamente criados. Para avaliar o algoritmo construído, os autores utilizaram as bases de dados CIFAR-10/SVHN, MNIST/FMNIST e CELEBA e foram avaliados utilizando as métricas G-Mean, ACSA e F1-Score.

Um dos problemas associados ao uso do SMOTE é a geração de dados ruidosos, fazendo com que muitas vezes imagens geradas como pertencentes a classe minoritária sejam classificadas como pertencentes a classe majoritária, ou o contrário.

Em (Maulidevi et al., 2022) os autores combinam o SMOTE com um algoritmo chamado de local *outlier factor* (LOF) com o objetivo de identificar e reduzir possíveis imagens ruidosas que poderiam ser classificadas incorretamente. Para avaliar o algoritmo proposto, comparando-o com o SMOTE original, os autores utilizaram os *datasets Prima*, contendo 768 imagens com 268 imagens pertencentes a classe minoritária, o *dataset Haberman*, contendo 306 imagens as quais 81 são pertencentes a classe minoritária e o *dataset Glass* com um total de 214 imagens das quais 76 são pertencentes a classe minoritária. As métricas escolhidas para avaliar o desempenho do trabalho foram a acurácia, precisão, Sensibilidade, *F-measure* e *Area under the Receiver Operating Characteristic Curve* (AUC), utilizando validação cruzada com *5-Folds*.

Um outro trabalho que apresenta um método para reduzir os dados ruidosos gerados pelo SMOTE foi apresentado em (Arafa et al., 2022), onde os autores propõem um algoritmo chamado de *Reduced Noise SMOTE* (RN-SMOTE). De forma resumida, a ideia do método é aplicar o DBSCAN nos dados gerados pelo SMOTE de modo que os dados ruidosos sejam eliminados antes de serem introduzidos ao dataset e em seguida, aplicar o SMOTE uma segunda vez para rebalancear a base de dados. O método proposto no trabalho foi executado utilizando 9 algoritmos de aprendizado de máquinas diferentes, dentre eles, pode-se citar o *RandomForest*, o *XGBoost* e o SVM. Os autores utilizaram um total de 9 bases de dados diferentes, todas com desbalanceamento de classe. As métricas escolhidas para avaliar o desempenho do trabalho foram as G-Means, MCC, *Kappa*, precisão, Sensibilidade e F1-Score.

(Ravi et al., 2022) utilizam uma abordagem de aprendizado sensível ao custo para li-

dar com um problema de desbalanceamento de classes na classificação de pneumonia pediátrica em imagens de Raio-X, obtidas da base de dados CXR. A abordagem mostrada no trabalho fez uso *transfer-learning* em redes pré-treinadas como *Xception*, *InceptionResnetV2* e *DenseNet201* para servirem de extrator de características e um algoritmo de regressão logística como classificador. Uma etapa intermediária de transformação dos dados feita no trabalho foi a aplicação do algoritmo KPCA para reduzir a dimensionalidade das imagens, antes de passá-las para o classificador. Como resultado, o método proposto conseguiu uma redução nas classificações incorretas e uma melhoria na generalização do modelo.

(Zhuang et al., 2020) utilizam uma combinação de uma técnica de aprendizado sensível ao custo com uma abordagem de fusão de múltiplos classificadores para classificar lesões na pele. O trabalho utilizou o *ISIC Challenge 2019 Dataset*, como base de dados de estudo, dividido nas proporções de 80% para treinamento, 5% para validação e 15% para testes. Foram utilizadas também um total de 12 arquiteturas diferentes de CNN, avaliadas em termos de acurácia, sensibilidade e especificidade. Os resultados demonstraram que a combinação de múltiplos classificadores com a técnica de aprendizado sensível ao custo permitiram a obtenção de resultados melhores e mais robustos.

(Ahmed et al., 2020) propõem a construção de um classificador de lesão de pele em imagens, com uma precisão superior a 89% para todas as classes. Para a construção do classificador, os autores utilizam um método *ensemble-based* com a combinação de três modelos diferentes de redes neurais convolucionais, sendo eles o *Xception*, *Inception-ResNet-V2* e *NasNetLarge*, sendo que o resultado final da classificação é feito por meio da votação dos três modelos. Para o treinamento e validação dos modelos os autores utilizaram a base de dados *ISIC2019*, que possui 8 categorias diferentes de imagens com um total de 25331 amostras. Os autores avaliaram os resultados dos modelos utilizando as métricas acurácia, auc, precisão e sensibilidade.

(Taherkhani et al., 2020b) apresentam o *AdaBoost-CNN*, um algoritmo adaptativo para redes neurais convolucionais para classificar conjuntos de dados multiclasse desbalanceados fazendo uso de *transfer learning* e adaptações no tradicional *AdaBoost* para torna-lo compatível com *deep learning*. A arquitetura da CNN utilizada foi uma arquitetura proposta pelos autores, não fazendo uso de CNNs clássicas disponíveis na literatura. Os autores realizaram comparativo de desempenho da arquitetura proposta no artigo com a CNN ResNet, concluindo que a arquitetura proposta tem desempenho superior que a ResNet para os experimentos realizados. Os autores utilizaram as bases de dados *synthetic dataset*, *CIFAR-10*, *Fashion-MNIST*, *EMNIST*, *EMNIST by-merge* e o *HAR dataset* para treinamento e validação dos modelos, fazendo uso da acurácia como métrica de avaliação.

# Capítulo 4

## Metodologia

Nesta seção são apresentados os métodos e estratégias que foram utilizados no desenvolvimento do trabalho. Nela estão descritos quais os dados que foram utilizados no projeto, bem como os tratamentos que foram aplicados aos mesmos. Também estão descritos os modelos, métodos de treinamento e avaliação de desempenho utilizados neste trabalho.

Com relação aos métodos para lidar com o desbalanceamento de dados, optou-se por seguir o caminho da amostragem, utilizando os métodos *Random Oversampling* e *Random Undersampling* estratificadas por corante e lesão, com o objetivo de preservar a distribuição original dos dados presentes na base de dados desbalanceada. Também foram realizados experimentos utilizando um modelo *Ensemble-Based*, por meio da combinação de diferentes modelos de redes neurais convolucionais consolidados na literatura, com o objetivo de avaliar se a combinação destes modelos resultaria em um classificador mais eficiente na classificação da amiloidose que os mesmos modelos individuais.

### 4.0.1 Ferramentas Utilizadas

Os modelos foram codificados em *Python v3.8* usando os *frameworks Keras v2.6*, *Tensorflow* (Variação GPU) na versão 2.4.1 e *Scikit learn*, versão 0.24.2. Uma GPU *Nvidia Titan RTX* foi utilizada para execução dos experimentos, a versão do *driver* da placa de vídeo foi o 11.6 e o sistema operacional utilizado foi o *Ubuntu 22.04.1*.

### 4.0.2 Dataset

A base de dados de imagens utilizada neste trabalho foi montada por pesquisadores do Instituto Gonçalo Moniz da Fundação Oswaldo Cruz - Bahia (IGM/FIOCRUZ), no qual a classificação das imagens como tendo amiloidose ou não foi sempre feita pela equipe médica, analisando a lâmina do corante específico para esta lesão (vermelho congo, mostrado na Figura 4.1), que é o padrão ouro para diagnóstico desta

lesão. Com isso, a base de dados é constituída por duas classes de imagens de glomérulos. A primeira contendo somente amiloidose, com um total de 374 imagens. A segunda classe é formado por 4 grupos de imagens: glomérulos sem nenhum tipo de lesão, com lesão do tipo esclerose pura sem crescente, com lesão do tipo hiperce-lularidade, e com lesão do tipo hiperce-lularidade pura sem crescente, com um total de 4015 imagens. Assim, tem-se uma base de dados com duas classes, a primeira chamada de classe Amiloidose e a segunda denominada de classe Não-Amiloidose. A decisão de agrupar na classe Não-Amiloidose diferentes tipos de lesões juntamente com imagens sem qualquer lesão nos glomérulos teve o objetivo de criar uma base de dados de treinamentos próxima com o cenário encontrado na prática pelos pato-logistas.

Como já foi dito, as imagens com amiloidose foram obtidas de pacientes em que as biópsias foram diagnosticadas com o corante específico para a lesão. Para estes casos, como em qualquer biópsia renal, diversas lâminas são obtidas para o mesmo paciente, cada uma tratada com um corantes específico para ressaltar outras lesões que não a amiloidose, montando assim a base de dados utilizada neste trabalho.

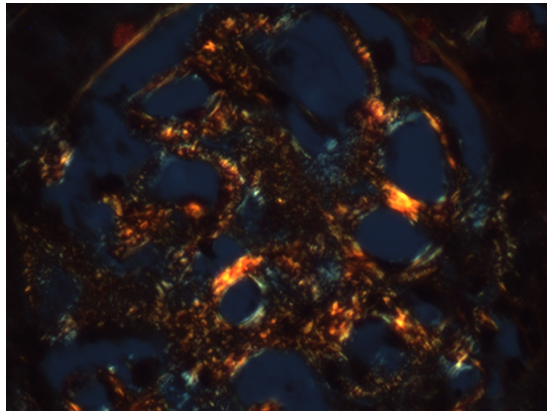


Figura 4.1: Exemplo de imagen com amiloidose na coloração vermelho congo.

A base de dados original, continha uma série de imagens fora do escopo para treina-mento de modelos, como por exemplo: imagens de lâminas completas, imagens com mais de um glomérulo, imagens com glomérulos cortados e imagens com diversas estruturas renais sem interesse para este estudo. A Figura 4.2 demonstram alguns dos exemplos de imagens fora do escopo citadas acima.

Devido a presença de imagens fora do escopo na base de dados original, foi necessária a realização de uma curadoria nas imagens para gerar uma base de dados adequado ao trabalho, eventualmente realizando um processamento para ajuste de tamanho ou remoção de informação de *background*. Este tratamento foi feito através da aplicação do seguintes passos:

1. Remover imagens de lâminas completa.

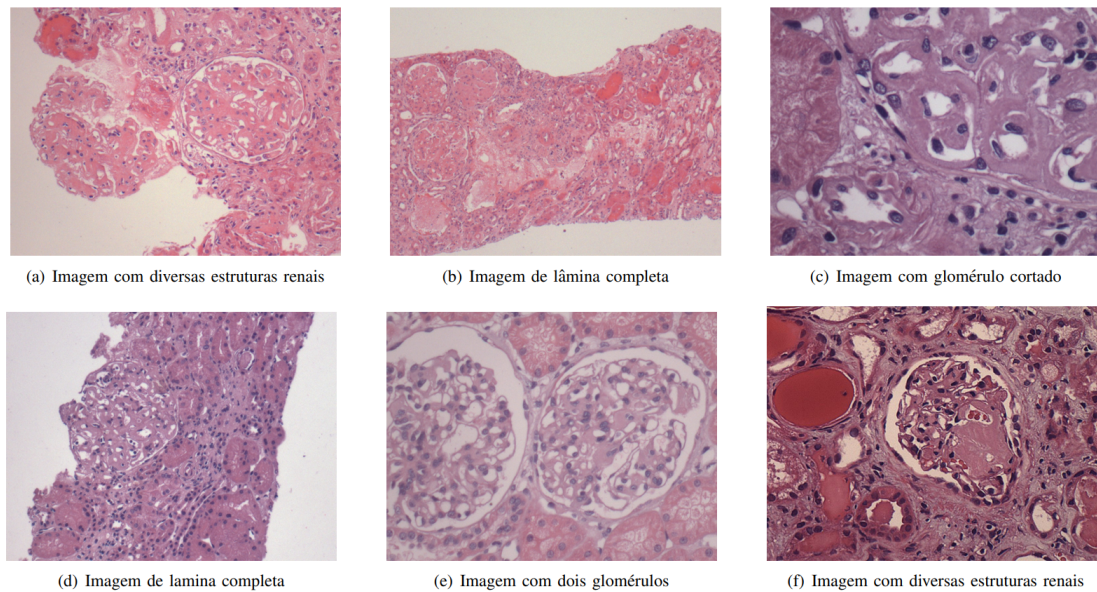


Figura 4.2: Exemplos de Imagens fora do escopo presentes na base de dados original.

2. Remover imagens com mais de um glomérulo para que cada amostra na base de dados apresenta somente um glomérulo por imagem.
3. Realizar recorte nas imagens com somente um glomérulo, centralizando o mesmo na imagem.

A Figura 4.3 mostra exemplos de imagens presentes na base de dados após tratamento, para as duas classes, Amiloidose e Não-Amiloidose, com amostras por corante.

A Tabela 4.1 mostra quantidade final de imagens presentes na base de dados após tratamento, dividida por classe e corantes não específicos para a detecção da lesão, sendo eles HE, AZAN, PAS, PAMS. A Tabela 4.2 mostra quantidade final de imagens, dividida por lesão e corantes não específicos para a amiloidose. Para ambas as tabelas pode-se ver também a porcentagem dos corantes para cada uma das classes e lesões. Como pode ser visto pela quantidade de imagens em cada uma das classes, tem-se uma base de dados com forte desbalanceamento para a classe Amiloidose. Para efeito de estudo, a classe Amiloidose será entendida como a classe positiva, enquanto que a classe Não-Amiloidose será vista como a classe negativa. Além disso, a base de dados original desbalanceada, após tratamento de dados, será chamada de *Dataset Baseline*.



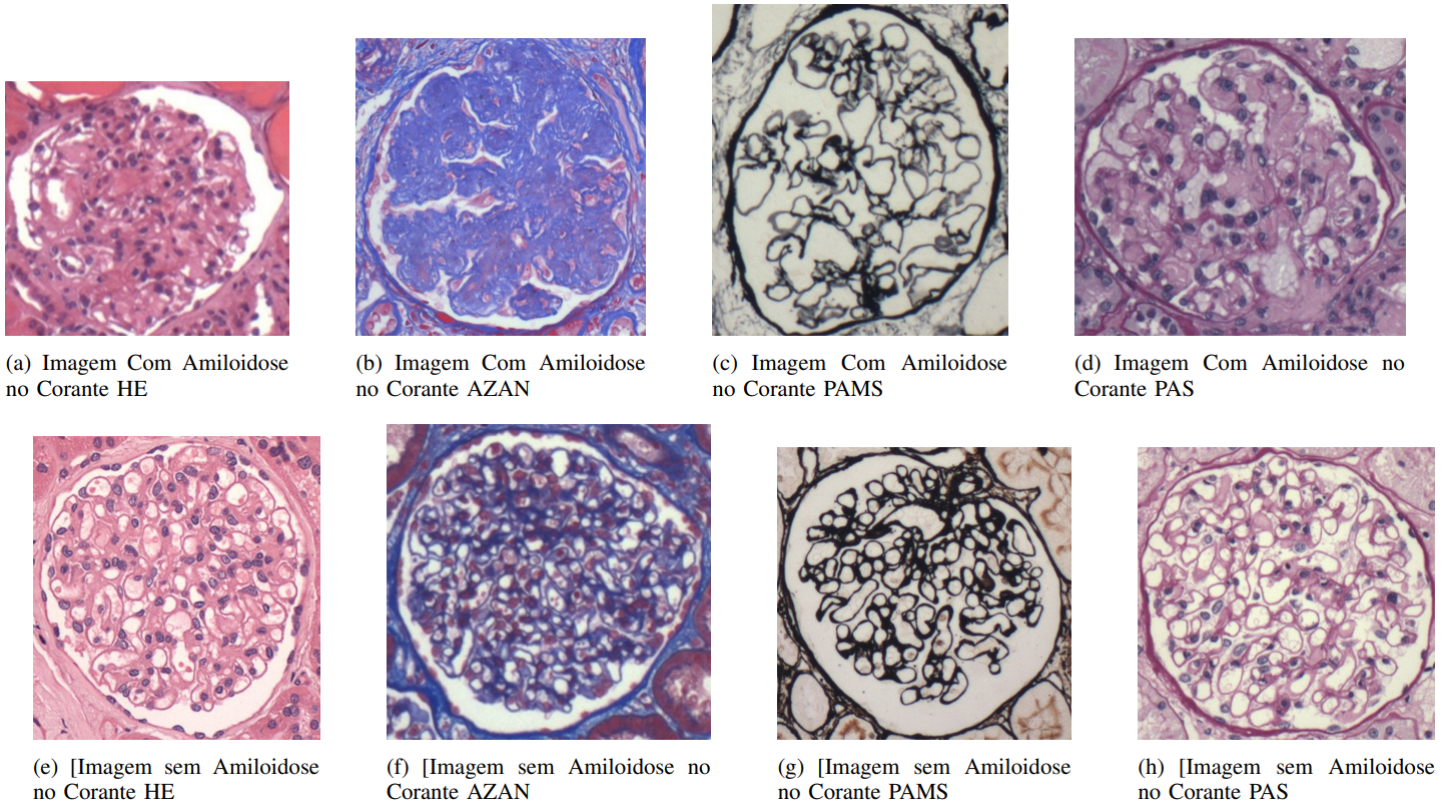


Figura 4.3: Exemplos de Imagens presentes no *Dataset Baseline*.

Tabela 4.1: Relação da Quantidade de Imagens por Corante para o Dataset RUS.

Classe	AZAN		HE		PAS		PAMS		Total
	N° de Imagens	%	N° de Imagens	%	N° de Imagens	%	N° de Imagens	%	
Amiloidose	31	8,3%	145	38,8%	96	25,7%	102	27,3%	374
Não Amiloidose	484	12,1%	2041	50,8%	260	6,5%	1230	30,6%	4015

Tabela 4.2: Quantidade de Imagens por Lesão e Corante presentes no *Dataset Baseline*.

Classe	Lesão	AZAN		HE		PAS		PAMS		Total
		N° de Imagens	%	N° de Imagens	%	N° de Imagens	%	N° de Imagens	%	
Positiva	Amiloidose	31	8,3%	145	38,8%	96	25,7%	102	27,3%	374
	Normal	716	17,8%	97	2,4%	156	3,9%	249	6,2%	1218
Negativa	Esclerose	672	16,7%	234	5,8%	472	11,8%	104	2,6%	1482
	Hiper celularidade	653	16,3%	93	2,3%	345	8,6%	0	-	1091
	Hiper celularidade Pura	0	-	60	1,5%	164	4,1%	0	-	224

Vale ressaltar que o projeto PathoSpotter, ao qual esse trabalho está vinculado, já recebeu aprovação dos Comitês de Ética para Pesquisa Envolvendo Seres Humanos do Instituto Gonçalo Moniz (FIOCRUZ), protocolos nº 188/09 e Nº 1.817.574, e da UEFS (Universidade Estadual de Feira de Santana) protocolo nº 2.637.620.

### 4.0.3 Modelos Utilizados

Para a construção dos modelos de classificação deste estudo e também para avaliar o ganho de desempenho obtido por meio da aplicação dos métodos destinados a lidar com o balanceamento de classes, foram utilizadas Redes Neurais Convolucionais (CNN) nas arquiteturas VGG16 e VGG19 (Simonyan e Zisserman, 2014), *Xception* (Chollet, 2017), *Inception* e *Inception-Resnet* (Szegedy et al., 2016), por serem modelos já bem consolidados na literatura. Para isso, foi realizado *transfer learning* em cada uma das redes, treinadas originalmente para resolver o desafio ImageNet (Russakovsky et al., 2015), adaptando-as ao problema em questão. O processo de *transfer learning* consistiu no congelamento do extrator de características, na remoção do topo das redes e na adição de um novo classificador ao topo dos modelos, ajustado para o problema de classificação binária.

### 4.0.4 Estratégias para lidar com o Desbalanceamento de Classes

Nessa sessão são apresentadas as estratégias para lidar com o desbalanceamento de classes presentes no problema estudado, são elas: *Random Oversampling*, *Random Undersampling* e Modelo *Ensemble-Based*.

#### *Random Oversampling e Random Undersampling*

Para a construção de um *dataset* balanceado foram aplicadas duas técnicas de tratamento de dados. São elas: Amostragem aleatória na classe majoritária, e a super amostragem para a classe minoritária.

Para a amostragem aleatória da classe majoritária, foi aplicado *Random Undersampling* na classe Não-Amiloidose com o objetivo de coletar uma quantidade de amostras com o mesmo número de imagens da classe Amiloidose. O *Random Undersampling* foi realizado utilizando a função *RandomUnderSampler* da biblioteca *imbalanced-learn* (Lemaître et al., 2017) na versão 0.10.0, utilizando o parâmetros *sampling strategy* como *majority*, indicando para a função fazer amostragem da classe majoritária, levando em consideração o número de amostras da classe minoritária. A nova base de dados construída com a aplicação do *Random Undersampling* será chamada de *Dataset RUS*.

A segunda abordagem utilizada foi uma aumento na quantidade de amostras da classe minoritária, por meio da aplicação de *Random Oversampling* da classe Amiloidose, com o objetivo de coletar uma quantidade de amostras com o mesmo número de imagens da classe Não-Amiloidose. O *Oversampling* foi realizado utilizando a função *RandomOverSampler* da biblioteca *imbalanced-learn* (Lemaître et al., 2017) na

versão 0.10.0, utilizando o parâmetros *sampling strategy* como *minority*, indicando para a função fazer super amostragem da classe minoritária, levando em consideração o número de amostras da classe majoritária. A nova base de dados construída com a aplicação do *Randon Oversampling* será chamada de *Dataset ROS*.

Em ambas as abordagens de balanceamento, foram aplicadas estratificações por corante com o objetivo de preservar a distribuição da base de dados original. Para o *Dataset RUS* foi realizado também estratificação por lesão para a classe Não-Amiloidose. A Tabela 4.3 mostra a quantidade de imagens, dividida por corante e classe, para o Dataset RUS, enquanto que a Tabela 4.4 também mostra a quantidade de imagens, dividida por corante e classe, para o *Dataset ROS*. Para o Dataset RUS tem-se um total de 748 imagens, já para o *Dataset ROS* tem-se um total de 8030 imagens.

Tabela 4.3: Relação da Quantidade de Imagens por Corante para o Dataset RUS.

	AZAN	HE	PAS	PAMS	Total
Amiloidose	31	145	96	102	374
Não Amiloidose	31	145	96	102	374

Tabela 4.4: Relação da Quantidade de Imagens por Corante para o Dataset ROS.

	AZAN	HE	PAS	PAMS	Total
Amiloidose	484	2041	260	1230	4015
Não Amiloidose	484	2041	260	1230	4015

### Modelo *Ensemble-Based*

Para a construção do Modelo *Ensemble-Based* foram utilizados 5 arquiteturas Redes Neurais Convolucionais, no modelo de *bagging*: VGG16, VGG19, *Xception*, *Inception* e *Inception-Resnet*, treinados utilizando os *emphDatasets* Baseline, RUS e ROS. O resultado final da classificação do Modelo *Ensemble-Based* é determinado por meio da combinação dos resultados de cada um dos modelo base (votação simples), tendo como resultado final a classe que foi prevista com mais frequência pelos 5 modelos.

A Figura 4.4 demonstra a arquitetura proposta para a construção do modelo *Ensemble-Based* para classificação de amiloidose renal.

### 4.0.5 Avaliação de Desempenho

Dado que o foco deste estudo está no problema de desbalanceamento de classe, as métricas de avaliação foram escolhidas de modo que a medida de desempenho não seja enviesada pela classe majoritária. Além disso, os modelos construídos foram treinados com o objetivo de reduzir a quantidade de classificações incorretas de imagens com amiloidose (classe positiva), uma vez que classificar uma imagem como Falso Negativo traz um maior potencial de risco ao paciente, já que pode direcionar

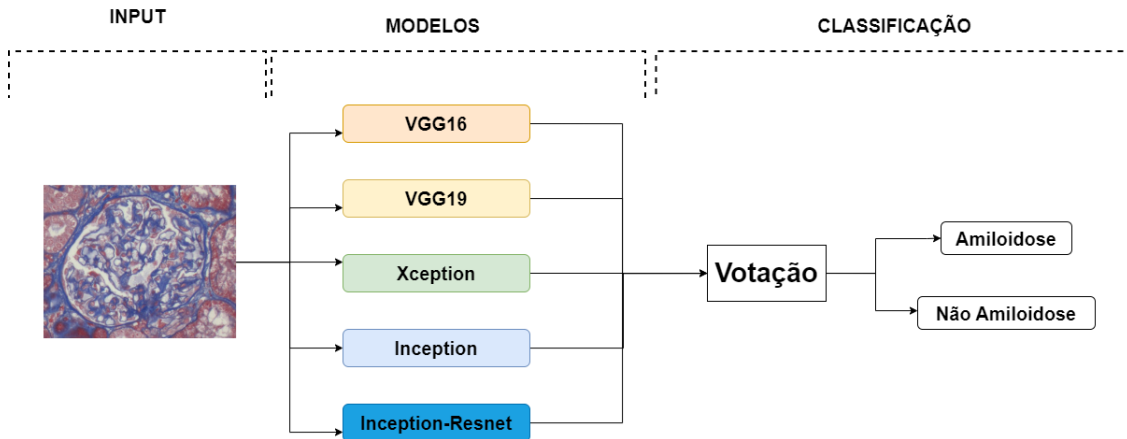


Figura 4.4: Arquitetura do Modelo *Ensemble-Based* Proposto.

para o não tratamento da doença, enquanto uma avaliação falso positiva é o ponto de partida para um curso de tratamento que será corroborado por outros exames.

A acurácia é uma métrica que trata ambas as classes como iguais, sendo assim, ela pode indicar que o modelo possui um elevado desempenho mesmo que ocorram erros de classificação para o grupo de imagens com amiloidose, que é a classe minoritária. Nesse sentido, para que seja possível ter uma avaliação que demonstre corretamente o impacto gerado por erros de classificação da classe minoritária, é necessário adicionar outras medidas de desempenho, além da acurácia, tais como precisão, especificidade, sensibilidade, F1-Score, G-Score e *AUC-ROC*. Além disso, uma vez que reduzir a quantidade de classificações incorretas de imagens com amiloidose é importante, será dada uma atenção especial à sensibilidade.

#### 4.0.6 Treinamento, Validação e Teste

Para realizar testes nos modelos, após a finalização do processo de treinamento, foi utilizado 25% das imagens presentes na base de dados. Estas imagens foram separadas dos dados originais para a criação de uma base de dados de teste, não sendo usadas para o treinamento de nenhum dos modelos. A Tabela 4.2 mostra a quantidade de imagens, dividida por corante e lesão. Dado que o problema estudado tem como uma forte característica o desbalanceamento de classes, a base de teste utilizada para testar os modelos após o treinamento foi mantida desbalanceada, mesmo para os experimentos nos quais os modelos serão treinados com as bases geradas após a aplicação de balanceamento de dados.

Tabela 4.5: Quantidade de Imagens por Lesão e Corante presentes na base de dados de teste.

Classe	Lesão	HE	AZAN	PAS	PAMS	Total
Positiva	Amiloidose	37	4	23	30	94
Negativa	Normal	176	22	54	34	286
	Esclerose	161	60	123	27	371
	Hiper celularidade	170	21	97	0	288
	Hiper celularidade Pura	0	11	48	0	59

Os demais 75% da base de dados foram utilizados para treinamento e validação dos modelos, utilizando validação cruzada de *5-folds*, com a proporção 60/15, ou seja, 60% da base de dados será utilizada para treinamento dos modelos e 15% para validação. Para ter uma maior confiança no resultado de avaliação dos modelos, todos eles foram treinados e validados com exatamente os mesmos dados.

# Capítulo 5

## Resultados

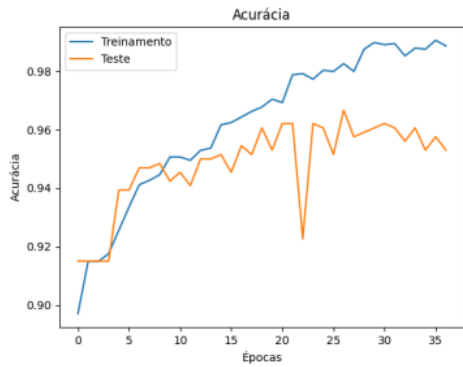
Nesta seção, são apresentados os resultados obtidos neste trabalho, os quais envolvem a análise de diversos modelos de redes neurais convolucionais, treinados no *Dataset Baseline*, caracterizado por desbalanceamento severo, e na mesma base de dados submetida a técnicas de balanceamento de classes, como *Dataset RUS* e *Dataset ROS*.

### 5.0.1 Treinamento dos Modelos utilizando o *Dataset Baseline*

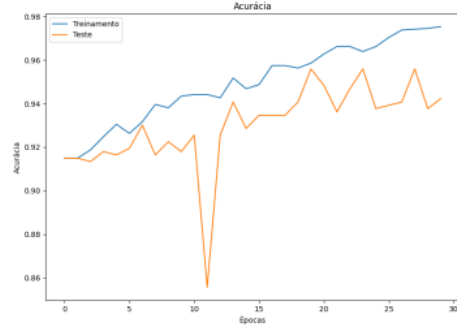
Na Figura 5.1 pode-se observar os gráficos de acurácia de treinamento e validação durante o processo de treinamento dos melhores modelos VGG-16, VGG-19, *Xception*, *Inception* e *Inception-Resnet*, dentre os treinados para os 5-*folds* da validação cruzada, utilizando a base de dados *Dataset Baseline*. Pode-se observar que todos os modelos obtiveram uma taxa de acertos elevada, tanto para a base de dados de treinamento quanto para a base de dados de validação.

Ao analisar as curvas de acurácia (Figura 5.1), observa-se que todos os modelos convergiram para uma acurácia máxima durante o processo de validação por volta da época 20, não havendo ganhos adicionais de desempenho posteriormente. Nota-se também que a curva de acurácia de validação dos modelos convergiu para um valor de aproximadamente 95% de acurácia, indicando que eles são capazes de generalizar o aprendizado obtido durante o treinamento, a fim de alcançar uma taxa de acurácia semelhante quando aplicados à base de dados de validação.

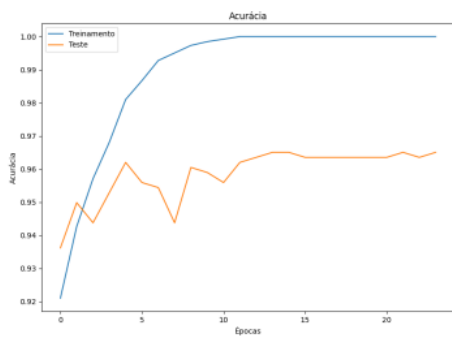
Apesar disso, para este experimento não se pode descartar um viés de aprendizado para a classe Não-Amiloidose, já que devido a sua preponderância em relação a classe Amiloidose, consequência do forte desbalanceamento de dados, os erros de classificação para a classe minoritária influenciam pouco na medida de avaliação de acurácia. Assim, ainda que a classe amiloidose possuísse mais erros de classificação, esses erros são pouco representativos na medida final da acurácia, uma vez que a classe majoritária tenha apresentado um desempenho de classificação elevado. Por conta disso, para a fase de teste dos modelos, serão incluídas outras métricas de avaliação, como



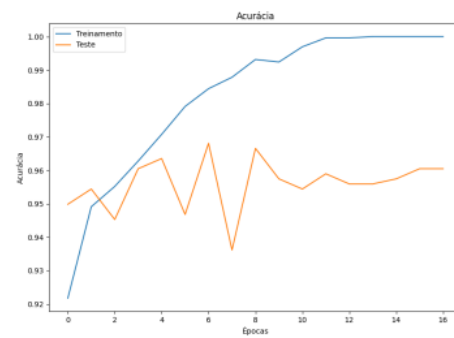
(a) Curva de Acurácia de Treinamento e Validação para o modelo VGG-16



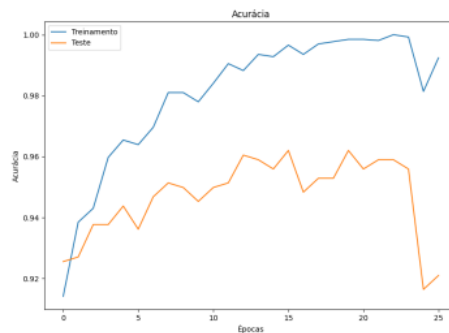
(b) [Curva de Acurácia de Treinamento e Validação para o modelo VGG-19



(c) [Curva de Acurácia de Treinamento e Validação para o modelo Xception



(d) [Curva de Acurácia de Treinamento e Validação para o modelo Inception



(e) [Curva de Acurácia de Treinamento e Validação para o modelo Inception-Resnet

Figura 5.1: Acurácia de Treinamento e Validação para o *Fold* com melhor Sensibilidade.



sensibilidade, *F1-Score* e *G-Score*, por serem mais apropriadas para avaliar modelos treinados com uma base de dados com situações de desbalanceamento de classe.

### 5.0.2 Testes dos Modelos utilizando o *Dataset Baseline*

Considerando a importância de reduzir a ocorrência de classificações incorretas em imagens com amiloidose (Falsos Negativos), dado que este tipo de erro é mais prejudicial para o paciente, os modelos estudados nesse trabalho serão avaliados com um peso adicional nos erros de Falsos Negativos, além das métricas padrões já discutidas. Com base nesse critério, foram definidos como os melhores modelos aqueles que obtiveram os melhores resultados na métricas de sensibilidade, quando avaliados na base de dados de teste.

Na Figura 5.2 pode-se observar as matrizes de confusão para os melhores resultados dos modelos VGG-16, VGG-19, *Xception*, *Inception* e *Inception-Resnet*, dentre os treinados para os 5-*fold* da validação cruzada.

Com base na Figura 5.2, nota-se que todos os modelos apresentaram uma baixa taxa de acerto para a classe positiva, enquanto que a classe negativa, com mais amostras, foi quase que completamente classificada corretamente. Estes dados expressam que o modelo está com um viés de classificação para a classe negativa, o que já era esperado, dado o forte desbalanceamento da base de dados *Dataset Baseline*.

Na Tabela 5.1 apresenta o número de erros de classificação, divididos por corante e lesão, para os modelos treinados com o *Dataset Baseline*. É possível perceber que dentre os corantes, o que mais apresentou erros de classificação foi o HE. Já entre as lesões presentes na classe Não-Amiloidose, a que apresentou mais erros de classificação foi a Esclerose.

A Tabela 5.7 apresenta a média dos resultados de teste dos modelos ao longo dos 5-*fold*, para as métricas Acurácia, Precisão, Sensibilidade, Especificidade, *AUC*, *F1-Score* e *G-Score*. Nela, observa-se que somente os modelos *Inception* e *Inception-Resnet* obtiverem resultados acima de 70% em termos de sensibilidade, com um pior resultado para o modelo VGG-19. Estes resultados são consequência de um elevado número de classificações incorretas para a classe positiva, resultando em um baixo desempenho de sensibilidade. Nota-se também que todos os modelos apresentam uma elevada acurácia, acima de 95%. Este acontecimento é consequência do elevado desbalanceamento de classes presente no problema estudado, reforçando que a métrica acurácia não é confiável para avaliar o desempenho do problema em estudo.

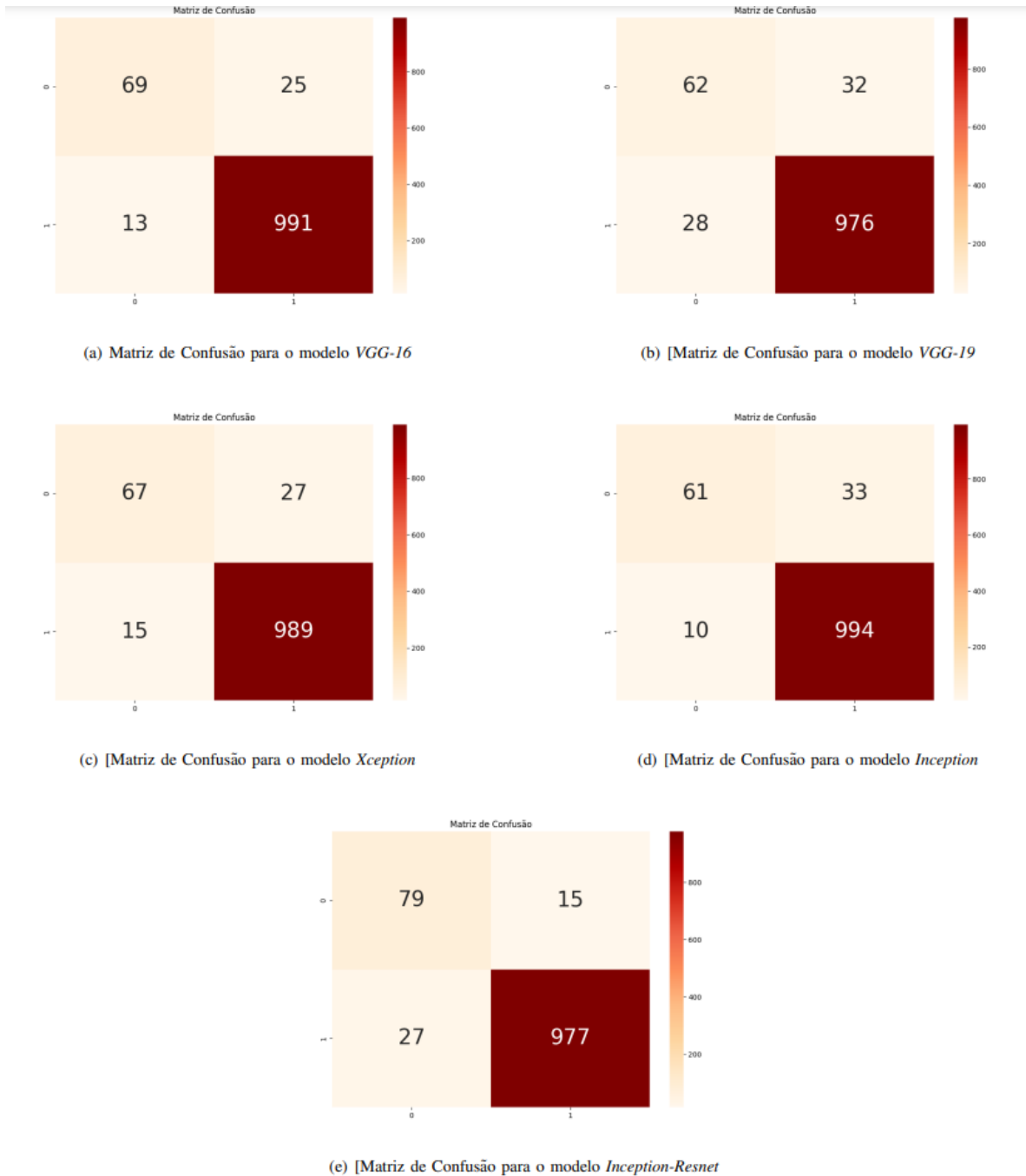


Figura 5.2: Matriz de Confusão para os melhores modelos treinados com o *Dataset Baseline*.

Tabela 5.1: Erros de Classificação por para os Melhores Modelos Treinados com o *Dataset Baseline*.

Tabela 5.2: Erros de classificação para o modelo VGG16

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	10,6%	1,1%	8,5%	5,3%
Negativa	Falso Positivo	Normal	0,3%	0,1%	0%	0,2%
		Esclerose	0,1%	0,3%	0,4%	0,3%
		Hiper celularidade	1%	0%	0%	-
		Hiper celulari. Pura	-	0%	0,1%	-

Tabela 5.3: Erros de classificação para o modelo VGG19.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	13,8%	1,1%	11,7%	7,4%
Negativa	Falso Positivo	Normal	0,4%	0,1%	0,2%	0,4%
		Esclerose	0,3%	0,1%	0,3%	0,3%
		Hiper celularidade	0,4%	0,0%	0,2%	-
		Hiper celulari. Pura	-	0,1%	0,0%	-

Tabela 5.4: Erros de classificação para o modelo *Xception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	16,0%	2,1%	7,4%	3,2%
Negativa	Falso Positivo	Normal	0	0	2	0
		Esclerose	0,2%	0,2%	0,5%	0,2%
		Hiper celularidade	0,0%	0,0%	0,1%	-
		Hiper celulari. Pura	-	0,1%	0,0%	-

Tabela 5.5: Erros de classificação para o modelo *Inception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	13,8%	1,1%	6,4%	5,3%
Negativa	Falso Positivo	Normal	1	1	1	0
		Esclerose	0,1%	0,3%	0,5%	0,0%
		Hiper celularidade	0,1%	0,0%	0,0%	-
		Hiper celulari. Pura	-	0,0%	0,0%	-

Tabela 5.6: Erros de classificação para o modelo *Inception-Resnet*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	4,3%	0%	3,2%	8,5%
Negativa	Falso Positivo	Normal	6	2	1	0
		Esclerose	0,2%	0,4%	0,6%	0,1%
		Hiper celularidade	0,0%	0,1%	0,2%	-
		Hiper celulari. Pura	-	0,1%	0,1%	-

Tabela 5.7: Média das Métricas de desempenho para o *Dataset Baseline*.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	95,6% ( $\pm 0,03$ )	81,4% ( $\pm 0,09$ )	65,1% ( $\pm 0,08$ )	98,4% ( $\pm 0,01$ )	71,5% ( $\pm 0,01$ )	79,9% ( $\pm 0,04$ )	81,8% ( $\pm 0,03$ )
VGG-19	94,0% ( $\pm 0,01$ )	70,6% ( $\pm 0,07$ )	54,5% ( $\pm 0,12$ )	97,6% ( $\pm 0,01$ )	60,2% ( $\pm 0,05$ )	72,5% ( $\pm 0,07$ )	76,1% ( $\pm 0,053$ )
Xception	95,8% ( $\pm 0,01$ )	82,9% ( $\pm 0,02$ )	63,8% ( $\pm 0,02$ )	98,8% ( $\pm 0,01$ )	72,1% ( $\pm 0,02$ )	79,4% ( $\pm 0,01$ )	81,3% ( $\pm 0,01$ )
Inception	96,6% ( $\pm 0,01$ )	86,9% ( $\pm 0,01$ )	71,5% ( $\pm 0,01$ )	98,9% ( $\pm 0,01$ )	78,4% ( $\pm 0,01$ )	84,1% ( $\pm 0,01$ )	85,2% ( $\pm 0,01$ )
InceptionResnet	96,1% ( $\pm 0,02$ )	83,1% ( $\pm 0,17$ )	77,2% ( $\pm 0,08$ )	97,9% ( $\pm 0,03$ )	78,4% ( $\pm 0,06$ )	86,8% ( $\pm 0,03$ )	87,6% ( $\pm 0,02$ )

Tabela 5.8: Intervalo de Confiança utilizando o *Dataset Baseline*.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	94,91% - 96,24%	65,05% - 97,72%	50,22% - 79,99%	96,34% - 100%	69,36% - 73,63%	71,77% - 88,02%	75,36% - 88,18%
VGG-19	92,91% - 94,99%	56,79% - 84,49%	33,52% - 75,42%	94,97% - 100%	51,11% - 69,29%	59,3% - 85,75%	66,81% - 85,3%
Xception	95,16% - 96,39%	78,5% - 87,25%	58,9% - 68,76%	98,45% - 99,08%	67,71% - 76,49%	76,24% - 82,53%	78,75% - 83,84%
Inception	96,4% - 96,86%	83,5% - 90,26%	68,15% - 74,83%	98,65% - 99,32%	76,83% - 79,98%	82,25% - 85,98%	83,68% - 86,79%
InceptionResnet	92,22% - 100%	52,91% - 100%	62,81% - 91,66%	92,34% - 100%	66,45% - 90,36%	81,38% - 92,21%	82,95% - 92,18%

Ao observar as métrica *F1-Score*, mostrada na Tabela 5.7, observa-se que os modelos *Inception* e *Inception-Resnet* apresentaram um melhor desempenho. Apesar disso, o desempenho geral dos modelos ainda é baixo, com um valor inferior a 80% de *F1-Score* para todos os casos. Estes resultados são consequência de um baixo valor da sensibilidade, demonstrando que todos os modelos ainda realizam uma quantidade elevada de classificações enganosas para a classe positiva. Por outro lado, ao se realizar uma análise em termos de *G-Score*, percebe-se que o modelo *Inception-Resnet* foi o que obteve o maior desempenho, com um valor de aproximadamente 87%, reforçando que este modelo apresentou um melhor resultado de classificações corretas para ambas as classes.

A Tabela 5.8 mostra os intervalos de confiança, com um nível de confiança de 95%, para as métricas analisadas, usando o *dataset* desbalanceado para o treinamento dos modelos. É possível visualizar que o modelo *Inception-Resnet* foi o que apresentou um intervalo de confiança mais amplo, seguido pelo modelo VGG19. Dentre os 5 modelos investigados, o que apresentou um melhor desempenho no intervalo de confiança para sensibilidade foi o *Inception*, com um limite inferior de 83,5%. Ao analisar as métricas acurácia e especificidade, nota-se que ambas apresentam um desempenho elevado, devido a elevada eficiência de todos os modelos para classificar corretamente imagens da classe Não-Amiloidose.

Com base nos dados mostrados na Matriz de Confusão (Figura 5.2), bem como das médias das métricas extraídas da mesma e seus intervalos de confiança (Tabelas 5.7 e 5.8), percebe-se que os modelos *InceptionResnet* e *Inception* foram os que apresentaram melhor desempenho entre os 5 estudados. Apesar disso, todos os modelos apresentam um baixo desempenho médio para a métrica sensibilidade, indicando que eles apresentaram uma taxa elevada de Falsos Negativos.

Como já foi apresentado, os modelos construídos para o problema de amiloidose, para serem considerados satisfatórios, é crucial que demonstrem uma taxa reduzida de Falso Negativos. Sendo assim, percebe-se que os 5 modelos, quando treinados em uma base de dados desbalanceada, não apresentam resultados adequados ao ponto de serem confiáveis nas classificações de imagens que possuem amiloidose, demonstrando a necessidade de estudos mais aprofundados sobre o problema.

### 5.0.3 Avaliação dos Modelos após Realização de Balanceamento de Dados por meio de *Random Undersampling*

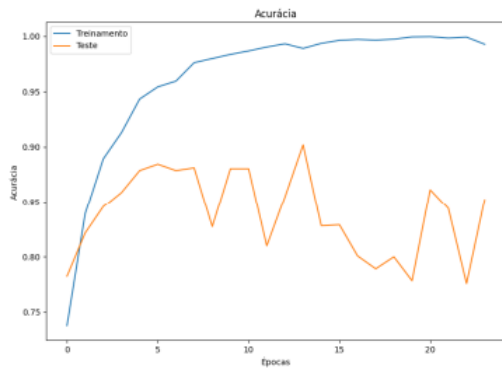
Na Figura 5.3 observa-se os gráficos de acurácia de treinamento e validação dos melhores modelos, dentre os treinados para os 5-*fold* da validação cruzada, utilizando a base de dados *Dataset RUS*. Em termos de acurácia, nota-se que todos os modelos obtiveram uma taxa de acertos superior a 90% para a base de dados de validação. Observa-se que a acurácia de validação dos modelos treinados com o *Dataset RUS* é inferior ao dos modelos treinados com o *Dataset Baseline*. Este acontecimento é consequência do processo de balanceamento de dados por amostragem aleatória, que reduziu a quantidade de amostras para treinamento e aprendizado dos modelos bem como equiparou a importância dos erros de classificação para ambas as classes.

É importante destacar que, apesar dos modelos *Xception*, *Inception* e *Inception-Resnet* serem mais suscetíveis a *overfitting* quando treinados em uma base de dados reduzida, devido aos mesmos serem mais complexos em termos de número de parâmetros de treinamento (Szegedy et al., 2016)(Chollet, 2017), essa ocorrência não se manifestou, visto que os valores de acurácia de treinamento e validação vistos Figura 5.3 são próximos. Assim, a não ocorrência de ajuste excessivo aos dados de treinamento, mesmo alcançando uma acurácia próxima de 100%, é um indicativo de que a base de treinamentos possui uma qualidade de amostras que permite a todos os modelos aprenderem corretamente os padrões necessários para uma classificação correta de ambas as classes.

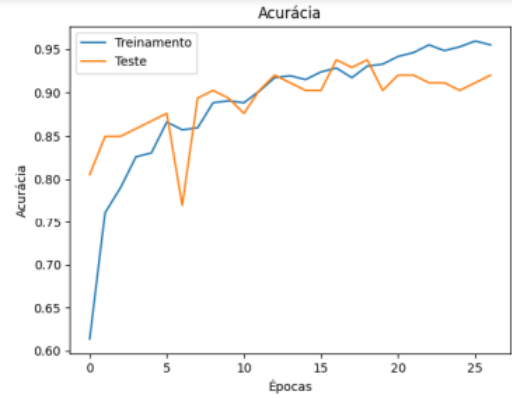
### 5.0.4 Testes dos Modelos após Realização de Balanceamento de Dados por meio de *Random Undersampling*

Na Figura 5.4 pode-se observar as matrizes de confusão para os melhores resultados dos modelos VGG-16, VGG-19, *Xception*, *Inception* e *Inception-Resnet*, dentre os treinados para os 5-*folds* da validação cruzada, com a base de dados *Dataset RUS*. O critério de escolha dos melhores modelos foi o mesmo utilizado nos experimentos utilizando a base de dados *Dataset Baseline*, ou seja, aqueles que obtiveram os melhores resultados na métricas de sensibilidade, quando avaliados na base de dados de teste.

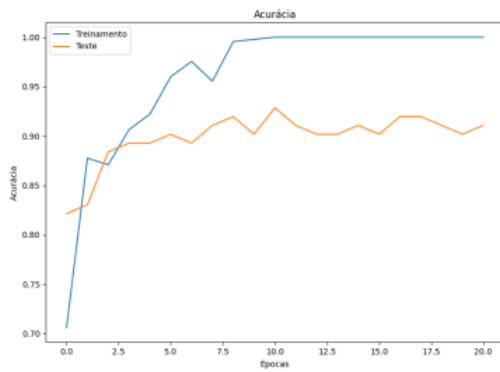
Com base na Matriz de Confusão, observa-se que todos os modelos obtiveram uma elevada taxa de acertos para a classe Amiloidose e dentre eles, o *Inception* foi o que obteve o menor número de Falsos Negativos, seguido pelo *Inception-Resnet* e *Xception*. Apesar disso, quando comparados com os resultados dos modelos treinado com o *Dataset Baseline*, nota-se que houve uma elevação significativa no número de amostras da classe Não-Amiloidose classificadas incorretamente, devido a aplicação do método *Random Unversample*, o que fez com que os modelos passassem a contar com menos amostras da classe Não-Amiloidose para treinamento, tendo como consequência um aprendizado fraco para mesma, resultando em um modelo pouco preciso e um elevado número de Falsos Positivos.



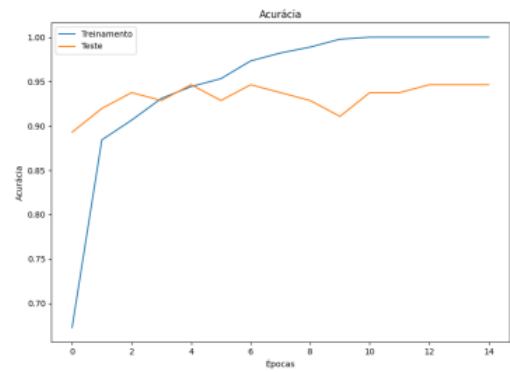
(a) Curva de Acurácia de Treinamento e Validação para o modelo *VGG-16*



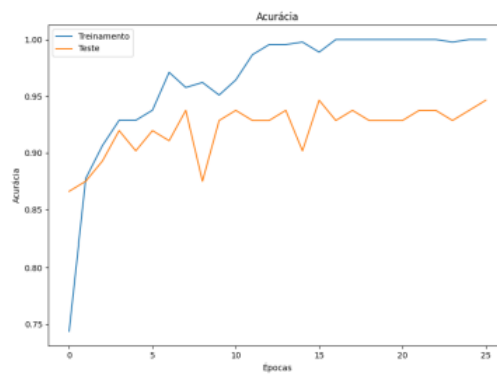
(b) Curva de Acurácia de Treinamento e Validação para o modelo *VGG-19*



(c) Curva de Acurácia de Treinamento e Validação para o modelo *Xception*



(d) Curva de Acurácia de Treinamento e Validação para o modelo *Inception*



(e) Curva de Acurácia de Treinamento e Validação para o modelo *Inception-Resnet*

Figura 5.3: Acurácia de Treinamento e Validação para os melhores Modelos.

Na Tabela 5.9 apresenta o número de erros de classificação, divididos por corante e lesão, para os modelos mesmos modelos discutidos na Figura 5.4. Quando se compara os resultados de classificação dos modelos treinados com o *Dataset Baseline* e o *Dataset RUS*, nota-se que houve uma melhoria no número de classificações incorretas por corante para a lesão amiloidose, especialmente para os corantes AZAN e PAS. Por outro lado, para as lesões presentes na classe Não-Amiloidose, percebe-se que houve um menor desempenho de classificação, já que o número de classificações incorretas aumentou para todas as lesões.

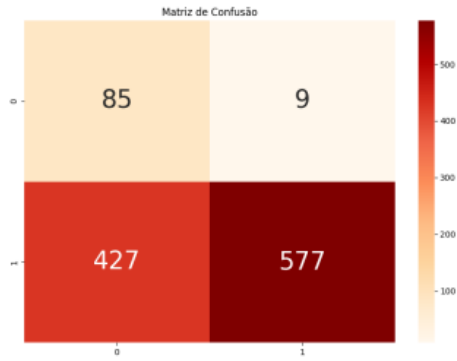
A Tabela 5.15 apresenta a média dos resultados de teste dos modelos para os 5-*folds*, treinados utilizando a base de dados *Dataset RUS*, para as métricas Acurácia, Precisão, Sensibilidade, Especificidade, *AUC*, *F1-Score* e *G-Score*. Nota-se que todos os modelos obtiveram valores elevados para a métrica sensibilidade, com ênfase para os modelos *Xception*, *Inception* e o *Inception-Resnet*. Apesar disso, nota-se que todos os modelos obtiveram valores de precisão muito baixos, indicando que todos eles apresentaram uma quantidade elevada de Falsos Positivos durante a fase de testes.

No que diz respeito a métrica *G-Score*, percebe-se que os modelos *Inception* e o *Inception-Resnet* obtiveram melhor desempenho, apontando que estes modelos apresentaram um aprimoramento nas classificações corretas para ambas as classes, se comparado com os modelos VGG e o *Xception*. Entretanto, devido a um baixo valor da métrica precisão, observa-se um baixo desempenho de *F1-Score* para todos os modelos, reforçando seu baixo desempenho de classificação para a classe Não-Amiloidose.

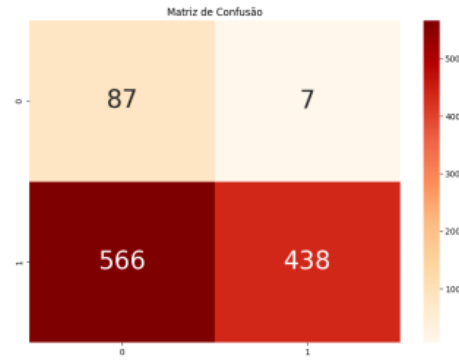
Ao se analisar a média da métrica *AUC* dos modelos, comparando-as com os modelos treinados anteriormente com o *Dataset Baseline*, nota-se que não houve um ganho de desempenho significativo. Isso aconteceu pois mesmo os modelos tendo demonstrado melhorias no sentido de reduzir a quantidade de erros de Falso Negativo, houve uma elevação nos erros de Falso Positivo, o que impediu a elevação dos resultados das métricas.

A Tabela 5.16 mostra os intervalos de confiança das métricas de avaliação utilizadas. É observado que para a métrica sensibilidade houve uma melhoria em termos de elevação dos limites inferiores e superiores de todos os modelos, quando comparado com os resultados de linha de base, reforçando a confiança nos modelos em termos de classificação correta da classe positiva. Nota-se também que além da elevação nos números de intervalo de confiança, os valores dos limites superiores e inferiores estão mais próximos, o que demonstra uma maior estabilidade nos resultados de classificação dos diferentes *folds*.

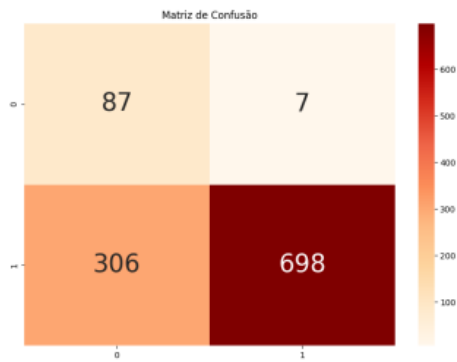




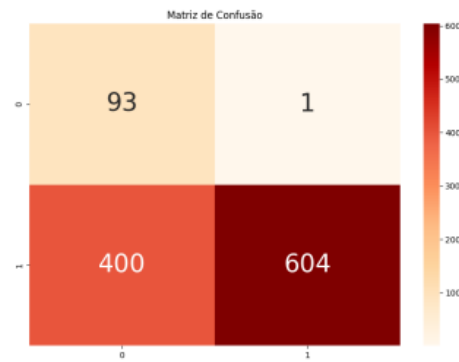
(a) Matriz de Confusão para o modelo *VGG-16*



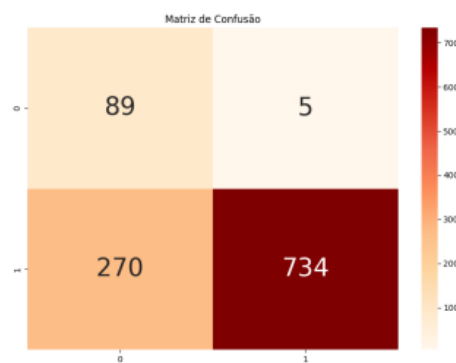
(b) [Matriz de Confusão para o modelo *VGG-19*



(c) [Matriz de Confusão para o modelo *Xception*



(d) [Matriz de Confusão para o modelo *Inception*



(e) [Matriz de Confusão para o modelo *Inception-Resnet*

Figura 5.4: Matriz de Confusão para os melhores modelos treinados com o *Dataset RUS*.

Tabela 5.9: Erros de Classificação por para os Melhores Modelos Treinados com o *Dataset RUS*.

Tabela 5.10: Erros de classificação para o modelo VGG16.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	3,2%	0,0%	1,1%	5,3%
Negativa	Falso Positivo	Normal	6,6%	1,0%	2,8%	0,8%
		Esclerose	5,0%	3,8%	6,0%	1,0%
		Hiper celularidade	5,7%	1,3%	4,7%	-
		Hiper celulari. Pura	-	0,6%	3,4%	-

Tabela 5.11: Erros de classificação para o modelo VGG19.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	3,2%	0,0%	1,1%	5,3%
Negativa	Falso Positivo	Normal	9,0%	1,1%	3,5%	1,5%
		Esclerose	7,1%	4,2%	7,8%	1,3%
		Hiper celularidade	8,3%	1,2%	6,7%	-
		Hiper celulari. Pura	-	0,8%	4,1%	-

Tabela 5.12: Erros de classificação para o modelo *Xception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	4,3%	0,0%	0,0%	3,2%
Negativa	Falso Positivo	Normal	4,3%	1,0%	0,9%	0,5%
		Esclerose	2,2%	3,7%	5,3%	0,3%
		Hiper celularidade	3,1%	0,9%	4,5%	-
		Hiper celulari. Pura	-	0,9%	2,0%	-

Tabela 5.13: Erros de classificação para o modelo *Inception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	0,0%	0,0%	0,0%	1,1%
Negativa	Falso Positivo	Normal	4,6%	0,6%	2,1%	0,3%
		Esclerose	3,9%	4,4%	6,7%	0,8%
		Hiper celularidade	5,8%	1,1%	6,1%	-
		Hiper celulari. Pura	-	0,8%	2,8%	-

Tabela 5.14: Erros de classificação para o modelo *Inception-Resnet*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	2,1%	0,0%	0,0%	3,2%
Negativa	Falso Positivo	Normal	2,5	0,3	1,1	0,3
		Esclerose	2,7%	3,0%	5,2%	0,8%
		Hiper celularidade	3,3%	0,8%	3,9%	-
		Hiper celulari. Pura	-	0,7%	2,3%	-

Tabela 5.15: Média das Métricas de desempenho Utilizando o *Dataset* RUS.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	63,6% ( $\pm 0,04$ )	17,4% ( $\pm 0,01$ )	85,1% ( $\pm 0,04$ )	61,6% ( $\pm 0,05$ )	28,8% ( $\pm 0,02$ )	72,3% ( $\pm 0,02$ )	73,4% ( $\pm 0,02$ )
VGG-19	60,9% ( $\pm 0,08$ )	16,7% ( $\pm 0,02$ )	86,6% ( $\pm 0,04$ )	58,5% ( $\pm 0,09$ )	27,9% ( $\pm 0,03$ )	70,9% ( $\pm 0,04$ )	72,6% ( $\pm 0,02$ )
Xception	69% ( $\pm 0,03$ )	20,8% ( $\pm 0,01$ )	92,3% ( $\pm 0,01$ )	66,8% ( $\pm 0,03$ )	33,9% ( $\pm 0,02$ )	78,5% ( $\pm 0,02$ )	79,6% ( $\pm 0,01$ )
Inception	68,8% ( $\pm 0,03$ )	21,2% ( $\pm 0,01$ )	96,8% ( $\pm 0,01$ )	66,2% ( $\pm 0,03$ )	34,8% ( $\pm 0,02$ )	80,0% ( $\pm 0,02$ )	81,5% ( $\pm 0,01$ )
InceptionResnet	76,3% ( $\pm 0,02$ )	25,8% ( $\pm 0,02$ )	93,0% ( $\pm 0,01$ )	74,7% ( $\pm 0,03$ )	40,3% ( $\pm 0,02$ )	83,3% ( $\pm 0,01$ )	83,8% ( $\pm 0,01$ )

Tabela 5.16: Intervalo de confiança utilizando o *Dataset* RUS.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	55,21% - 72,04%	14,15% - 20,57%	78,0% - 92,21%	51,96% - 71,26%	24,6% - 32,96%	68,25% - 76,35%	69,9% - 76,82%
VGG-19	46,42% - 75,41%	12,68% - 20,73%	79,18% - 94,01%	41,99% - 75,03%	22,54% - 33,27%	63,16% - 78,55%	67,77% - 77,33%
Xception	63,29% - 74,64%	17,85% - 23,67%	90,78% - 93,9%	60,5% - 73,04%	30,01% - 37,74%	75,03% - 81,96%	76,68% - 82,43%
Inception	62,82% - 74,74%	18,34% - 24,15%	94,52% - 99,09%	59,5% - 72,81%	30,94% - 38,68%	76,44% - 83,54%	78,76% - 84,21%
InceptionResnet	71,22% - 81,31%	21,45% - 30,15%	90,15% - 95,81%	68,94% - 80,46%	35,41% - 45,24%	81,29% - 85,33%	82,19% - 85,49%

### 5.0.5 Avaliação dos Modelos após Realização de Balanceamento de Dados por meio de *Random Oversampling*

Na Figura 5.5 pode-se observar os gráficos de acurácia de treinamento e validação, durante o processo de treinamento dos modelos, dentre os treinados para os *5-folds* da validação cruzada, utilizando o *Dataset ROS*.

Ao observar a curva de acurácia de treinamento dos modelos, nota-se que todos os modelos convergiram para 100% de acurácia. Entretanto, ao analisar as curvas de acurácia de validação, nota-se que somente nos modelos VGG-16, VGG-19 e *Inception-Resnet* obtiveram uma acurácia de 90%, ou superior, indicando que estes modelos conseguem generalizar melhor o conhecimento adquirido para obter uma taxa de acerto elevada nos dados de validação. Já os modelos *Xception* e *Inception* foram mais sensíveis a *overfitting*, uma vez que não conseguiram obter taxas de acerto elevadas em termos de acurácia na etapa de validação.

### 5.0.6 Testes dos Modelos após Realização de Balanceamento de Dados por meio de *Random Oversampling*

Na Figura 5.6 pode-se observar as matrizes de confusão para os melhores resultados dos modelos VGG-16, VGG-19, *Xception*, *Inception* e *Inception-Resnet*, dentre os treinados para os *5-folds* da validação cruzada utilizando base de dados *Dataset ROS*. O critério de escolha dos melhores modelos foi o mesmo utilizado no experimento com a base de dados *Dataset Baseline*, ou seja, aqueles que obtiveram os melhores resultados na métricas de sensibilidade, quando avaliados na base de dados de teste.

Com base na Matriz de Confusão, observa-se que o modelo *Inception* foi o que obteve uma maior taxa de acerto para a classe positiva, ou seja, a classe de imagens com amiloidose. Quando se analisa a taxa de acerto para a classe negativa, nota-se que o modelo *Inception-Resnet* foi o que obteve um maior desempenho, apresentando 48 Falsos Positivo. Apesar de maiores taxas de Falsos Negativo ser mais danoso do que de Falso Positivo, é importante ressaltar que este cenário também não deve ser desejado, uma vez que pode ocasionar erros no auxílio ao diagnóstico. Nesse sentido, dado que a diferença de Verdadeiros Positivos entre os modelos *Inception* e *Inception-Resnet* é de apenas 4 amostras, pode-se considerar que o modelos *Inception* apresentou uma maior estabilidade na classificação correta de ambas as classes.

Na Tabela 5.17 apresenta o número de erros de classificação, divididos por corante e lesão, para os mesmos modelos discutidos na Figura 5.6. Quando se compara os resultados de classificação dos modelos treinados com o *Dataset Baseline* e o *Dataset ROS*, nota-se que houve uma melhoria no número de classificação corretas para a lesão amiloidose, especialmente para os corantes AZAN e PAS. Por outro lado, para as lesões presentes na classe Não-Amiloidose, percebe-se que houve um decaimento no desempenho de classificação, já que o número de classificações incorretas aumentou para todas as lesões. Apesar disso, nota-se que para os modelos *Xception*

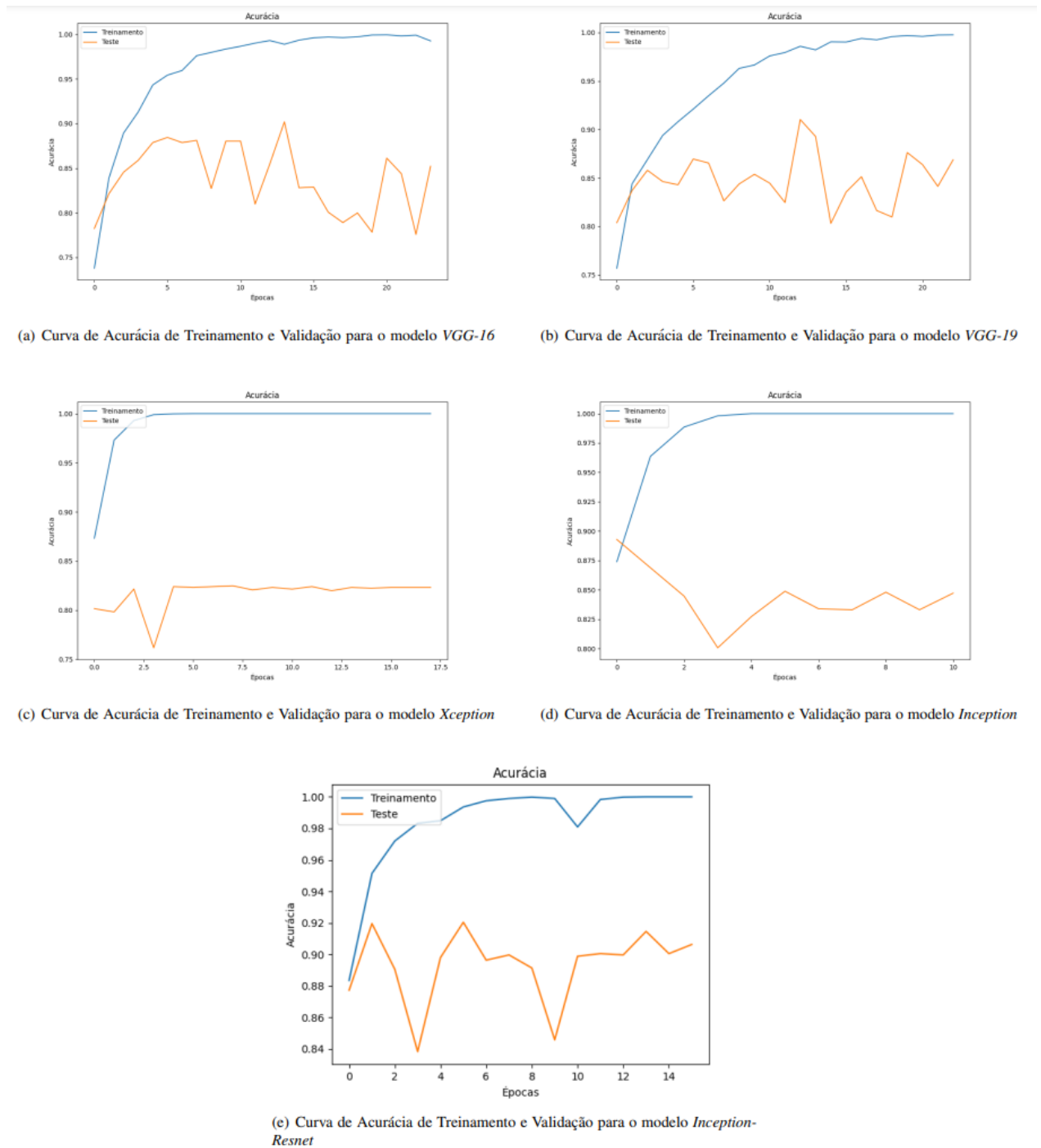


Figura 5.5: Acurácia de Treinamento e Validação para o *Fold* com melhor Sensibilidade

e *Inception-Resnet* a quantidade de classificações incorretas pode ser considerada baixa, como pode ser visualizada na Figura 5.6.

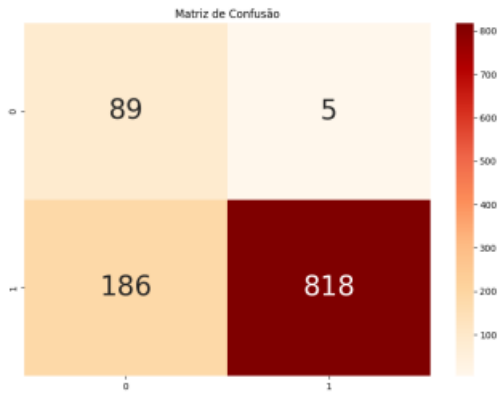
A Tabela 5.23 apresenta a média dos resultados de teste dos modelos para os 5-*folds* para as métricas Acurácia, Precisão, Sensibilidade, Especificidade, *AUC*, *F1-Score* e *G-Score*, utilizando a base de dados *Dataset ROS*. Ao analisa-se a métrica sensibilidade, pode-se perceber que, com exceção do modelo *Xception* todos os modelos obtiveram superior a 84%, com destaque para o modelo *Inception*, que obteve uma sensibilidade de 90,2%.

No que diz respeito a precisão, nota-se que os modelos *Xception* e *Inception-Resnet* foram os que obtiveram valores mais elevados. Apesar disso, é visto que o desempenho geral dos modelos para esta métrica ainda é insuficiente, com valores menores ou iguais a 70% de precisão. Entretanto, ao se analisar a métrica *G-Score*, nota-se que todos os modelos obtiveram resultados iguais ou superiores a 85%, demonstrando um bom balanço de desempenho para ambas as classes, uma vez que mesmo que classe Não-Amiloidose tenha apresentado mais erros de classificação que a classe Amiloidose, o número de amostras para a classe Negativa é proporcionalmente maior.

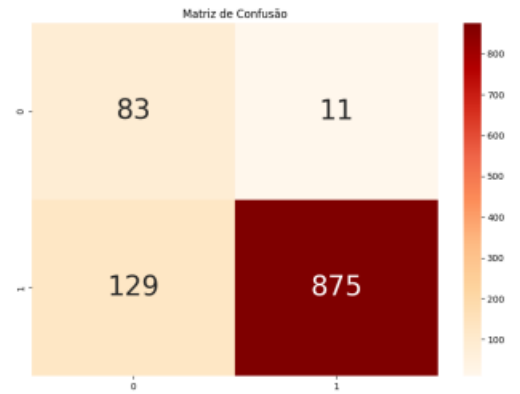
Ao se comparar a média da métrica *AUC* dos modelos treinados com o *Dataset ROS* com os treinados com o *Dataset Baseline*, nota-se que houveram melhorias para esta métrica em todos os modelos, consequência a elevação do desempenho da sensibilidade, apesar de ter havido perdas de desempenho para a precisão. Este acontecimento reforça que os modelos apresentaram um bom balanço de desempenho para ambas as classes.

Ao se realizar um comparativo com os resultados dos modelos treinados com o *Dataset Baseline*, nota-se que não somente houve uma melhoria de desempenho para todos os modelos, no sentido de melhor classificação de imagens com amiloidose, como também ocorreu uma melhoria no balanço de desempenho entre as classes, dado que houve uma elevação desempenho na métrica *G-Score*, ao mesmo tempo em que não houve uma queda significativa em termos de acurácia. Este acontecimento é positivo, no sentido de que apesar de que um modelo mais sensível ser mais vantajoso para o paciente, também é importante um bom desempenho de classificação para ambas as classes.

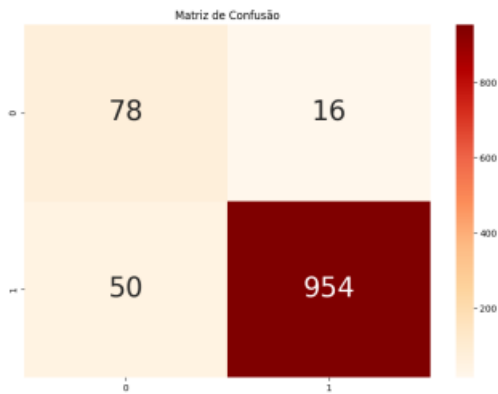
A Tabela 5.24 mostra os intervalos de confiança das métricas de avaliação de desempenho para modelos treinados com o *Dataset ROS*.



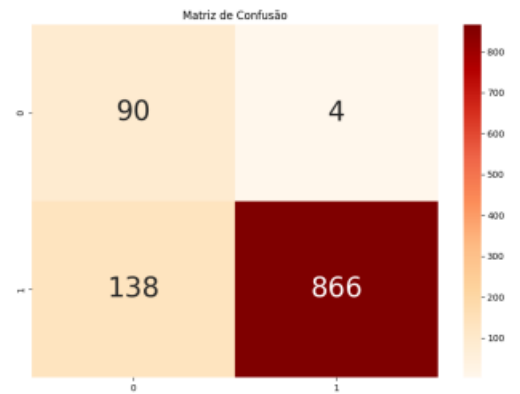
(a) Matriz de Confusão para o modelo *VGG-16*



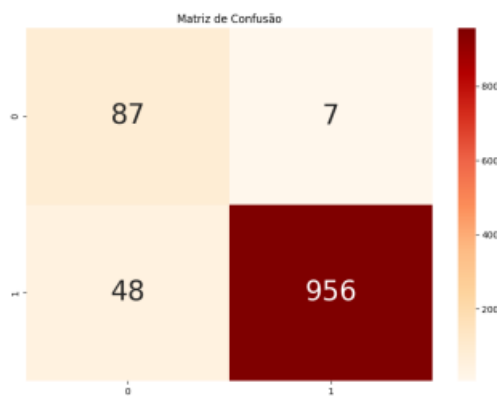
(b) [Matriz de Confusão para o modelo *VGG-19*



(c) [Matriz de Confusão para o modelo *Xception*



(d) [Matriz de Confusão para o modelo *Inception*



(e) [Matriz de Confusão para o modelo *Inception-Resnet*

Figura 5.6: Matriz de Confusão para os melhores modelos treinados com o *Dataset ROS*.

Tabela 5.17: Erros de Classificação por para os Melhores Modelos Treinados com o *Dataset ROS*.

Tabela 5.18: Erros de classificação para o modelo VGG16.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	2,1%	0,0%	0,0%	1,1%
Negativa	Falso Positivo	Normal	2,7%	3%	1,1%	0,9%
		Esclerose	1,3%	2,4%	3,3%	1,4%
		Hiper celularidade	2,1%	0,8%	1,1%	-
		Hiper celulari. Pura	-	0,3%	0,9%	-

Tabela 5.19: Erros de classificação para o modelo VGG19.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	7,4%	0,0%	3,2%	1,1%
Negativa	Falso Positivo	Normal	1,8%	0,1%	0,3%	0,1%
		Esclerose	1,5%	1,5%	2,1%	1,1%
		Hiper celularidade	2,1%	0,5%	1,4%	-
		Hiper celulari. Pura	-	0,2%	1,2%	-

Tabela 5.20: Erros de classificação para o modelo *Xception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	8,5%	2,1%	3,2%	3,2%
Negativa	Falso Positivo	Normal	0,1%	0,0%	0,7%	0,4%
		Esclerose	0,4%	0,9%	1,3%	0,3%
		Hiper celularidade	0,2%	0,1%	0,4%	-
		Hiper celulari. Pura	-	0,2%	0,0%	-

Tabela 5.21: Erros de classificação para o modelo *Inception*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	1,0%	0,0%	3,2%	0,0%
Negativa	Falso Positivo	Normal	1,4%	0,2%	1,0%	1,1%
		Esclerose	1,0%	2,9%	2,0%	1,3%
		Hiper celularidade	1,2%	0,3%	0,9%	-
		Hiper celulari. Pura	-	0,2%	0,3%	-

Tabela 5.22: Erros de classificação para o modelo *Inception-Resnet*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	2,1%	0,0%	3,2%	2,1%
Negativa	Falso Positivo	Normal	1,1	0,2	0,3	0,2
		Esclerose	0,3%	1,1%	0,5%	0,4%
		Hiper celularidade	0,0%	0,1%	0,4%	-
		Hiper celulari. Pura	-	0,1%	2,1%	-



Tabela 5.23: Média das Métricas de desempenho Utilizando o *dataset* ROS.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	89,3% ( $\pm 0,04$ )	46,9% ( $\pm 0,12$ )	86,6% ( $\pm 0,05$ )	89,6% ( $\pm 0,05$ )	59,8% ( $\pm 0,09$ )	88,0% ( $\pm 0,01$ )	88,1% ( $\pm 0,01$ )
VGG-19	87,6% ( $\pm 0,02$ )	40,3% ( $\pm 0,05$ )	84,3% ( $\pm 0,05$ )	87,9% ( $\pm 0,03$ )	54,2% ( $\pm 0,03$ )	86,0% ( $\pm 0,01$ )	86,1% ( $\pm 0,01$ )
Xception	95,1% ( $\pm 0,01$ )	70,8% ( $\pm 0,07$ )	76,0% ( $\pm 0,04$ )	96,9% ( $\pm 0,01$ )	72,8% ( $\pm 0,02$ )	85,7% ( $\pm 0,02$ )	86,4% ( $\pm 0,01$ )
Inception	92,8% ( $\pm 0,03$ )	58,3% ( $\pm 0,14$ )	90,2% ( $\pm 0,05$ )	93,0% ( $\pm 0,04$ )	69,6% ( $\pm 0,09$ )	91,5% ( $\pm 0,01$ )	91,6% ( $\pm 0,01$ )
InceptionResnet	95,7% ( $\pm 0,01$ )	70,4% ( $\pm 0,07$ )	88,3% ( $\pm 0,04$ )	96,4% ( $\pm 0,01$ )	78,0% ( $\pm 0,03$ )	92,2% ( $\pm 0,01$ )	92,3% ( $\pm 0,01$ )

Tabela 5.24: Intervalo de confiança utilizando o *Dataset* ROS.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
VGG-16	80,95% - 97,74%	25,82% - 67,92%	77,59% - 95,6%	79,66% - 99,54%	43,17% - 76,41%	85,85% - 90,07%	85,95% - 90,24%
VGG-19	83,25% - 91,98%	30,93% - 49,63%	75,47% - 93,04%	82,42% - 93,44%	47,42% - 60,92%	83,56% - 88,43%	83,88% - 88,31%
Xception	93,69% - 96,55%	56,95% - 84,57%	67,66% - 84,26%	94,64% - 99,18%	68,73% - 76,97%	82,02% - 89,48%	83,32% - 89,54%
Inception	93,69% - 96,55%	56,95% - 84,57%	67,66% - 84,26%	94,64% - 99,18%	68,73% - 76,97%	82,02% - 89,48%	83,32% - 89,54%
InceptionResnet	93,79% - 97,57%	57,2% - 83,52%	81,07% - 95,52%	93,87% - 98,88%	71,17% - 84,83%	89,19% - 95,25%	89,5% - 95,17%

Comparando estes resultados com os mostrados na Tabela 5.8, percebe-se que a aplicação do *Random Oversampling* resultou em uma melhoria tanto nos limites inferiores quanto superiores de cada modelo para a métrica sensibilidade, sem prejudicar de forma significativa a acurácia do modelo. Esta melhoria é importante no sentido de reforçar a confiança dos resultados dos modelos em termos de classificação correta da classe Amiloidose. Apesar disso, nota-se que para a métrica precisão, a diferença entre os limites superiores e inferiores do intervalo de confiança ainda é elevada, demonstrando que os resultados dos modelos nos 5 *folds* para a métrica foi instável.

### 5.0.7 Resultados Modelos *Ensemble-Based*

O modelo *Ensemble-Baseline* é o modelo construído com base na arquitetura proposta na Figura 4.4, treinado utilizando a base de dados *Dataset Baseline*, que apresenta um desbalanceamento severo entre as classes. Os modelos *Ensemble-ROS* e o *Ensemble-RUS* também são modelos construídos com base na arquitetura proposta na Figura 4.4, porém, foram treinados com o *Dataset ROS* e o *Dataset RUS*, respectivamente.

Na figura 12 pode-se observar as matrizes de confusão para os melhores resultados de teste dos modelos, dentre os treinados para os 5-*folds* da validação cruzada. Percebe-se que o modelo *Ensemble-Baseline* apresentou apenas 7 Falsos Positivos, entretanto foi o que mais cometeu erros do tipo Falso Negativo. Comparando os modelos *Ensemble-ROS* e *Ensemble-RUS*, entende-se que o número de Falsos Negativos de ambos é próximo, porém, o número de Falsos Positivos do *Ensemble-RUS* é de 5,3 vezes maior que o do *Ensemble-ROS*. Com isso, considerando a importância de um baixo número de classificações incorretas para a classe Amiloidose e a proporção de classificações corretas para ambas as classes, verifica-se que para este experimento o modelo *Ensemble-ROS* foi o que atingiu um melhor desempenho.

Na Tabela 5.25 apresenta o número de erros de classificação, divididos por corante e lesão. É possível perceber que dentre os corantes, os que mais apresentaram erros de classificação foram o HE e o PAS. Já entre as lesões presentes na classe Não-Amiloidose, a que apresentou mais erros de classificação foi a Esclerose.

A tabela 5.29 mostra a média dos resultados de teste dos modelos *Ensemble-Based*, avaliado com base nas métricas de avaliação utilizadas neste estudo. Dentre os três modelos, o *Ensemble-Baseline* foi o que apresentou melhor desempenho de classificação para a classe negativa, entretanto, sua sensibilidade é inferior a dos demais. Em termos de sensibilidade, percebe-se que o modelo *Ensemble-RUS* foi o que alcançou melhor desempenho, porém, ao analisar a precisão do modelo, percebe-se que é o que apresenta a maior quantidade de erros de Falso Positivo dentre os 3 modelos. Ao analisar a métrica *G-Score*, entende-se que o modelo *Ensemble-ROS* foi o que apresentou um melhor desempenho, com 92,6%, indicando que em ele foi o que mostrou um maior desempenho em termos de equilíbrio de classificações corretas para ambas as classes.

A Tabela 5.30 mostra os intervalos de confiança, com um nível de confiança de 95%, para os três modelos *Ensemble-Based* propostos. É importante observar na tabela que os limites superiores e inferiores dos intervalos de confiança para as métricas são mais estreitos, demonstrando que os modelos apresentaram uma estabilidade no resultado de classificação nos 5 *folds*.

Ao analisar a métrica sensibilidade, nota-se que o modelo *Ensemble-RUS* foi o que alcançou os maiores limites inferiores e superiores. Apesar disso, como já foi discutido acima, este modelo apresenta níveis de precisão menores que os outros dois. Com isso, entende-se que o modelo *Ensemble-RUS* foi o que apresentou o maior número de classificações corretas para a classe amiloidose, porém, devido a sua baixa precisão de 69,2%, apresentou um elevado número de Falsos Positivos. Dentre os 3 modelos, o que apresentou um melhor intervalo de confiança para a métrica G-Score, foi o *Ensemble-ROS*, dado que ele é o que apresenta o melhor balanço de classificações corretas para ambas as classes.

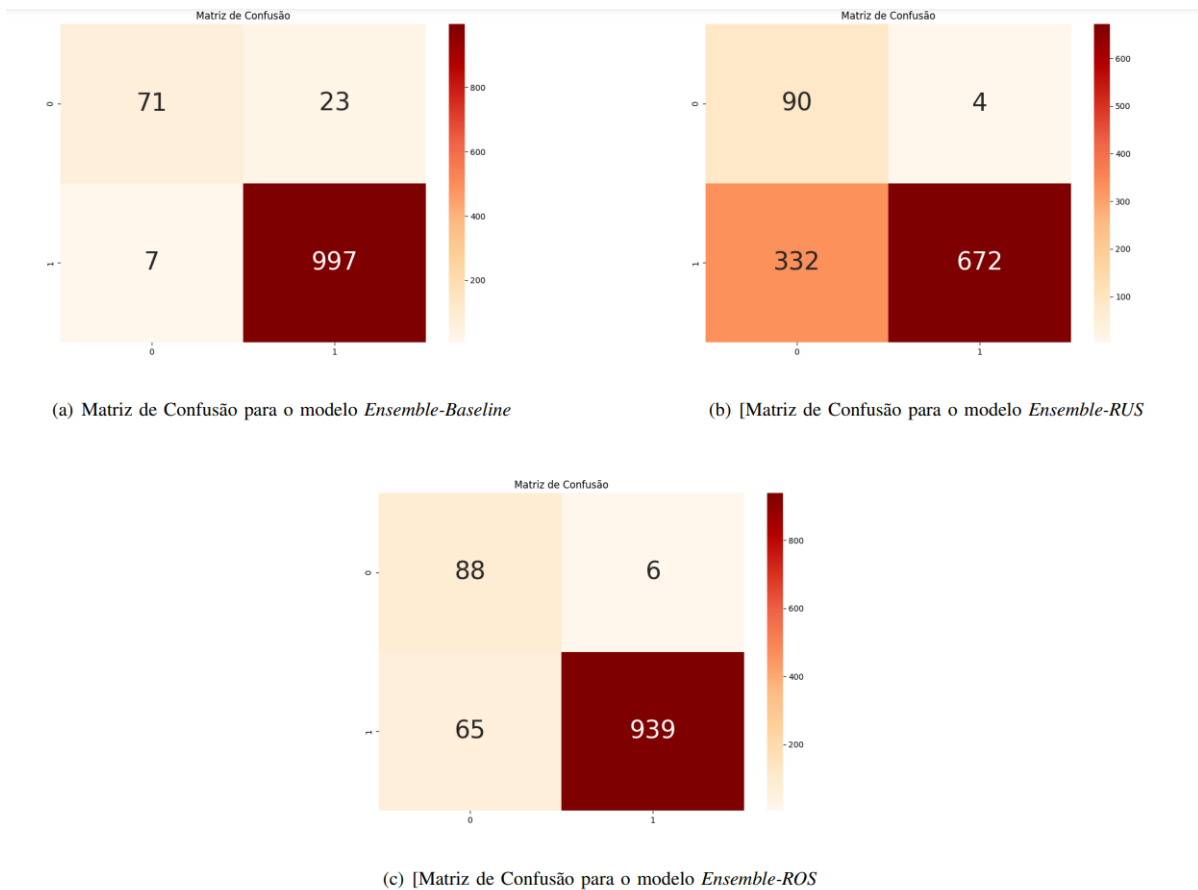


Figura 5.7: Matriz de Confusão para os melhores modelos.

Tabela 5.25: Erros de classificação por para o melhor modelos *Ensemble*

Tabela 5.26: Erros de classificação para o modelo *Ensemble-Baseline*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	10,6%	0,0%	7,4%	6,4%
Negativa	Falso Positivo	Normal	0,0%	0,0%	0,0%	0,0%
		Esclerose	0,1%	0,1%	0,3%	0,1%
		Hiper celularidade	0,0%	0,0%	0,0%	-
		Hiper celulari. Pura	-	0,1%	0,0%	-

Tabela 5.27: Erros de classificação para o modelo *Ensemble-RUS*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	7,4%	0,0%	3,2%	1,1%
Negativa	Falso Positivo	Normal	1,8%	0,1%	0,8%	0,9%
		Esclerose	1,5%	1,5%	2,1%	1,1%
		Hiper celularidade	2,1%	0,5%	1,4%	-
		Hiper celulari. Pura	-	0,2%	1,2%	-

Tabela 5.28: Erros de classificação para o modelo *Ensemble-ROS*.

Classe	Erro	Lesão	HE	AZAN	PAS	PAMS
Positiva	Falso Negativo	Amiloidose	3,2%	0,0%	1,1%	2,1%
Negativa	Falso Positivo	Normal	0,9%	0,0%	0,5%	0,3%
		Esclerose	0,6%	1,1%	1,4%	0,4%
		Hiper celularidade	0,4%	0,5%	0,2%	-
		Hiper celulari. Pura	-	0,3%	0,0%	-

Tabela 5.29: Média das Métricas de desempenho para o modelo *Ensemble-Based*.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
Ensemble-Baseline	96,8% ( $\pm 0,01$ )	91,3% ( $\pm 0,01$ )	68,7 ( $\pm 0,06$ )%	99,4% ( $\pm 0,01$ )	78,3% ( $\pm 0,04$ )	82,6% ( $\pm 0,04$ )	84,1% ( $\pm 0,03$ )
Ensemble-RUS	69,2% ( $\pm 0,01$ )	21,1% ( $\pm 0,01$ )	94,7% ( $\pm 0,01$ )	66,8% ( $\pm 0,02$ )	34,5% ( $\pm 0,01$ )	79,5% ( $\pm 0,01$ )	80,7% ( $\pm 0,01$ )
Ensemble-ROS	95,1% ( $\pm 0,01$ )	66,5% ( $\pm 0,08$ )	89,8% ( $\pm 0,02$ )	95,6% ( $\pm 0,01$ )	76,1% ( $\pm 0,04$ )	92,6% ( $\pm 0,01$ )	92,7% ( $\pm 0,01$ )

Tabela 5.30: Intervalo de confiança para o modelo *Ensemble-Based*.

<b>Modelo</b>	Acurácia	Precisão	Sensibilidade	Especificidade	F1-Score	G-Score	AUC
Ensemble-Baseline	95,87%-97,64%	88,93%-93,62%	57,03%-80,41%	99,15% - 99,61%	70,65% - 85,85%	75,53% - 89,59%	78,28% - 89,83%
Ensemble-RUS	65,87% - 72,52%	19,33% - 22,91%	93,36% - 96,0%	63,17% - 70,46%	32,12% - 36,93%	77,32% - 81,74%	78,85% - 82,64%
Ensemble-ROS	92,65% - 97,51%	52,15% - 80,9%	85,11% - 94,47%	92,52% - 98,63%	67,89% - 84,29%	91,41% - 93,82%	91,55% - 93,82%

# Capítulo 6

## Discussão

Como era esperado, o desbalanceamento de classes influenciou negativamente no resultado da classificação, especialmente para a classe positiva, que possui menos amostras disponíveis para treinamento. Ao analisar a acurácia, durante a fase de testes dos modelos treinados utilizando o *Dataset Baseline*, nota-se que todos eles obtiveram uma acurácia média igual ou superior a 94%, enquanto que a especificidade foi igual ou superior a 97%, o que pode ser considerado bom, dada a limitação de dados do problema. Entretanto, quando se analisa a métrica sensibilidade, que para este estudo tem um peso maior que as demais, percebe-se que ela não alcançou valores tão elevados quanto a acurácia, com um valor de 77% para o melhor modelo. Esta circunstância é considerada indesejável, tanto do ponto de vista de análise de resultados quanto do ponto de vista da saúde do paciente, o que por si só justificam um estudo mais aprofundado de técnicas que permitam uma melhoria na taxa de acerto da classificação para a classe positiva.

Após a aplicação da técnica de *Random Undersampling* na classe majoritária, notou-se que houve uma melhoria de 30% na Sensibilidade para os modelos VGG-16, de 58% o modelos VGG-19, de 44% o modelos *Xception*, de 35% *Inception* e de de 21% *Inception*, quando comparado com os resultados dos mesmos modelos treinados com o *Dataset Baseline*. Esta elevação de desempenho de sensibilidade dos modelos é importante no sentido de que houve uma redução nas classificações incorretas de imagens com amiloidose. Entretanto, nota-se que houveram quedas elevadas de precisão para todos os modelos, com um valor médio 74% de redução, demonstrando que apesar deles estarem mais sensível para a classe amiloidose, todos apresentam uma elevada taxa de erros de Falso Positivo. Nesse sentido, devido ao elevado número de erros de Falso Positivo, pode-se perceber uma queda nas métricas *G-Score*, *F1-Score* e *AUC*. Estes resultados indicam que para os experimentos realizados com o *Dataset RUS*, devido a redução de amostras de treinamento para a classe Não-Amiloidose, os modelos não foram capazes de aprender os padrões necessários para realizar uma correta classificação das imagens presentes na base de testes para a classe negativa. Apesar disso, a aplicação do balanceamento de dados reduziu o viés

de treinamento para a classe majoritária, visto no *Dataset Baseline*, uma vez que pode-se perceber melhorias na correta classificação da classe Não-Amiloidose.

Após a aplicação da técnica de *Random Oversampling* na classe minoritária, notou-se que houve uma melhoria de 33% na Sensibilidade para os modelos VGG-16, de 54% o modelos VGG-19, de 19% o modelos *Xception*, de 26% *Inception* e de de 14% *Inception*, quando comparado com os resultados dos mesmos modelos treinados com o *Dataset Baseline*. Nota-se que também houveram quedas de precisão para todos os modelos, com uma redução média de 30%. Estes acontecimentos indicam que a aplicação da técnica de *Random Oversampling* proporcionou o treinamento de modelos com menos erros de classificação para a classe Amiloidose, mais que erra mais para a classe Não-Amiloidose. Também houveram melhorias em termos de *G-Score* e *AUC* e uma vez que ambas as métricas são sensíveis à ambas as classes e houve uma melhoria de desempenho de taxa de acerto para a classe Amiloidose, era esperado que ambas as métricas também apresentem melhorias.

Quando se realiza um comparativo de desempenho entre as médias das métricas modelos treinados para os *5-folds* da validação cruzada com os *Dataset RUS* e o *Dataset ROS*, percebe-se que o experimento realizado com o *Dataset RUS* apresentou um melhor resultado de sensibilidade para todos os modelos. Ao analisar a média de precisão e especificidade dos modelos, notou-se que o desempenho do experimento realizado com o *Dataset ROS* foi superior em termos de resultados. Por fim, ao analisar-se a métrica *G-Score*, pode-se entender que dentre os dois experimentos, os modelos treinados com *Dataset ROS* apresentam um melhor balanço de classificação correta para ambas as classes. Com base nestes resultados, pode-se entender que os modelos treinados com o *Dataset RUS* são superiores no sentido de classificação correta da classe Amiloidose, porém, quando se analisa um balanço de classificações corretas para ambas as classes, os modelos treinados com o *Dataset ROS* obtém melhores resultados.

Ao se comparar a matriz de confusão para os melhores resultados dos modelo dentre os treinados para os *5-folds* da validação cruzada com os *Dataset RUS* e o *Dataset ROS*, verifica-se que a maior diferença entre os dois está na maior quantidade de erros de Falso Positivo para os modelos treinados com o *Dataset RUS*. Um dos motivos para este acontecimento é a redução na quantidade de amostras disponíveis para treinamento dos modelos, decorrente da aplicação da técnica de balanceamento de dados. Nesse sentido, a menor quantidade de exemplos visto pelos modelos treinados com o *Dataset RUS* não possibilitou que ele aprendesse os padrões necessários para classificar corretamente a classe Não-Amiloidose com o mesmo nível de taxa de acerto dos modelos treinados com o *Dataset ROS*, que processaram mais exemplos de dados. Assim, percebe-se que a redução na quantidade de amostras de treinamento para o *Dataset RUS*, associado ao fato da classe Não-Amiloidose ser composta por diferentes grupos de imagens, gerou um viés de treinamento para a classe Amiloidose, que pode ser percebido com o elevado número de erros de Falso Negativo vistos durante a fase de teste dos modelos.

Com relação aos modelos *EnsembleBased-Baseline*, *EnsembleBased-RUS* e *EnsembleBased-ROS*, pode-se perceber que, mesmo que eles preservem os problemas identificados nos experimentos de base, treinados utilizando os *Datasets Baseline*, *RUS* e *ROS*, houveram melhorias de desempenho de classificação correta para a classe Amiloidose. Para o modelo *EnsembleBased-Baseline*, percebeu-se uma baixa quantidade de erros de Falso Positivo, apenas 7 erros, porém uma maior quantidade de erros de Falso Negativo, um total de 23 das 94 imagens com Amiloidose. O maior ganho de desempenho foi identificado no modelo *EnsembleBased-Baseline*, que apresentou um G-Score 92,6%, maior valor identificado neste trabalho.

Ao comparar o desempenho dos modelos *EnsembleBased-Baseline*, *EnsembleBased-RUS* e *EnsembleBased-ROS*, com os modelos base, VGG16, VGG19, *Xception*, *Inception* e *Inception-Resnet*, treinados com a mesma base de dados, entende-se que não houve um ganho de desempenho significativo de sensibilidade, ao mesmo tempo em que os modelos *Inception* e *Inception-Resnet* foram os apresentando melhor desempenho para a métrica, quando treinados com o *Dataset ROS*.

Ao realizar uma análise de erros por corante, notou-se que os corantes que mais apresentam erros de classificação foram os HE e PAS. Uma das possíveis causas para o maior número de classificações incorretas para os corante HE e PAS é a maior dificuldade dos modelos em capturar detalhes que possam identificar a imagem como amiloidose tal como nos outras corantes. Com relação aos erros de classificação por lesões percebeu-se que a lesão que apresentou mais erros de classificação foi a esclerose. Em conversa com o líder médico do projeto PathoSpotter, especula-se que o maior número de classificações incorretas para esclerose tem relação com suas características visuais e pela forma como ela se apresenta no glomérulo, que faria com que as imagens com esclerose se assemelhassem visualmente com a amiloidose, gerando confusões de classificação para os modelos. Entretanto, para que esta hipótese seja confirmada ou não, são necessários mais experimentos e análises posteriores.



# Capítulo 7

## Conclusões

Este trabalho teve como proposta realizar classificação automática de amiloidose renal em imagens digitais de biópsias utilizando corantes não específicos para a lesão. Além disso, dado que a classificação de amiloidose trouxe como desafio secundário um problema de desbalanceamento de classes, foram realizadas também análises comparativas entre diferentes métodos para lidar com esta situação, utilizando como referencial teórico abordagens disponíveis na literatura.

Como etapa inicial deste trabalho foi montada uma base de dados de imagens constituída por duas classes de imagens de glomérulos. A primeira contendo somente amiloidose e a segunda formado por 4 grupos de imagens: glomérulos sem nenhum tipo de lesão, com lesão do tipo esclerose pura sem crescente, com lesão do tipo hiper celularidade, e com lesão do tipo hiper celularidade pura sem crescente. Esta base de dados foi construída a partir de imagens de biópsias coletadas por pesquisadores do Instituto Gonçalo Moniz da Fundação Oswaldo Cruz - Bahia (IGM/FIOCRUZ).

Para o desenvolvimento do classificador de amiloidose, foram utilizados os modelos de Redes Neurais Convolucionais VGG16, VGG19, *Xception*, *Inception* e *Inception-Resnet*. A fim de estabelecer um ponto de partida inicial para a pesquisa, os modelos utilizados foram treinados com a base de dados original, com desbalanceamento para a classe amiloidose. Os resultados deste primeiro experimento apresentaram números elevados de Falso Negativo, demonstrando a necessidade da aplicação de métodos capazes de lidar com o problema de desbalanceamento de classes da amiloidose. No que diz respeito às estratégias para abordar o problema de desbalanceamento de classes presente na amiloidose, foram investigados os métodos *Random Undersampling*, *Random Oversampling* e o *Ensemble-Based*.

Nesse contexto, constatou-se a viabilidade em obter resultados de classificação capazes de identificar lesão em imagens com corantes não específicos, com uma taxa de falsos negativos de até 4,5%, identificados nos modelos *Inception*, quando treinado com o *Dataset RUS* e para o modelo o *Ensemble-RUS*, modelos *Ensemble-Based* treinados com o *Dataset RUS*. Foi identificado que o melhor desempenho em termos

de sensibilidade foi alcançado pelo modelo *Inception*, quando treinado com uma base de dados balanceada utilizando a técnica *Random Undersampling*, que alcançou uma sensibilidade de 96,8%. Apesar disso, observou-se que o modelo *Ensemble-ROS* se saiu melhor em termos de classificações corretas para ambas as classes, alcançando um *G-Score* de 92,6%.

Com relação aos erros de classificação por corantes, pode-se observar uma predominância maior de erros para os corantes HE e PAS, mesmo após a aplicação das estratégias para lidar com o desbalanceamento de classes. Uma das possíveis causas para o maior número de classificações incorretas para estes corantes é a maior dificuldade dos modelos em capturar características nas imagens que possam identificá-las como amiloidose. Em relação aos erros por lesão, para a classe Não-Amiloidose, pode-se perceber que os modelos apresentaram maiores taxas de classificação incorreta para a lesão esclerose, se comparado com as demais lesões.

Apesar de ser um estudo inicial, os resultados deste trabalho proporcionam a identificação de pontos relevantes para a construção de um modelo de classificação de amiloidose a partir de corantes não específicos, em um cenário onde a classe Não-Amiloidose é composta por diferentes tipos de lesões juntamente com imagens sem qualquer lesão nos glomérulos, que é o cenário comum encontrado na prática pelos patologistas. Com isso, estes resultados abrem espaço para automação da classificação de imagens com amiloidose, viabilizando ganhos operacionais importantes do ponto de vista da nefropatologia, uma vez que a solução proposta aqui tem potencial para simplificar o protocolo usado na análise desta lesão, considerando que pode dispensar a necessidade de corantes especiais para o processamento da biópsica, visando detectar a presença de amiloidose, o que resultará em benefícios operacionais significativos do ponto de vista da nefropatologia.

Com relação a base de dados, entende-se que ela possui vieses oriundos do método de obtenção dos dados, que apesar de não inviabilizarem os estudos realizados aqui, devem ser levados em consideração no momento de melhorar a base de dados em trabalhos futuros. Nesse sentido, pretende-se aumentar a quantidade e diversidades das imagens, incluindo-se imagens coletadas de outros centros de patologia, bem como com maior variedade de classes de lesões, com o objetivo de construir uma base de dados ainda mais próxima de um cenário visto pelos médicos patologistas.

Como outros trabalhos futuros, pretende-se também aprimorar os classificadores desenvolvidos, aplicando técnicas como métodos de atenção ou métodos de treinamento semi-supervisionado, que possam auxiliar na obtenção de um modelo melhor para a tarefa, bem como aplicar métodos de interpretação dos resultados dos modelos, como mapa de calor e Grad-CAM. Outro trabalho futuro será a utilização de classificadores para corantes individuais em lugar de usar todos juntos como foi feito nesse trabalho, para investigar se apenas usando um corante individual é possível extrair características mais discriminatórias da amiloidose, viabilizando um melhor desempenho do classificador.

# Referências

- Abd Elrahman, S. M. e Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013):332–340.
- Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., et al. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294.
- Agibetov, A., Kammerlander, A., Duca, F., Nitsche, C., Koschutnik, M., Donà, C., Dachs, T.-M., Rettl, R., Stria, A., Schrutka, L., et al. (2021). Convolutional neural networks for fully automated diagnosis of cardiac amyloidosis by cardiac magnetic resonance imaging. *Journal of Personalized Medicine*, 11(12):1268.
- Ahmed, S. A. A., Yanikoğlu, B., Göksu, Ö., e Aptoula, E. (2020). Skin lesion classification with deep cnn ensembles. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, páginas 1–4. IEEE.
- Alex, S. A., Jhanjhi, N., Humayun, M., Ibrahim, A. O., e Abulfaraj, A. W. (2022). Deep lstm model for diabetes prediction with class balancing by smote. *Electronics*, 11(17):2737.
- Arafa, A., El-Fishawy, N., Badawy, M., e Radad, M. (2022). Rn-smote: Reduced noise smote based on dbscan for enhancing imbalanced data classification. *Journal of King Saud University-Computer and Information Sciences*, 34(8):5059–5074.
- Calumby, R. T., Duarte, A. A., Angelo, M. F., Santos, E., Sarder, P., Dos-Santos, W. L., e Oliveira, L. R. (2023). Toward real-world computational nephropathology. *Clinical journal of the American Society of Nephrology: CJASN*, 18(6):809–812.
- Chagas, P., Souza, L., Araújo, I., Aldeman, N., Duarte, A., Angelo, M., Dos-Santos, W. L., e Oliveira, L. (2020). Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artificial intelligence in medicine*, 103:101808.
- Chagas, P., Souza, L., Pontes, I., Calumby, R., Angelo, M., Duarte, A., dos Santos, W. L., e Oliveira, L. (2021). Deep-learning-based membranous nephropathy

- classification and monte-carlo dropout uncertainty estimation. In *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, páginas 257–268. SBC.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1251–1258.
- da Silva, J. M., Angelo, M. F., dos Santos, W. L., e Loula, A. C. (2021). Aprendizado profundo na classificação de lesões crescentes glomerulares: modelos e condições. In *Anais Estendidos do XXXIV Conference on Graphics, Patterns and Images*, páginas 162–165. SBC.
- Dablain, D., Krawczyk, B., e Chawla, N. V. (2022). Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*.
- Devi, D., Biswas, S. K., e Purkayastha, B. (2020). A review on solution to class imbalance problem: undersampling approaches. In *2020 international conference on computational performance evaluation (ComPE)*, páginas 626–631. IEEE.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., e Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.
- He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778.
- Johnson, J. M. e Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.
- Lemaître, G., Nogueira, F., e Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., e Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.
- Maulidevi, N. U., Surendro, K., et al. (2022). Smote-lof for noise identification in imbalanced data classification. *Journal of King Saud University-Computer and Information Sciences*, 34(6):3413–3423.

- Medela, A., Picon, A., Saratxaga, C. L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., e Glover, B. (2019a). Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, páginas 1860–1864. IEEE.
- Medela, A., Picon, A., Saratxaga, C. L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., e Glover, B. (2019b). Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, páginas 1860–1864. IEEE.
- Monteiro, N. F. e Diz, M. C. E. (2015). Difficulties in the diagnosis of primary amyloidosis: case report. *Rev Med Minas Gerais*, 25(2):268–274.
- Moraes, C. A. d. e Colicigno, P. R. C. (2007). Estudo morfofuncional do sistema renal.
- Palladini, G., Campana, C., Klersy, C., Balduini, A., Vadacca, G., Perfetti, V., Perlini, S., Obici, L., Ascari, E., d’Eril, G. M., et al. (2003). Serum n-terminal pro-brain natriuretic peptide is a sensitive marker of myocardial dysfunction in al amyloidosis. *Circulation*, 107(19):2440–2445.
- Ravi, V., Narasimhan, H., e Pham, T. D. (2022). A cost-sensitive deep learning-based meta-classifier for pediatric pneumonia classification using chest x-rays. *Expert Systems*, página e12966.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.
- Shinde, P. P. e Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, páginas 1–6. IEEE.
- Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sipe, J. D., Benson, M. D., Buxbaum, J. N., Ikeda, S.-i., Merlini, G., Saraiva, M. J., e Westermarck, P. (2016). Amyloid fibril proteins and amyloidosis: chemical identification and clinical classification international society of amyloidosis 2016 nomenclature guidelines. *Amyloid*, 23(4):209–213.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., e Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 2818–2826.
- Taherkhani, A., Cosma, G., e McGinnity, T. M. (2020a). Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, 404:351–366.

- Taherkhani, A., Cosma, G., e McGinnity, T. M. (2020b). Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, 404:351–366.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., e Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.
- Timm, L. d. L. (2005). Técnicas rotineiras de preparação e análise de lâminas histológicas. *Caderno La Salle XI*, 2(1):231–239.
- Wang, Y., Yao, Q., Kwok, J. T., e Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.
- Zhang, Y.-P., Zhang, L.-N., e Wang, Y.-C. (2010). Cluster-based majority under-sampling approaches for class imbalance learning. In *2010 2nd IEEE International Conference on Information and Financial Engineering*, páginas 400–404. IEEE.
- Zhuang, D., Chen, K., e Chang, J. M. (2020). Cs-af: A cost-sensitive multi-classifier active fusion framework for skin lesion classification. *arXiv preprint arXiv:2004.12064*.