Metodologia para a Classificação Multiclasse de Imagens Histológicas baseada em Inteligência Artificial Explicável

Tiago Pereira de Faria

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2022

Tiago Pereira de Faria

Metodologia para a Classificação Multiclasse de Imagens Histológicas baseada em Inteligência Artificial Explicável

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Luiz Gustavo Almeida Martins
Coorientador: Prof. Dr. Marcelo Zanchetta do Nascimento

Uberlândia
2022

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

F224 Faria, Tiago Pereira de, 1997-
2022 Metodologia para a Classificação Multiclasse de Imagens Histológicas baseada em Inteligência Artificial Explicável [recurso eletrônico] / Tiago Pereira de Faria. - 2022.

Orientador: Luiz Gustavo Almeida Martins.
Coorientador: Marcelo Zanchetta do Nascimento.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2022.198
Inclui bibliografia.

1. Computação. I. Martins, Luiz Gustavo Almeida,1974-, (Orient.). II. Nascimento, Marcelo Zanchetta do,1976-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado 7/2022, PPGCO				
Data:	29 de abril de 2022	Hora de início:	9:04	Hora de encerramento:	11:14
Matrícula do Discente:	11922CCP013				
Nome do Discente:	Tiago Pereira de Faria				
Título do Trabalho:	Metodologia para a Classificação Multiclasse de Imagens Histológicas baseada em Inteligência Artificial Explicável				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Bruno Augusto Nassif Travençolo - FACOM/UFU, Alessandro Santana Martins/IFTM, Marcelo Zanchetta do Nascimento - FACOM/UFU (coorientador) e Luiz Gustavo Almeida Martins - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Alessandro Santana Martins - Ituiutaba/MG; Bruno Augusto Nassif Travençolo, Marcelo Zanchetta do Nascimento e Luiz Gustavo Almeida Martins - Uberlândia/MG. O discente participou da cidade de Uberlândia/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Luiz Gustavo Almeida Martins, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Luiz Gustavo Almeida Martins**, **Professor(a) do Magistério Superior**, em 05/05/2022, às 15:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 05/05/2022, às 15:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 05/05/2022, às 18:51, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alessandro Santana Martins**, **Usuário Externo**, em 06/05/2022, às 10:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3539221** e o código CRC **C4BFDFA4**.

Para meus pais, que possibilitaram isso tudo

Agradecimentos

Agradeço ao meu orientador, Luiz Gustavo Almeida Martins, e ao meu co-orientador Marcelo Zanchetta do Nascimento, por terem aceitado esse desafio junto comigo, e pelo imenso apoio que me deram durante o desenvolvimento deste trabalho. Testemunhei, durante este período, a dedicação e empenho que eles têm para aprender, ensinar, e contribuir para o avanço da ciência.

Também devo agradecimentos aos professores da Faculdade de Computação da Universidade Federal de Uberlândia, pelo trabalho impecável desempenhado nos programas de graduação e pós-graduação que participei. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Projeto 11404/2021-9, Projeto 145081/2019-2 e Projeto 311404/2021-9), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Projeto APQ-00578-18) e ao Programa de Pós-Graduação em Computação pelos apoios financeiros.

À meus pais, irmãos e avó, que me acompanharam de perto nessa jornada, e foram o apoio perfeito que eu precisava. À João Paulo, Luiz Henrique, Maycon e Gabriel, que a mais de dez anos, servem de inspiração para que eu continue tentando ser sempre uma pessoa melhor. Aos mestres Douglas e Márcio, que trilharam esse caminho junto comigo, e têm participação em cada parte deste trabalho. À todos os amigos que ajudaram, cada um do seu jeito, a me manter são e motivado para seguir em frente (me desculpem por não citar todos os nomes).

Resumo

Sistemas de diagnóstico apoiado por computador (CAD) têm sido estudados como uma ferramenta para diminuir a variabilidade entre especialistas e agilizar o processo de diagnóstico. Nesse tipo de sistema, técnicas de visão computacional e aprendizado de máquina costumam ser empregadas no diagnóstico automático a partir de imagens histopatológicas. Apesar de resultar em classificadores de alta precisão, muitos desses métodos são caixas-pretas, em que o conhecimento usado na decisão está implícito no modelo, ou é representado de forma complexa. Esse aspecto diminui a confiabilidade do sistema e dificulta sua depuração. Nesse contexto, este trabalho propõe uma metodologia que integra métodos de classificação multiclasse e de inteligência artificial explicável (XAI) para a construção de sistemas CAD mais interpretáveis e confiáveis. A predição é feita a partir de descritores morfológicos e não-morfológicos, extraídos dos núcleos celulares identificados nas imagens segmentadas. Esses descritores são pré-processados e empregados na construção dos modelos. A fim de melhorar a compreensão acerca das classificações, nossa metodologia integra diferentes técnicas de XAI. Uma estratégia baseada em preditores binários estima a confiabilidade das decisões do modelo, categorizando-as em confiáveis, incertas ou inconclusivas. Então, métodos baseados em âncoras (*anchors*) e no *Shapley Additive Explanations* (SHAP) são usados para explicar o comportamento global e local do modelo. Uma nova forma de exibição das âncoras fornece uma alternativa para interpretar a decisão do modelo preditivo. Por fim, uma abordagem de exibição dos dados, baseada em histogramas e visualizações 3D, auxilia no diagnóstico dos casos incertos ou inconclusivos. Essa metodologia foi avaliada na classificação de imagens histológicas de linfomas não Hodkin e de displasias orais, resultando em modelos com uma acurácia média em torno de 94% e 92%, respectivamente. Com base nas análises de interpretação, foi possível melhorar a compreensão sobre o comportamento de classificadores multiclasse.

Palavras-chave: Inteligência artificial explicável. Classificação multiclasse. Descritores morfológicos e não-morfológicos. Imagens histológicas. Sistema de apoio ao diagnóstico.

Abstract

Computer-aided diagnostic (CAD) systems have been studied as a tool to decrease inter-expert variability and streamline the diagnostic process. In this type of system, computer vision and machine learning techniques are usually employed in the automatic diagnosis from histopathological images. Despite resulting in high precision classifiers, many of these methods are black boxes, in which the knowledge used in the decision is implicit in the model, or is represented in a complex way. This aspect decreases the reliability of the system and makes it difficult to debug. In this context, this work proposes a methodology that integrates multiclass classification and explainable artificial intelligence (XAI) methods to build more interpretable and reliable CAD systems. The prediction is made from morphological and non-morphological descriptors, extracted from the cell nuclei identified in the segmented images. These descriptors are pre-processed and used in the construction of the models. In order to improve the understanding of classifications, our methodology integrates different XAI techniques. A strategy based on binary predictors estimates the reliability of model decisions, categorizing them as reliable, uncertain or inconclusive. Then, methods based on Anchors and Shapley Additive Explanations (SHAP) are used to explain the global and local behavior of the model. A new way of displaying anchors provides an alternative to interpret the predictive model's decision. Finally, a data display approach, based on histograms and 3D visualizations, helps in the diagnosis of uncertain or inconclusive cases. This methodology was evaluated in the classification of histological images of non-Hodkin lymphomas and oral dysplasias, resulting in models with an average accuracy of around 94% and 92%, respectively. Based on the interpretation analyses, it was possible to improve the understanding of the behavior of multiclass classifiers.

Keywords: Explainable artificial intelligence. Multiclass classification. Morphological and non-morphological descriptors. Histological images. Diagnostic support system.

Lista de ilustrações

Lista de tabelas

Lista de siglas

ANOVA *Analysis of variance* (análise de variância)

CAD *Computer Aided Diagnosis* (diagnose guiada por computador)

CP *Ceteris Paribus*

CLAHE *Contrast Limited Adaptive Histogram Equalisation*(equalização adaptativa de histograma com contraste limitado)

CLL *Chronic Lymphocytic Leukemia* (Leucemia Linfocítica Crônica)

DOE Displasia Oral Epitelial

EDA *Exploratory Data Analysis* (analise exploratória de dados)

FP Falsos positivos

FN Falsos negativos

FL *Follicular Lymphoma* (linfoma Folicular)

GI Ganho de Informação

GBDT *Gradient Boosting Decision Trees*

H&E Hematoxilina-Eosina

LH Linfoma Hodkin

LIME *Local Interpretable Model-Agnostic Explanations* (explicações modelo-agnósticas locais e interpretáveis)

LNH Linfomas Não Hodkin

MCL *Mantle Cell Lymphoma* (linfoma de células do manto)

MLP *Multilayer Perceptron* (perceptron multicamadas)

PCA *Principal Component Analysis* (análise de componentes principais)

PDP *Partial Dependence Plot* (mapeamento de dependência parcial)

RL Regressão Linear

ReLU *Rectified Linear Unit* (unidade linear retificada)

SHAP *SHapley Additive exPlanations* (explicações aditivas de shapley)

SVM *Support Vector Machine* (máquina de vetores de suporte)

WND *Weighted Neighbours Distance* (distância de vizinhos balanceados)

VP Verdadeiros positivos

XAI *Explainable Artificial Intelligence* (inteligência artificial explicável)

Sumário

Capítulo 1

Introdução

Dentre os diversos métodos empregados no apoio ao diagnóstico médico por meio de sistema de imagens, a histopatologia é uma área que investiga o uso de imagens microscópicas de seções de tecidos que as doenças podem afetar (BENTAIEB; HAMARNEH, 2018). As amostras de tecidos são coradas com Hematoxilina-Eosina (H&E) e submetidas à análise de um especialista por meio de microscópio, que pode identificar características nas células que indiquem a incidência de uma doença. Essas observações são essenciais ao permitirem o acompanhamento da doença, a identificação de seu estágio e orientação a tratamentos eficazes para o paciente (ORLOV et al., 2010). A análise de imagens histológicas pode evidenciar informações importantes para o diagnóstico de diversas doenças. Porém, é uma tarefa complexa que demanda um tempo considerável e um alto grau de conhecimento. Ela pode ser influenciada pela subjetividade de cada especialista, resultando em variabilidade de diagnósticos inter e intra-patologistas (MARTINS et al., 2020). Sistemas de apoio ao diagnóstico, também conhecidos como sistemas de diagnóstico apoiado por computador (do inglês, *computer aided diagnosis* - CAD) são ferramentas empregadas com objetivo de auxiliar os especialistas na análise de diversos tipos de lesões na área médica. O estudo de metodologias empregando técnicas baseadas em visão computacional e aprendizado de máquina podem ser usadas para classificar tecidos lesionados, permitindo maior objetividade, diminuindo a variabilidade no processo de análise de imagens. Além disso, a análise de descritores obtidos por técnicas computacionais de extração de características pode ajudar na descoberta de novos aspectos biológicos em tecidos com presença de câncer (BECK et al., 2011).

Um dos tipos de câncer com mais casos diagnosticados é o linfoma, que ataca células relacionadas ao sistema linfático, responsável pela defesa imunológica do organismo (ORLOV et al., 2010). Dentre os diversos tipos dessa doença, os que se encaixam no grupo de Linfomas Não Hodkin (LNH) são os mais comuns, em contraste com os Linfomas Hodkin (LH), mais raros. O Linfoma de Células do Manto (MCL), o Linfoma Folicular (FL) e a Leucemia Linfocítica Crônica (CLL) são três tipos de LNH, e podem ser diferenciados entre si por características morfológicas, imunofenotípicas, genéticas e clínicas (ORLOV

et al., 2010). Diagnosticar o tipo de um linfoma é crucial para um tratamento efetivo da doença, porém não é uma tarefa trivial nem mesmo para os especialistas.

Outra doença para qual um médico pode utilizar da histopatologia em seu diagnóstico é a displasia. Ela é um tipo comum de lesão pré câncer (que pode originar um câncer), que também pode ser diagnosticada realizando análises de imagens histológicas do tecido lesionado e classificada de acordo com a sua gravidade entre leve, moderada ou severa. O termo displasia se refere a um crescimento desordenado e mudanças morfológicas de células afetadas. Quando ocorre em células epiteliais da cavidade bucal, é chamada de Displasia Oral Epitelial (DOE). Assim como os linfomas, o diagnóstico e análise de imagens histológicas dessa doença de forma manual é uma tarefa complexa e sujeita a subjetividade do especialista. Por isso, o uso de CAD pode ajudar no correto diagnóstico e tratamento da doença, provendo ferramentas para evidenciar informações contidas nessas imagens e reduzir a subjetividade na classificação das lesões.

Boa parte das abordagens existentes na literatura buscam aprimorar os modelos de classificação entre as lesões apresentadas (SHAMIR et al., 2008; CODELLA et al., 2016; MARTINS et al., 2020; MENG et al., 2010; NASCIMENTO et al., 2015; SONG et al., 2016; NASCIMENTO et al., 2018; BAI et al., 2019; ADEL et al., 2018; SILVA et al., 2022), combinando técnicas ou adotando descritores mais complexos. Entretanto, apesar de melhorar o desempenho dos modelos, tais recursos dificultam a interpretação ou explicação dos seus resultados. Quando um especialista usa um sistema CAD, mais do que uma predição acurada, ele também precisa entender como a decisão foi tomada. Portanto, é importante para um CAD produzir resultados que permita ao especialista compreender melhor o contexto que levou a cada diagnóstico (ARRIETA et al., 2020). Neste contexto, um sistema que produz uma explicação para suas decisões é visto como mais confiável pelos médicos e pacientes. Esse tipo de abordagem também possibilita destacar as informações importantes que devem ser observadas pelos médicos em seu diagnóstico e no tratamento da doença em cada paciente. Além disso, uma análise destas explicações pode ser utilizada para entender possíveis problemas que o sistema possa apresentar e guiar eventuais melhorias (LUNDBERG et al., 2020).

Nossa pesquisa propõe uma metodologia de classificação multiclasse de imagens histológicas capaz de fornecer explicações locais e globais acerca dos diagnósticos, além de análises dos dados que foram utilizados durante a criação do modelo. Para extrair características numéricas das imagens médicas, foi utilizado o método de segmentação de núcleos e extração de características proposto em (NASCIMENTO et al., 2018). Descritores morfológicos e não morfológicos são extraídos das imagens segmentadas, os quais são baseados na forma desses núcleos e no nível de brilho de seus pixels, respectivamente. Em seguida, serão avaliadas diferentes técnicas de pré-processamento (engenharia de atributos e tratamento de dados) e classificação supervisionada. Por fim, o sistema utilizará de métodos de explicação dos dados, que indicam estruturas e relações existentes nos

dados utilizados na construção do modelo; de explicação local, que evidenciam o contexto e as informações que levaram o sistema a fazer cada predição; e de explicação global, que buscam destacar, de forma interpretável, informações relevantes ao funcionamento do modelo preditivo como um todo.

1.1 Motivação

Muitos trabalhos têm sido publicados sobre a utilização de métodos de visão computacional e aprendizado de máquina na criação de sistemas para auxiliar no diagnóstico médico. Esses estudos podem, além de prover uma ferramenta que diminui a subjetividade no diagnóstico médico, evidenciam características nas células que podem ajudar na descoberta de novos aspectos dessas doenças. Para algumas doenças, como os linfomas e a displasia oral epithelial, os sistemas CAD são recomendados, pois seu diagnóstico manual é geralmente complexo e sujeito à subjetividade do especialista, o que pode levar a variabilidade do diagnóstico inter e intra-patológico. Apesar de diversos estudos já terem sido realizados neste assunto, a maioria têm focado no aprimoramento do desempenho dos classificadores, adotando descritores que não são facilmente interpretados e compreendidos pelos médicos, ou métodos de classificação que encapsulam o conhecimento utilizado nas predições, como por exemplo, algoritmos baseados em máquinas de vetores de suporte (do inglês, *support vector machine* - SVM) ou redes neurais artificiais, sem a utilização de técnicas de interpretação.

Mais recentemente, pesquisas têm sido realizadas a fim de proporcionar formas de melhorar a interpretabilidade de sistemas preditivos. Esse campo de estudo é chamado de Inteligência Artificial Explicável (do inglês, *explainable artificial intelligence* - XAI) (ARRIETA et al., 2020). Os esforços nessa área podem ser classificados em duas categorias principais: (i) métodos que visam criar modelos preditivos que sejam naturalmente interpretáveis, ou seja, técnicas nas quais a informação utilizada na predição e a forma como ela é utilizada seja transparente para o usuário, tais como: árvores de decisão [2, 48] e modelos de regressão [2, 49]; e (ii) técnicas que buscam explicar a decisão de modelos "caixa preta" (não interpretáveis), mas que costumam resultar em classificadores com maior acurácia. Essa segunda categoria, foca em desenvolver técnicas que possam ser aplicadas a modelos já existentes, com o objetivo de explicar seu funcionamento como um todo, ou predições individuais (ARRIETA et al., 2020). Assim, este trabalho propõe a criação de uma metodologia que integra um modelo de classificação de imagens histológicas baseado em descritores mais simples (morfológicos e não morfológicos) à técnicas de explicação, relativas ao modelo de forma geral, aos dados utilizados na sua construção, e a cada predição especificamente, a fim de facilitar a interpretação dos diagnósticos pelos usuários.

1.2 Objetivos e Desafios da Pesquisa

O principal objetivo deste trabalho é investigar técnicas de inteligência artificial interpretável a fim de entender a influência de descritores morfológicos e não morfológicos na classificação de imagens histológicas relacionadas à amostras de diferentes lesões de tecidos histológicos.

Para isso, busca-se alcançar os seguintes objetivos específicos:

i) A formulação de uma metodologia de classificação multiclasse para auxiliar em diagnósticos médicos baseados em imagens histológicas, bem como sua avaliação em duas bases de dados. A primeira contém imagens de tecidos do sistema linfático afetados por linfoma e classificadas em três tipos diferentes da doença. A segunda contém imagens de tecidos da língua afetados ou não com displasia epitelial oral, e categorizada em três graus de evolução da doença;

ii) A utilização de métodos *post-hoc* de explicação de modelos preditivos, a fim de construir um sistema de apoio ao diagnóstico mais confiável e transparente, provendo mais informação ao médico especialista acerca das causas de cada decisão;

iii) O desenvolvimento de uma metodologia de explicação de predições multiclasse, adaptando técnicas de XAI que são criadas para classificação entre duas classes;

iv) A formulação e avaliação de uma metodologia de estimação da confiabilidade de classificações multiclasse baseado em modelos de classificação binária auxiliares;

v) Uma análise da explicação de modelos de classificação de diferentes casos de uso, e de predições de amostras de cada caso, demonstrando como as explicações formuladas pela metodologia proposta podem ser apresentadas, interpretadas e utilizadas em contextos reais.

vi) A formulação de uma metodologia, utilizando técnicas de explicação locais, para análise de predições erradas, a fim de guiar possíveis melhorias em sistemas de classificação e melhor entendimento do problema e dos dados.

1.3 Hipóteses

Na área do estudo e tratamento de lesões histológicas, os pontos em aberto são:

i) Há um subconjunto de descritores morfológicos e não morfológicos de menor tamanho que o conjunto original que é suficiente para criação de um classificador multiclasse eficiente para os problemas propostos, e este subconjunto pode ser estimado por meio de algoritmos de filtragem de atributos;

ii) Técnicas podem ser usadas para criar modelos de classificação multiclasse interpretáveis, capazes de explicitar o conhecimento utilizado nas tomadas de decisão, sem comprometer de forma significativa o desempenho em relação ao estado da arte;

iii) Utilizando predições de modelos de classificação binária auxiliares, é possível estimar a confiabilidade de uma predição multiclasse, categorizando-a em diferentes níveis de confiança.

1.4 Contribuições

Este trabalho contribui com a apresentação de uma metodologia nova para a criação de sistemas CAD explicáveis, combinando diferentes técnicas já existentes com ferramentas para adaptá-las ao contexto de imagens histológicas multiclasse. O método se mostrou robusto o suficiente quando utilizado em um segundo caso de uso, sem a necessidade de adaptações ou mudanças de parâmetros. Como a área de classificação de imagens histológicas de linfomas e displasias apresenta uma escassez de publicações focadas em XAI, este trabalho se apresenta como parte da fundação dos estudos nesta área, e espera-se que ele sirva de base para futuros avanços científicos.

1.5 Organização da Dissertação

Neste capítulo, uma introdução ao trabalho foi apresentada, com uma contextualização sobre o tema e os problemas abordados, uma dissertação sobre importância do trabalho, e uma apresentação dos objetivos da pesquisa, das hipóteses levantadas, e das contribuição alcançadas com a pesquisa. O restante dessa dissertação consiste em:

- Fundamentação Teórica: uma introdução aos casos de uso abordados, aos algoritmos de classificação multiclasse e à Inteligência Artificial Explicável, além da contextualização dos trabalhos do estado da arte sobre classificação de linfomas não Hodkin e displasia oral epitelial, e sobre a aplicação de métodos de explicação em sistemas de classificação de dados médicos multiclasse;

- Metodologia Proposta: descrição da solução proposta e das técnicas utilizadas na sua construção, assim como uma apresentação das bases de dados utilizadas;

- Experimentos e Análises dos Resultados: descrição da metodologia usada na avaliação da abordagem proposta, apresentação dos resultados obtidos e discussão sobre cada um deles bem como uma demonstração de como cada explicação podem ser interpretada e utilizada;

- Conclusões: discute as implicações dos resultados obtidos, as limitações encontradas na metodologia, e apresenta trabalhos futuros.

Capítulo 2

Fundamentação Teórica

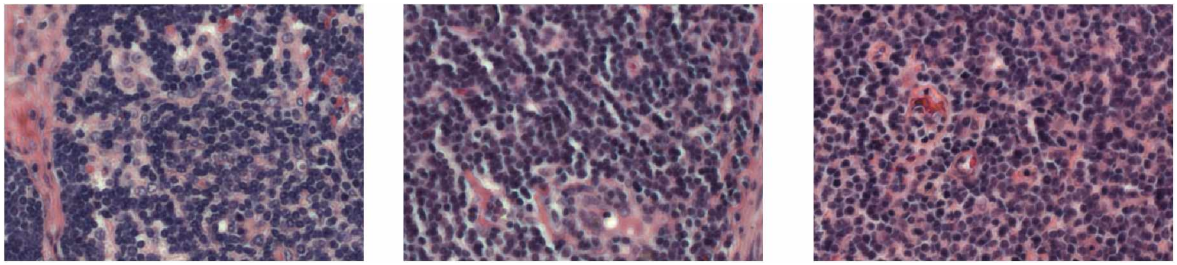
Neste capítulo, são apresentados os conceitos e técnicas utilizados na pesquisa, incluindo alguns tópicos da área de diagnóstico baseado em imagens histológicas, de sistemas de diagnóstico auxiliado por computador, de aprendizado de máquina e de inteligência artificial explicável. Finalmente, alguns trabalhos correlatos são exibidos.

2.1 Diagnóstico Baseado em Imagens Histológicas

Dentre os diversos métodos utilizados como suporte ao diagnóstico médico baseado em sistemas de imagens, a histologia é a área que pesquisa o uso de imagens microscópicas de cortes de tecidos possivelmente afetados (BENTAIEB; HAMARNEH, 2018). As amostras de tecido são geralmente coradas com Hematoxilina e Eosina (H&E) e submetidas à análise de um especialista por meio de um microscópio, o qual pode identificar características nas células observadas que indiquem a presença de alguma doença. Essas observações são essenciais para permitir o diagnóstico correto, a identificação do estágio da doença e a orientação do melhor tratamento para cada paciente (ORLOV et al., 2010). Esse tipo de imagem pode ser obtido de diversos tecidos. O presente estudo foca as investigações e avaliações sobre os tecidos histológicos de linfoma e displasia.

2.1.1 Linfomas

Linfoma é um dos tipos mais comuns de câncer registrados em humanos (NASCIMENTO et al., 2018). Essa variante da doença acomete células responsáveis pela defesa imunológica do organismo, chamadas linfócitos (ORLOV et al., 2010). De acordo com a Organização Mundial da Saúde, essa doença possui mais de 38 variantes, que podem ser divididas entre Linfomas Hodkin (LH) e Linfomas Não Hodkin (LNH). Sua diferenciação pode ser feita pela combinação de fatores morfológicos, imunofenotípicos, genéticos e clínicos. O LNH é o mais comum, representando cerca de 90% dos casos de linfoma, e já foi considerado o quinto tipo de câncer mais diagnosticado no Reino Unido (SHANKLAND;



(a) FL (b) CLL (c) MCL

Figura 1 – Regiões ampliadas de imagens de tecido linfático de cada tipo de linfoma não hodkin: a) Linfoma Folicular (FL), b) Leucemia Linfocítica Crônica (CLL) e c) Linfoma de Células do Manto (MCL).

ARMITAGE; HANCOCK, 2012). De acordo com estatísticas publicadas, 1.590 e 6.580 casos novos foram estimados no Brasil, respectivamente, para LH e LNH no ano de 2020 (BRAZIL, 2020). Nos Estados Unidos, foram estimados 8.480 novos casos de LH e 77.240 novos casos de LNH para o ano de 2020 (SOCIETY, 2020). Dentre os diversos tipos de LNH, estão o Linfoma de Células do Manto (do inglês: *mantle cell lymphoma* - MCL), o Linfoma Folicular (do inglês: *folicular lymphoma* - FL) e a Leucemia Linfocítica Crônica (do inglês: *chronic lymphocytic leukemia* - CLL). Esses subtipos mencionados são responsáveis por 85% dos casos de linfomas, o que torna seu estudo e investigação importantes tarefas no diagnóstico de LNH (SANTOS; FERNANDES, 2008).

A Figura 1(a) mostra uma pequena parte de uma imagem de tecido linfático do tipo FL ampliada. Essa neoplasia ocorre em células que revestem as cavidades dos linfonodos, chamadas centros foliculares, e se caracteriza por apresentar má definição, e baixa concentração de linfócitos B (TOSTA et al., 2016). Na Figura 1(b), é apresentado uma região de um tecido linfático que apresenta o linfoma do tipo CLL. Essa neoplasia é caracterizada pela proliferação de células linfóides que não se desenvolvem completamente, e portanto, não cumprem sua função imunológica. Ela pode ser caracterizada pela presença de linfócitos menores, com núcleos regulares e cromatina condensada, sem nucléolo visível. Além disso, menos de 10% dos prolinfócitos possuem tamanho médio, citoplasma abundante e nucléolo aparente (TOSTA et al., 2016). O exemplo apresentado na Figura 1-c apresenta neoplasias do tipo MCL, que se caracterizam por apresentar células um pouco maior que linfócitos saudáveis, com tamanhos parecidos entre si. Estas células também podem apresentar cromatinas aglutinadas, pouco citoplasma, nucléolos imperceptíveis e contorno nuclear geralmente irregular e clivado.

2.1.2 Displasia Oral Epitelial

A Displasia Oral Epitelial (DOE) é um tipo comum de lesão pré câncer (que pode originar um câncer), que pode ser diagnosticada realizando análises de imagens histológicas do tecido lesionado (ADEL et al., 2018). O termo displasia se refere a um crescimento

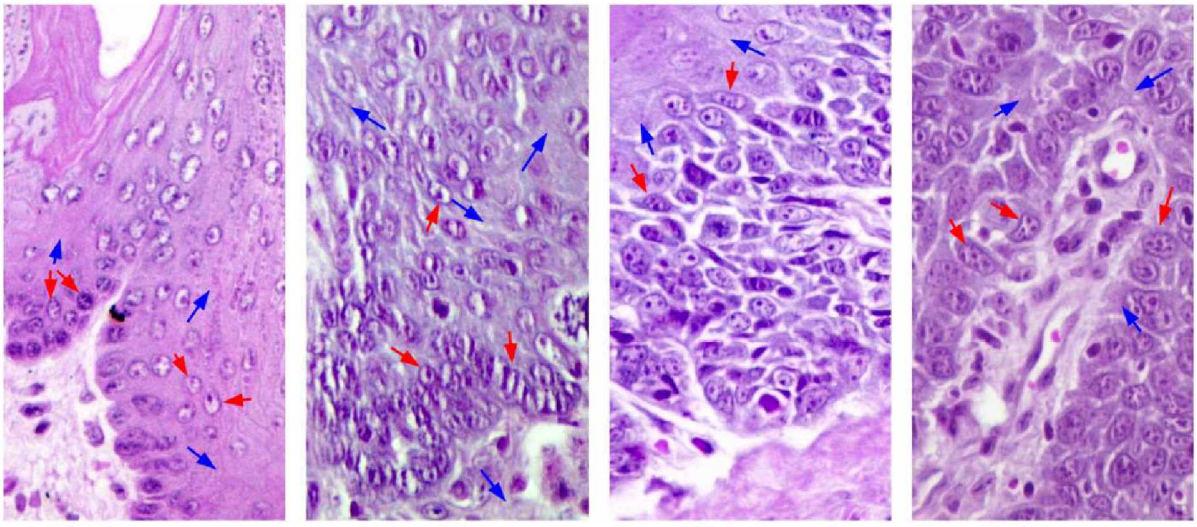


Figura 2 – Regiões ampliadas de imagens de tecidos histológicos orais que apresentam: (a) tecido saudável; (b) displasia leve; (c) displasia moderada; e (d) displasia severa. As setas vermelhas apontam para núcleos celulares e as azuis apontam para o epitélio celular. Adaptada de Silva et al. (2019).

desordenado e mudanças morfológicas de células afetadas, e quando ocorre em células epiteliais da cavidade bucal, é chamada de DOE (SILVA et al., 2019). Este tipo de lesão é caracterizado pela alteração de características morfológicas das células, como formato, tamanho e cor do núcleo celular (SILVA et al., 2019). A intensidade e frequência dessas alterações no tecido epitelial são utilizadas para classificar as lesões em leve, moderada e severa. Exemplos referentes aos diferentes estágios da doença estão representados na Figura 2. O risco de evolução da doença para um câncer, o que ocorre com probabilidade de 6% a 36% (SMITH et al., 2009), faz necessária sua identificação e seu diagnóstico precoce, já que a qualidade de vida e as chances de sobrevida do paciente aumentam se a lesão for tratada neste estágio (DOST et al., 2014). A histopatologia, estudo do impacto de uma doença em um tecido celular, é considerada o padrão ouro tanto para o diagnóstico da DOE quanto para a análise do risco de evolução da doença (DOST et al., 2014).

2.2 Diagnóstico apoiado por computador

Sistemas de diagnóstico apoiado por computador são comumente utilizados para auxiliar patologistas a analisar exames médicos. Tipicamente, eles são utilizados na histopatologia, a área que estuda doenças por meio da análise dos tecidos afetados. Neste casos, os sistemas CAD podem ser projetados para classificar imagens histológicas, de forma a prover ao patologista uma diagnóstico automático que o auxilia a analisar os casos de forma mais consistente (KAUSHAL et al., 2019). Esses sistemas podem ser divididos em cinco etapas: aquisição das imagens; pré-processamento; segmentação; extração de

atributos; e classificação. Para realizar a coleta das imagens, é necessário realizar alguma forma de coloração das amostras. Na histopatologia, é comumente utilizado para este fim, um protocolo baseado em hematoxilina e eosina. A hematoxilina é uma substância de coloração púrpura ou azul que se liga ao DNA das células, pintando, desta forma, o núcleo celular com esta cor. A eosina se prende a proteínas, atribuindo a cor rosa a várias estruturas ao redor do núcleo celular. Essa coloração provê uma distinção necessária para o reconhecimento dos núcleos celulares, como pode ser visto nas Figuras 1 e 2 (KAUSHAL et al., 2019). As imagens, então, são realizadas através de um microscópio e uma câmera digital. A etapa de pré-processamento das imagens consiste na aplicação de técnicas para diminuir ruídos, falhas e variações que podem ocorrer durante sua coloração e captura. Essas imperfeições são tratadas, por exemplo, com filtros, que destacam as estruturas contidas nas imagens, normalizações de cores, que reduzem o impacto da variação no processo de colorização da amostra, filtros de ruído e aumentadores de contraste (DEMIR; YENER, 2005). A segmentação da imagem refere-se ao processo de particiona-la em regiões não sobrepostas, separando objetos de interesse do fundo da imagem (NASCIMENTO et al., 2018). A extração de características tem como objetivo calcular atributos das imagens que têm pouca variância ligada a mudanças irrelevantes nas imagens. Os tipos mais comuns são baseados em fractais, texturas, intensidade, e morfologia das regiões de interesse (KAUSHAL et al., 2019). Por último, acontece a classificação dessas imagens, utilizando um modelo preditivo. Esta última etapa será explorada com mais detalhes na seção seguinte.

2.3 Algoritmos para classificação multiclasse

Em aprendizado de máquina, é chamado de classificador um sistema que, para um vetor de atributos x, determina uma entre K classes (AGGARWAL, 2015). Seguindo este conceito é possível observar uma separação do problema em duas instâncias, baseada no número de classes existentes. O caso mais simples, chamado de classificação binária, trata de apenas duas classes, enquanto classificação multiclasse é utilizado para se referir a problemas com um número maior de rótulos. Embora algumas técnicas permitam selecionar diretamente entre as diferentes classes de um problema (ex: redes neurais), muitos métodos tratam este problema do aumento no número de classes utilizando alguma forma de junção de modelos de classificação binária. O método chamado de um-contra-todos é utilizado no contexto de classificação para construir um modelo multiclasse utilizando K classificadores binários. Cada um deles é treinado na tarefa de diferenciar as amostras entre uma das K classes e o restante delas, de forma binária, sendo a classe escolhida tratada como o rótulo 1, e todas as outras como o rótulo 0. Para rotular uma amostra nova, utiliza-se uma comparação das predições dos K modelos, escolhendo a classe do modelo com o maior valor saída, ou seja, com maior probabilidade de classificação (BISHOP,

2006).

Neste trabalho, foram utilizados quatro algoritmos de predição e aqui são tratados as principais características teóricas dessas abordagens: *i*) *Gradient Boosting Decision Trees* (GBDT); *ii*) máquina de vetores de suporte (SVM); *iii*) regressão linear ; e *iv*) perceptron multicamadas (MLP). Foram escolhidos métodos baseadas em técnicas comumente empregadas em problemas de classificação e que possuem diferentes características e hiperplanos (FACELI et al., 2021). Por exemplo, modelos baseados em árvore de decisão e regressão linear costumam ser mais interpretáveis que aqueles construídos a partir de redes neurais e SVM. Por outro lado, redes neurais possibilita hiperplanos com áreas convexas, enquanto árvores de decisão geram hiperplanos retangulares e regressão linear e SVM projetam uma fronteira de separação entre 2 classes. Dessa forma, os algoritmos GBDT e MLP têm estratégias próprias para lidar com o problema de classificação multiclasse, mas para criar modelos compatíveis com este tipo de problema, a partir dos algoritmos SVM e regressão linear, foi necessário empregar a técnica de um-contra-todos. Neste capítulo, serão apresentados os funcionamentos básicos destas quatro técnicas de predição.

2.3.1 *Gradient Boosting Decision Trees* (GBDT)

Uma técnica de comitê de classificadores (*ensemble*) é uma agregação de vários modelos mais simples, que realiza uma decisão final em função das decisões individuais de cada modelo (pela média, ou votação por exemplo). Este tipo de técnica já se mostrou mais eficiente que a utilização de apenas um modelo em diversas aplicações (OZA; TUMER, 2008). Apesar dessa melhora, esses métodos costumam ter maior complexidade, o que dificulta a interpretação do modelo e de suas decisões de classificação.

Um dos modelos de aprendizado de máquina comumente utilizados em técnicas baseadas em comitês é a árvore de decisão. Trata-se de um dos algoritmos mais simples de aprendizado supervisionado, e também está dentre os de maior sucesso na literatura. Cada nó da árvore tem uma função de decisão baseada em apenas um atributo, e direciona o algoritmo a um de seus nós filhos, ou, se for um nó folha, classifica a amostra em uma das classes (RUSSELL; NORVIG, 2002). Um exemplo de árvore de decisão pode ser visto na Figura 3.

Os nós da árvore são construídos individualmente e sequencialmente, e uma vez definidas, suas regras de decisão não são alteradas. Para decidir qual atributo será utilizado, e qual regra será formada, o algoritmo é guiado por uma medida de desempenho baseada na impureza dos conjuntos de dados obtidos após sua aplicação. Esta avaliação é realizada com o conjunto de amostras de uma base de treino que chegaram a este nó ao percorrer a árvore. A função que calcula a impureza deve ser baseada na probabilidade de uma amostra pertencer a cada classe, considerando o conjunto de amostras presentes. A impureza assume o valor máximo se a amostra tiver a mesma probabilidade de pertencer a qualquer classe. Em contrapartida, essa métrica tende ao valor mínimo à medida que a

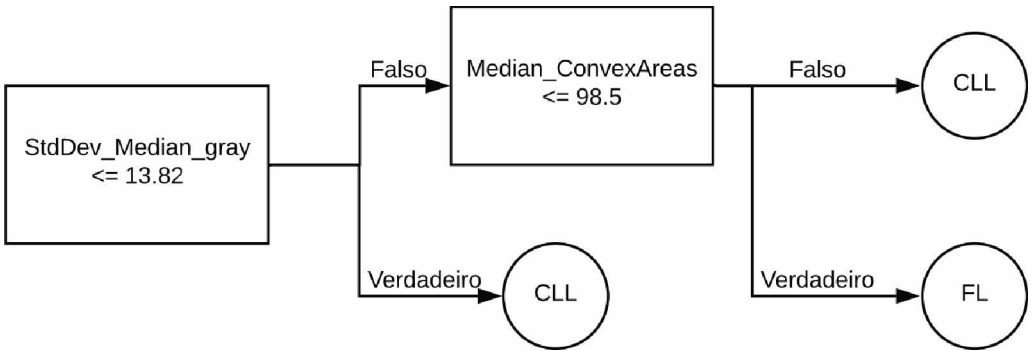


Figura 3 – Exemplo de árvore de decisão, utilizando classes de linfoma e os atributos propostos na metodologia deste trabalho.

probabilidade da amostra pertencer a uma determinada classe se aproximar de 1. O algoritmo de construção de árvores de decisão CART (do inglês, *classification and regression trees* - árvores de classificação e regressão) define a impureza com base na função Gini, definida pela Equação 1:

$$i(t) = 1 - \sum_i p_i^2, \tag{1}$$

sendo p_i a probabilidade de cada classe i. A avaliação de uma regra de corte é calculada pela média ponderada das impurezas de cada subconjunto de dados gerado pelo corte. Esta ponderação é baseada na quantidade de amostras em cada subconjunto.

O GBDT é um algoritmo baseado em comitês de árvores de decisão. Em um comitê, os modelos treinados precisam ser diferentes entre si. Para garantir variabilidade entre as árvores, cada uma é treinada a partir de um conjunto de dados distinto. Assim, os atributos escolhidos, e os valores de corte calculados são diferentes em cada caso. Cada árvore subsequente a primeira é adicionada ao modelo iterativamente de forma a diminuir seu erro. Para isso, o problema é modificado entre cada iteração, utilizando um conjunto de amostras com foco naquelas com maior chance de serem classificadas erroneamente pelo modelo (KE et al., 2017).

2.3.2 Maquina de vetores de suporte (SVM)

A máquina de vetores de suporte, do inglês *Support Vector Machines* (SVM) é um algoritmo de aprendizado de máquina baseado na teoria de aprendizado estatístico (SAIN, 1996). Essa área de estudo estabelece estratégias e princípios para a construção de modelos preditivos generalizáveis, que conseguem classificar corretamente objetos que não estavam presentes da base de dados de treinamento.

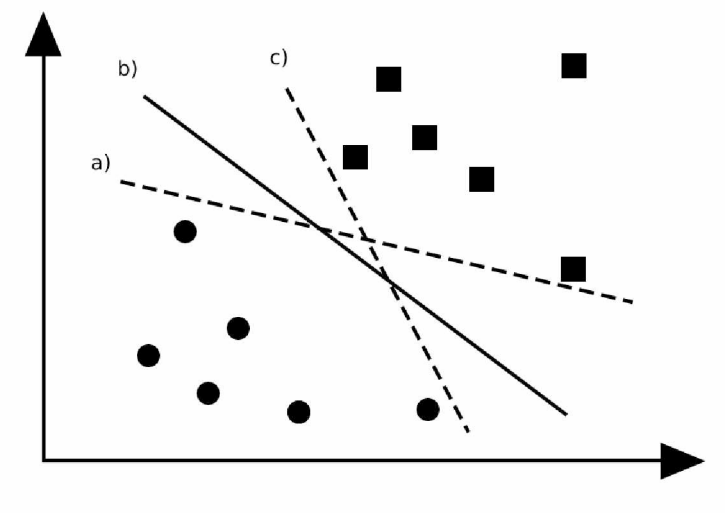


Figura 4 – Hiperplanos construídos por uma SVM para um problema hipotético linearmente separável, com duas classes representadas por círculos e quadrados. As linhas tracejadas a) e c) são hiperplanos com margens não ótimas; e a linha contínua b) é um hiperplano com margem ótima

Ele assume que a variável a ser predita pertence ao intervalo [-1, 1], e cria um hiperplano que funciona como uma borda de decisão entre duas classes (AGGARWAL, 2015). O método busca uma solução que maximize o tamanho da margem entre este hiperplano e qualquer amostra, e minimize a quantidade de amostras no lado errado da divisão.

O conceito de margem, ou borda utilizado pelo método se remete a distância entre o hiperplano de separação e as amostras mais próximas de cada classe. Isso é possível apenas em um cenário simples, em que as classes sejam linearmente separáveis, como na Figura 4. Na mesma imagem, é possível perceber como é possível criar, neste cenário, uma quantidade infinita de diferentes planos de separação entre as classes, como demonstrado pelas linhas tracejadas *a*) e *c*). Neste cenário, o método baseado em SVM define e utiliza a reta com a maior margem, representada pela linha contínua *b*), o que aumenta a capacidade de generalização do modelo, melhorando sua acurácia (AGGARWAL, 2015).

Entretanto, em problemas não linearmente separáveis, não é possível traçar um hiperplano que separe completamente os objetos das duas classes. Nestes casos, utiliza-se uma evolução do conceito de margem, chamada de margem suave, em que, amostras que se encontram do lado errado do hiperplano não o invalidam, mas contribuem com uma penalização na função de otimização, utilizando um parâmetro de regularização (AGGARWAL, 2015).

2.3.3 Regressão Linear

Classificadores baseados em regressão linear são formados por um ou mais deste modelo de regressão, combinados com funções de ativação, que transformam os resultados numéricos em rótulos binários (AGGARWAL, 2015). Uma regressão linear é uma função

da forma

$$y(x) = w_0 + \sum_{i=1}^{n} x_i w_i, \tag{2}$$

onde y é a variável predita, x é uma amostra, x_i é seu i-ézimo atributo, n é o número de atributos da amostra e w_i são os parâmetros do modelo, definidos na fase de treino. O parâmetro w_0 é geralmente chamado de viés, utilizado para deslocar a equação, e o restante dos parâmetros w_i são pesos relacionados aos atributos x_i.

Para definir estes parâmetros, pode ser empregado o método dos mínimos quadrados ordinários, que tem como objetivo minimizar a soma dos quadrados das diferenças entre os valores preditos e os valores reais de cada amostra utilizada no treino.

Para que esse método seja utilizado na criação de um classificador binário, uma função de ativação baseada em um limite é utilizada após a aplicação da função. Resultados com valores menores que $0,5$ são transformados em 0, enquanto valores iguais ou maiores que este, são transformados em 1. Para se adequar a técnica, as classes do problema também devem ser transformadas, utilizando os valores 0 e 1 para representar duas classes.

2.3.4 *Perceptron* multicamadas

Também chamado de rede neural (RUSSELL; NORVIG, 2002), o *perceptron* multicamadas (do inglês: *multilayer perceptron* - MLP) é inspirado no formato e comportamento de um sistema neural biológico. Ele é formado pela combinação de perceptrons, estruturas análogas ao neurônio biológico, e que funcionam de forma parecida a uma regressão linear, processando os valores de entrada através de uma função polinomial como vista na Equação 2, e uma função ativadora não linear (BISHOP, 2006).

Existem diversas funções de ativação, como a sigmoide, tangente hiperbólica, ou a Unidade Linear Retificada (do inglês: *Rectified Linear Unit* (ReLU)). Esta terceira foi utilizada na construção das MLP's presentes neste trabalho, já que é constantemente utilizada na literatura, para a classificação de displasias (SILVA et al., 2019) e linfomas (SOMARATNE et al., 2019; GANGULY; DAS; SETUA, 2020). A ReLU transforma os valores resultantes do processamento de *perceptron*, garantindo que o resultado seja não negativo, utilizando a Equação 3:

$$f(x) = max(0, x), \tag{3}$$

onde x é o resultado do *perceptron*.

Na Figura 5 é possível notar como uma MLP pode ser dividida entre diferentes tipos de camadas. Dois tipos são essenciais para a construção desse algoritmo, as camadas de entrada e de saída (na Figura 5, foi utilizada uma forma específica de camada de saída, construída para classificação, conhecida como camada de classificação) (HAYKIN, 2010). A primeira camada não realiza nenhum tipo de cálculo, e funciona apenas como uma

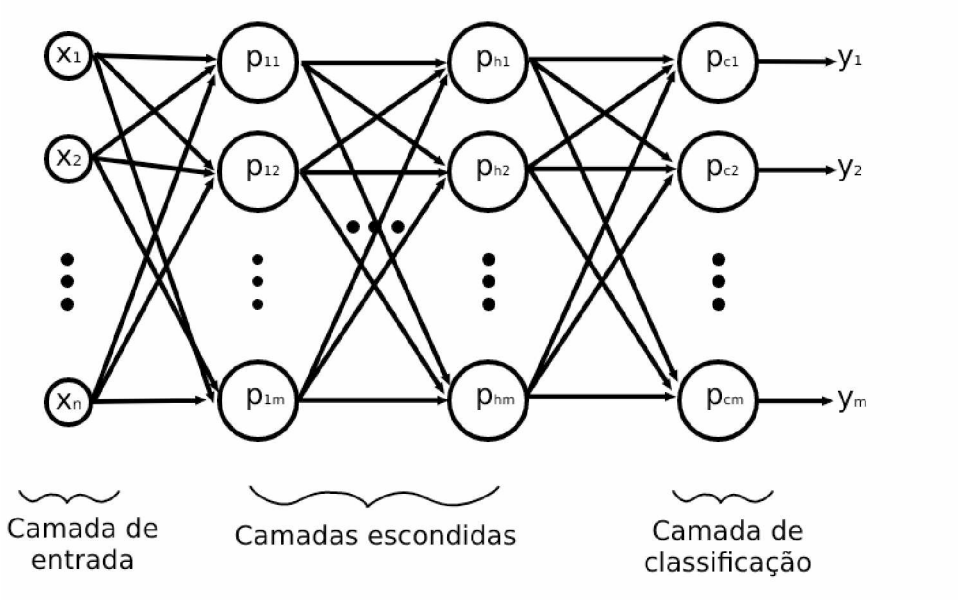


Figura 5 – Organização de um modelo de *perceptron* multicamadas projetado para classificação multiclasse.

padronização de como os dados devem ser introduzidos no sistema. Já a camada de saída, deve entregar os resultados do processamento da amostra, e geralmente se constituem de um neurônio para cada valor necessário na saída da rede. Entre essas duas camadas essenciais, podem haver um número variável de camadas escondidas, que também são constituídas de perceptrons, mas a sua quantidade não é limitada, nem pela quantidade de atributos, nem pela quantidade de valores de saída da rede (HAYKIN, 2010).

A camada de classificação recebe este nome por ser utilizada em problemas de classificação. Geralmente, em um cenário de classificação binária, esta camada é composta de apenas um *perceptron*, enquanto em um cenário multiclasse, utiliza-se um *perceptron* para cada classe. É comum que a função de ativação utilizada nesta última camada seja diferente das demais. Para uma camada de classificação de um problema multiclasse, pode-se utilizar uma função *softmax* (BISHOP, 2006), que transforma os m valores de saída em m probabilidades, que, quando somadas se igualam a 1. Para conseguir este resultado, aplica-se para cada classe, a Equação 4:

$$\sigma(c_m|x) = \frac{\exp(y_m)}{\sum_{i=1}^{m} \exp(y_i)}, \tag{4}$$

Em que $\sigma(c_m|x)$ estima a probabilidade da amostra x pertencer a classe c_m, exp() é a função exponencial de base e, y_m é o resultado obtido pelo *perceptron* correspondente a classe c_m, e y_i é o resultado do *perceptron* correspondente a classe c_i.

O treinamento desse modelo de aprendizado de máquina pode ser visto como um problema de otimização, no qual os pesos dos perceptrons devem ser ajustados para diminuir uma função de perda (HAYKIN, 2010). Também chamada de função de custo, ela representa, matematicamente, o quão próxima a resposta do modelo está da variável

alvo. Para classificação multiclasse, um exemplo comumente empregado é a perda de entropia cruzada, utilizada para comparar a distribuição de probabilidade das classes estimada pelo preditor, e as probabilidades reais da amostra (JANOCHA; CZARNECKI, 2017). Este calculo é realizado seguindo a Equação 5:

$$Entropia = -\sum_{i=1}^{m} \overline{y_i} \log(y_i), \tag{5}$$

sendo m a quantidade de classes; y_i a probabilidade prevista da amostra pertencer a classe i e $\overline{y_i}$ a probabilidade real da amostra pertencer a classe i.

Um método comumente utilizado para realizar o treinamento da rede é a retro-propagação (HAYKIN, 2010), que utiliza o gradiente da função erro como uma direção a um ponto de ótimo local. Derivando a equação de erro em função de um peso, é possível descobrir se o peso deve aumentar ou diminuir para que o erro diminua. Desta forma, o treinamento se da alterando os pesos da rede, somando ou diminuindo-os iterativamente, utilizando as amostras de uma base de treino para o calculo do erro. *Adam* é um algoritmo de retro-propagação que altera o valor dos pesos utilizando uma taxa de aprendizado multiplicada a um valor de momento, que se altera de acordo com os resultados consecutivos da função de erro. Se um peso é alterado múltiplas vezes para a mesma direção, o momento aumenta, e portanto, suas alterações seguintes são maiores, e caso essa direção mude, o momento diminui (KINGMA; BA,).

2.4 Inteligência Artificial Explicável

Técnicas de inteligência artificial tem sido utilizadas em uma quantidade crescente de sistemas de computação, com o objetivo de automatizar ou auxiliar seres humanos na realização de diversas tarefas. Essa amplitude de usos se estende a problemas críticos, que exigem que os usuários confiem no sistema, ou que tenham métodos que possibilitem auditar os resultados, como por exemplo, sistemas de automação de veículos, diagnóstico médico e serviços financeiros. Nesse contexto, ser capaz de prover uma explicação de como ou porque uma decisão foi feita, se tornou uma qualidade valiosa em sistemas de inteligência artificial (CONFALONIERI et al., 2021).

Essa área recebe o nome de inteligencia artificial explicável (do inglês: *explainable artificial inteligence* - XAI), o campo de estudo sobre esta propriedade. Nesta seção, serão introduzidos alguns conceitos básicos do tema; e alguns termos importantes utilizados na literatura; e por fim, serão apresentados dois métodos de XAI, que foram utilizados na construção da metodologia proposta neste trabalho.

2.4.1 Conceitos básicos

Uma inteligência artificial é definida como explicável se consegue formular justificativas ou prover informações complementares que ajudem uma audiência alvo a ter um melhor entendimento de seu funcionamento (LUNDBERG; LEE, 2017). Outro termo bastante utilizado ao se falar sobre XAI é a interpretabilidade, que apesar de correlacionados, são utilizados para se referir a conceitos diferentes dentro do tema. Um modelo é considerado interpretável se a sua arquitetura e funcionamento, assim como sua forma de armazenar o conhecimento aprendido são fáceis de serem entendidos pelo usuário, sem a necessidade de simplificações ou alterações no comportamento do modelo. Ao analisar um sistema interpretável, deve-se conseguir entender todo o processo de decisão, sem a necessidade de cálculos complexos ou ferramentas externas. Já um sistema explicável pode ser construído por um modelo de aprendizado de máquina não interpretável, mas utiliza de ferramentas e métodos para facilitar o entendimento. Nestes conceitos, a interpretabilidade é tratada como uma qualidade passiva, que é atribuída a um algoritmo de aprendizado de máquina pela forma como ele foi projetado, enquanto a explicabilidade é atingida de forma ativa, utilizando técnicas externas ao algoritmo preditor (ARRIETA et al., 2020).

Em metodologias propostas para problemas suficientemente complexos de aprendizado de máquina, é possível notar uma tendência da melhora dos resultados obtidos juntamente ao aumento da complexidade dos modelos preditivos utilizados (ARRIETA et al., 2020). Porém, modelos mais complexos também tendem a ser menos interpretáveis que alternativas mais simples. Modelos baseados na construção de uma árvore de decisão, por exemplo, podem criar estruturas simples o suficiente para que, lendo-a, seja possível entender seu processo de decisão, e inferir como o modelo se comportaria em diferentes ocasiões, utilizando apenas comparações entre valores. Já uma rede neural multicamadas pode utilizar uma quantidade tão grande de operações matemáticas, que inviabiliza este tipo de interpretação (ARRIETA et al., 2020).

Há vários motivos pelos quais alguém pode querer tornar um modelo preditivo explicável, mas entre eles, destacam-se o aumento da confiança do usuário nas predições do sistema; a capacidade do uso dessas explicações no processo de contestação de um resultado; a possibilidade de encontrar relações causais sobre o tema abordado ao investigar ligações entre variáveis encontradas pelas técnicas de XAI; sua habilidade de encontrar e destacar problemas de viés nos dados e modelos treinados, que podem evitar problemas éticos; e sua função em ajudar os desenvolvedores a construir modelos preditivos melhores, destacando problemas de viés e parcialidade, encontrando potenciais alvos de ataques adversários, e avaliando se apenas atributos verdadeiramente significativos sejam utilizados na decisão, garantindo que as relações entre variáveis utilizadas pelo modelo sejam baseadas em relações causais verdadeiras (ARRIETA et al., 2020).

2.4.2 Taxonomia

Há uma grande variedade de técnicas de XAI, que se diferenciam por qual parte do modelo preditivo é explicada e pela forma na qual as explicações são construídas. Com o objetivo de criar uma taxonomia comum às pesquisas do tema, uma gramática foi criada (BANIECKI; BIECEK, 2020), classificando os métodos segundo dois critérios: i) o que o método se propõe a explicar e ii) como o método explica.

O primeiro critério divide os métodos de explicação em três grupos com base na parte do modelo preditivo analisada. O primeiro, chamado de exploração dos dados, tem origens na área de análise exploratória de dados (TUKEY et al., 1977), e tem como objeto de explicação os dados utilizados no treinamento da metodologia. O segundo é a explicação global, que é criada para explicar o comportamento e a estrutura de um modelo treinado, destacando as informações com mais impacto no modelo, e as relações entre variáveis encontradas no treinamento. As explicações aditivas de Shapley (do inglês: SHapley Additive exPlanations - SHAP) (LUNDBERG; LEE, 2017) são um exemplo de método que pode realizar explicações globais. Por fim, há o grupo das explicações locais, que focam em explicar a predição de uma amostra específica. Este tipo de explicação pode ser utilizado para realizar testes mais aprofundados no modelo e para prover ao usuário uma justificativa para a decisão. Exemplos de métodos neste grupo incluem o SHAP (LUNDBERG; LEE, 2017), o método de âncoras (Anchors) (RIBEIRO; SINGH; GUESTRIN, 2018) e as explicações modelo-agnósticas locais e interpretáveis (do inglês: *Local Interperetable Model-agnostic Explanations* - LIME) (RIBEIRO; SINGH; GUESTRIN, 2016).

O segundo critério utilizado para agrupar métodos de XAI diz respeito a uma questão de como suas explicações são formadas. O primeiro grupo é formado por técnicas que utilizam análises de partes do modelo preditivo, principalmente quantificando a importância de componentes, como atributos ou grupos de atributos. Os modelos SHAP (LUNDBERG; LEE, 2017) e LIME (RIBEIRO; SINGH; GUESTRIN, 2016) são estratégias definidas nessa categoria. Outro grupo é formado pelos métodos que analisam o perfil da relação entre duas variáveis. Este método pode ser utilizado para analisar detalhadamente o comportamento de um modelo, e descobrir como cada atributo se relaciona com sua predição. Exemplos incluem a *Ceteris Paribus* (CP) (BIECEK; BURZYKOWSKI, 2021), e o mapeamento de dependência parcial (pdp) (MOLNAR, 2022). Por último, há o grupo de explicação que utiliza análises de distribuição dos dados (BANIECKI; BIECEK, 2020).

Além destes grupos, destaca-se também, como forma de caracterizar e agrupar os métodos de explicação, as formas como interagem com o modelo de predição treinado. Um algoritmo que pode criar explicações para qualquer tipo de modelo de predição, sem a necessidade de acessar os atributos e métodos internos do mesmo, é chamado de modelo-agnóstico, enquanto sua contraparte, projetada para explicar um algoritmo, ou um grupo de algoritmos específicos, é chamada de modelo-específico (ARRIETA et al., 2020). Por

Se	Então
StdDev_Areas <= 51.11 E Median_StdDev_b <= 15.36 E Avr_EquivDiameters > 11.64	Predição: CLL Precisão: 98.9% Cobertura: 0.02

Figura 6 – Exemplo de uma explicação fornecida pelo método *Anchors*, com a regra de decisão, o rótulo da predição, a precisão e a cobertura da explicação

não possuir acesso a estrutura interna do modelo, nem as suas informações gravadas, é comum que os algoritmos modelo-agnósticos façam predições com os modelos alvo de diversas amostras, tanto reais, quanto artificialmente criadas, e utilizem os resultados obtidos para entender o comportamento do modelo.

2.4.3 Métodos de explicação investigados

A metodologia proposta neste trabalho utiliza dois algoritmos de explicação modelo-agnósticos do estado da arte com diferentes abordagens. Nesta seção, são apresentados o funcionamento de cada um deles. Estes métodos foram escolhidos considerando sua extensa utilização no estado da arte, inclusive de trabalhos envolvendo sistemas CAD (LUNDBERG et al., 2018; WU et al., 2021; ELSHAWI; AL-MALLAH; SAKR, 2019; MA et al., 2020; ELSHAWI et al., 2021; YOO et al., 2020).

2.4.3.1 *Anchors*

Anchors (RIBEIRO; SINGH; GUESTRIN, 2018) é uma ferramenta de explicação local modelo-agnóstica. O método foca em oferecer ao usuário uma interpretação da decisão tomada para uma única amostra, e consegue fazer isso de forma transparente ao modelo preditivo utilizado. Como pode ser observado na Figura 6, a explicação consiste em uma regra de decisão, formada por um conjunto de predicados lógicos envolvendo os atributos utilizados na classificação. Também são informados para o usuário a precisão e a cobertura da regra na base de treino. Esta explicação é construída como um modelo de decisão simples, que funciona nas proximidades da amostra explicada x. As amostras para quais a regra de decisão retorna verdadeiro podem ser classificadas como da mesma classe de x, com a precisão estimada apresentada na explicação. A cobertura é definida como a probabilidade da âncora ser aplicada a uma amostra.

Para fornecer uma explicação a uma amostra x, o método realiza uma busca por uma âncora A (conjunto de regras de decisão), maximizando precisão e cobertura. A cobertura de uma âncora é formalmente definida como apresentado na Equação 6:

$$cov(A) = E_{D(z)}[A(z)], \tag{6}$$

sendo D uma base de amostras de validação. Já para calcular a precisão, o método cria um conjunto $D(z|A)$, que é gerado aplicando-se perturbações na amostra x. Para dados

tabulares, por exemplo, os autores em (RIBEIRO; SINGH; GUESTRIN, 2018) realizam essa perturbação mantendo os atributos de x que são utilizados na regra de decisão da âncora A e substituindo o restante pelos valores de uma amostra de D.

Dessa forma, dado um modelo de decisão f, uma amostra x, um conjunto de validação D, e um valor de precisão desejado τ, uma âncora é definida como um conjunto de regras de decisão que conseguem precisão maior ou igual a τ. Como apenas a estimativa da precisão pode ser calculada, ela deve ter probabilidade mínima de $1-\delta$. Dentre as âncoras que cumprem esta condição, o método busca por aquela com maior cobertura. Assim, o problema de explicação é definido como um problema de otimização combinatorial, que satisfaça a Equação 7:

$$\max_{As.t.P(prec(A)\geq\tau)\geq 1-\delta} cov(A). \tag{7}$$

Como a quantidade de combinações de regras de decisão cresce exponencialmente à quantidade de atributos, é impraticável avaliar todas as âncoras possíveis, então os autores em (RIBEIRO; SINGH; GUESTRIN, 2018) utilizam de um método heurístico para realizar esta busca, chamado de *Beam-Search*. O algoritmo se inicia com um conjunto de regras vazio, e iterativamente acrescenta regras até que se encontre âncoras com precisão suficiente. Nesse processo, em cada iteração, são selecionados os B conjuntos de regras com maior precisão, e estes serão utilizados na próxima iteração, onde novas regras serão inseridas nestes conjuntos e avaliadas seguindo o algoritmo KL-LUCB com múltiplos braços (KAUFMANN; KALYANAKRISHNAN, 2013). Ao atingir o valor mínimo de precisão, a âncora encontrada com maior cobertura é utilizada.

2.4.3.2 SHapley Additive exPlanations (SHAP)

SHAP (LUNDBERG; LEE, 2017) é um método de explicação modelo-agnóstico que pode ser utilizado de forma local ou global. As explicações são realizadas pelo cálculo da importância de cada atributo, seja em uma predição, no contexto local, ou em um conjunto de predições, no contexto global. O algoritmo é baseado no método de valores de Shapley, originário da teoria dos jogos cooperativos (NOWAK; RADZIK, 1994). Este método calcula a contribuição de cada um dos M jogadores para um objetivo em comum. Para utilizar a técnica em um contexto de aprendizado de máquina, cada atributo ou grupo de atributos é tratado como um jogador, e o valor de saída do método preditor é o resultado do jogo (MOLNAR, 2022). O valor de Shapley de um jogador é calculado pelo impacto médio observado ao adicioná-lo a todos os possíveis sub grupos dos outros $M-1$ jogadores, chamados de coalizões (MOLNAR, 2022).

O número de coalizões cresce exponencialmente com o aumento do número de jogadores, e com isso, o tempo necessário para o cálculo dos valores de Shapley segue o mesmo crescimento. SHAP tem algumas alternativas para aproximar esses valores gastando um tempo consideravelmente menor. *KernelSHAP* é um deles, que consegue fazer isso de forma modelo-agnóstica, e calcula os valores de impacto de cada atributo de tal

forma que eles podem ser utilizados como os pesos de um modelo de regressão linear que aproxima o comportamento do modelo preditivo na vizinhança da amostra explicada.

Esta vizinhança é formada utilizando o conceito de coalizões, do cálculo do valor de Shapley: Novas instâncias são criadas pela utilização de um sub-grupo dos atributos da amostra x explicada. Muitos algoritmos de predição não conseguem processar uma amostra com valores faltantes, então uma função h deve ser aplicada para preenchê-los. Um exemplo desta função é selecionar uma amostra aleatória da base de dados de treino e utilizar seus valores.

Cria-se então uma base de amostras vizinhas Z, e rotula-se elas utilizando a predição do modelo a ser explicado, $\hat{f}$. Este conjunto de amostras é utilizado para treinar um modelo de regressão linear ponderada g, uma forma de regressão linear em que, amostras com pesos maiores têm maior influência no calculo da função de perda. Dessa forma, a função linear obtida é obtida pela Equação 8:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi z'_j, \tag{8}$$

sendo z' uma instância derivada da amostra x, formada por uma coalizão dos seus atributos; M a quantidade de atributos; z'_j o j-ésimo atributo de z'; e ϕ_j o j-ésimo peso da função, e também, em caso de $j > 0$, a importância do atributo j para a predição de z.

Para calcular os pesos desta equação, otimiza-se a função de perda da Equação 9:

$$L(\hat{f}, g, \pi_x) = \sum_{z' \epsilon Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z'), \tag{9}$$

sendo π_x a função que calcula o peso de cada instância z'. A ideia utilizada para construir π_x foi a de que as coalizões que possuem uma quantidade de atributos próxima a zero ou próxima a quantidade máxima contribuem com mais informação sobre o valor dos atributos presentes ou faltantes respectivamente, do que instâncias com valores parecidos de atributos presentes e faltantes (MOLNAR, 2022). Então, pode ser definida pela Equação 10:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}, \tag{10}$$

sendo $|z'|$ a quantidade de atributos presentes na coalizão que originou z'.

2.5 Trabalhos Correlatos

Não foram encontrados trabalhos que utilizassem técnicas de XAI na criação de modelos de classificação de imagens histológicas de linfoma não Hodkin, ou de displasia oral. A pesquisa, então, foi dividida em três áreas, relacionadas ao tema proposto: i) A segmentação e a classificação de imagens histológicas de linfomas não Hodkin; ii) A segmentação de imagens histológicas de displasia oral, e a classificação destes dados; e iii) A aplicação de

técnicas de XAI para aumentar a interpretabilidade de sistemas CAD. Esta seção expõe as principais características dos trabalhos analisados de cada área, permitindo uma visão geral do estado da arte no momento da pesquisa.

2.5.1 Classificação de linfomas Não Hodgkin

A análise de descritores obtidos por técnicas computacionais de extração de características pode ajudar na descoberta de novos aspectos biológicos em tecidos com presença de câncer (BECK et al., 2011). Com a finalidade de prover uma ferramenta para auxiliar nas pesquisas de linfoma não Hodkin, os autores em (SHAMIR et al., 2008) propuseram uma base de dados com imagens microscópicas de pacientes com três tipos da doença: MCL, FL e CLL. Vários trabalhos recentes investigaram diferentes abordagens para obtenção de informação por meio de descritores e classificação dessas imagens.

Em (ORLOV et al., 2010), os autores propuseram utilizar descritores baseados em wavelets de Chebyshev e Fourier. Para classificação, foram avaliados um classificador baseado em distância de vizinhos balanceados (WND, *weighted neighbours distance*), um classificador baseado em funções de base radial e uma *Naive Bayes Network*. Em (NASCIMENTO et al., 2015) foi investigada a classificação dos linfomas baseado em características retiradas de transformadas de ondas estacionárias. Já em (SONG et al., 2016) e (SONG et al., 2017), os autores extraíram características morfológicas da base de dados. Em (BAI et al., 2019), os autores combinaram o uso de um classificador baseado em floresta aleatória, utilizando características formuladas a mão, e um classificador de imagens pré-treinado (*GoogLeNet*). Em (NANNIA; GHIDONI; BRAHNAM, 2020) e (ROBERTO et al., 2021), os autores propuseram uma metodologia utilizando ensemble de redes neurais convolucionais, combinando características obtidas através de técnicas de aprendizado profundo e descritores formulados a mão. Em (DIF; ELBERRICHI, 2020), uma combinação de métodos de aumento artificial dos dados, validação cruzada, uso de métodos pequenos e um ensemble de diferentes estados chaves do treino de uma rede de aprendizado profundo baseado na arquitetura *MobileNet2* foi utilizada para conseguir resultados com maior capacidade de generalização na classificação das imagens histológicas de linfoma. Uma característica em comum a estas técnicas é a falta de uma segmentação das imagens a fim de retirar informações do núcleo das células. Essas informações de regiões locais são relevantes para os especialistas numa investigação das características morfológicas e não morfológicas dos núcleos. O trabalho de (NASCIMENTO et al., 2018) utiliza de técnicas de visão computacional para fazer uma segmentação das imagens e detectar as regiões de núcleo das células, e propõem uma análise da diferença do uso de descritores morfológicos e não morfológicos, retiradas dos núcleos celulares das amostras, na sua classificação binária.

2.5.2 Classificação de imagens histológicas de displasia oral

Em (SILVA et al., 2019), os autores propuseram um método de segmentação e classificação de núcleos celulares de tecidos afetados por diversos graus de displasia oral. Um conjunto de 296 imagens foram coletadas de amostras de tecido da língua de camundongos. essas imagens foram classificadas entre tecido saudável, displasia leve, displasia moderada e displasia severa. Cada uma destas classes contém 74 imagens. O método de segmentação proposto utiliza a rede neural convolucional pré-treinada *ResNet50*. Os autores utilizaram um processo de transferência de conhecimento para atualizar os pesos desta rede utilizando uma base de imagens histológicas de displasia oral. Para estas imagens utilizadas na fase de treino, foram criadas, com a ajuda de um patologista, máscaras binárias que indicam a localidade dos núcleos celulares.

Após a segmentação, foram extraídas características morfológicas e não morfológicas dos núcleos celulares. As características morfológicas obtidas foram formadas pela área, excentricidade, orientação, perímetro e solidez. Já as características não morfológicos foram formadas pelo calculo da entropia, utilizando 7 tamanhos diferentes de vizinhança (3×3, 5×5, 7×7, 9×9, 11×11, 13×13 e 15×15 pixels) e pelo Índice de Moran. Após calculadas estas 13 medidas para cada núcleo celular, 26 descritores foram formados através das médias e desvios padrão dos núcleos de cada amostra.

Em (ADEL et al., 2018), foi proposta uma metodologia de classificação de displasias utilizando descritores baseados no algoritmo FAST orientado e BRIEF rotacionado, e um modelo de classificação baseado em SVM. O método obteve acurácia de 92,8% em experimentos utilizando um conjunto de 46 imagens histológicas coradas com H&E. Dentre este conjunto de imagens, apenas sete eram de pacientes saudáveis, e as outras 39 pertenciam a casos de displasia leve, moderada ou severa. Apesar de apresentar bons resultados, o método não realiza uma segmentação dos núcleos celulares, e por ter sido avaliado em um conjunto pequeno, e com quantidades desproporcionais de imagens de cada classe, pode ter resultados diferentes na classificação de classes pouco representadas. Em (SILVA et al., 2022), foram utilizados 23 descritores morfológicos e não morfológicos e um classificador polinomial para classificar imagens histológicas de DOE. O sistema obteve acurácia média de 92,4%.

2.5.3 Técnicas de Explicação Aplicadas na Classificação Multiclasse de Dados Médicos

Em um estudo comparativo (ELSHAWI et al., 2021), seis algoritmos de explicação local foram avaliados de acordo com sete métricas propostas. Os testes foram realizados com bases de dados médicas, com problemas de classificação tabular e problemas de classificação de linguagem natural. O método âncora (RIBEIRO; SINGH; GUESTRIN, 2018) obteve o melhor desempenho no quesito confiabilidade, uma medida relacionada

a concordância entre as explicações e o comportamento do modelo. O modelo LIME (RIBEIRO; SINGH; GUESTRIN, 2016) atingiu o maior grau de separabilidade. Esta métrica avalia se o método de explicação constrói explicações diferentes para amostras diferentes. Já o SHAP (LUNDBERG; LEE, 2017) foi a técnica com menor tempo de execução entre os algoritmos avaliados.

A técnica de explicação SHAP foi utilizada como uma forma de explicação global de modelos de apoio ao diagnóstico baseado em classificação de imagens histológicas em (BI et al., 2020; MA et al., 2020). Ambos os trabalhos utilizaram a técnica para calcular os 20 atributos mais importantes na classificação, e construíram um gráfico que mostra, para cada um destes atributos, a influência do atributo na classificação ao longo do espaço amostral. Ressalta-se que nenhum destes trabalhos trata de problemas multiclasse. Em (IRFAN; BASUKI; AZHAR, 2021) o método SHAP foi utilizado para listar os atributos mais impactantes em um modelo de classificação multiclasse. Já para explicar as predições, de forma local, o método proposto utilizou três explicações construídas pelo algoritmo LIME (RIBEIRO; SINGH; GUESTRIN, 2016), uma para cada classe, de forma um-contra-todos. As explicações do LIME mostram os atributos que tiveram mais impacto na predição, e indicam se aquele atributo contribui a favor ou contra a predição realizada.

Em (YOO et al., 2020), os autores apresentam uma metodologia para classificação multiclasse, com o objetivo de recomendar tratamentos de erros refrativos. O método recomenda entre três tipos de cirurgia ou indica a não realização deste tipo de tratamento. As informações utilizadas na predição são retiradas de um questionário com perguntas de múltipla escolha e de dados numéricos calculados em dois exames clínicos nos olhos do paciente. Para classificação, são treinados 11 modelos diferentes, todos utilizando o método de *gradient boosting* da biblioteca XGBoost: um modelo multiclasse, que indica uma das quatro classes possíveis; um modelo treinado de forma um-contra-todos para cada classe; e um modelo binário para cada par de classes existentes. A explicação de cada predição foi realizada através da técnica de explicação local SHAP (LUNDBERG; LEE, 2017), que indica quais descritores foram mais influentes na classificação de uma amostra. Em uma eventual predição, se os modelos não são unânimes na decisão, o resultado e o caso são encaminhados para um especialista para que uma análise mais detalhada seja realizada. Caso os modelos consigam decidir entre uma das cirurgias de forma unânime, o modelo a recomenda e realiza uma explicação utilizando SHAP. Como este algoritmo de explicação consegue ser aplicado diretamente apenas a problemas de duas classes, os autores utilizam várias explicações, utilizando os modelos binários e um-contra-todos. Primeiramente, utilizando o modelo preditivo do tipo um-contra-todos relativo a classe predita, a técnica explica porque aquela classe foi escolhida. Então, utilizando os modelos de classificação binária, foram formuladas explicações sobre o porquê a classe predita foi escolhida ao invés de cada uma das outras possíveis. Apesar de ter bons resultados,

e conseguir boas explicações para modelos multiclasse, esta técnica é custosa, já que a quantidade de modelos necessária pode se tornar impraticável com o aumento do número de classes.

2.5.4 Considerações Finais

Com a finalidade de promover o desenvolvimento de modelos capazes de auxiliar nesse diagnóstico, uma base de dados com imagens microscópicas de pacientes com três tipos de linfoma não Hodkin: MCL, FL e CLL foi apresentada em (SHAMIR et al., 2008). Vários trabalhos recentes investigaram diferentes abordagens para obtenção de informação por meio de descritores e classificação dessas imagens. No entanto, muitos deles não realizam a segmentação de imagens visando a extração de informações do núcleo da célula ou utilizam descritores complexos e que não são triviais ao médico (CODELLA et al., 2016; MARTINS et al., 2020; MENG et al., 2010). Em (NASCIMENTO et al., 2018), os autores exploram informações internas e externas dos núcleos das células, mas realizam apenas classificações binárias. Abordagens multiclasse também foram propostas na literatura. No entanto, geralmente são baseados em métodos "caixa preta", nos quais o conhecimento utilizado na tomada de decisão está implícito no modelo, como aprendizado profundo (BAI et al., 2019; NANNIA; GHIDONI; BRAHNAM, 2020) ou máquinas de vetores de suporte (SVM) (SONG et al., 2016; NASCIMENTO et al., 2015), o que torna difícil para o especialista interpretar o resultado.

Uma base de dados de imagens histológicas de tecido epitelial oral foi apresentada em (SILVA et al., 2020), contendo tanto amostras de tecido saudável quanto amostras contendo DOE, categorizados como leves, moderados ou avançados. O trabalho propôs uma metodologia para segmentação dos núcleos celulares nas imagens, porém não abordou a tarefa de classificação das lesões. Alguns trabalhos propuseram abordagens de segmentação e classificação das lesões (ADEL et al., 2018; SILVA et al., 2022), porém todos utilizaram métodos do tipo "caixa preta", o que dificulta sua interpretação pelo especialista.

Há diversas metodologias para classificação multiclasse de LNH e DOE com alta acurácia publicadas. Porém, não foi encontrado trabalhos que integrem métodos de explicação nos sistemas CAD propostos, e poucos deles utilizam descritores interpretáveis. Entre os trabalhos encontrados que tratam da explicação de sistemas de classificação de imagens histológicas, poucos deles utilizam modelos multiclasse. Na maioria das vezes, métodos propostos para este problema utilizaram alguma combinação de múltiplas explicações locais para prover a interpretação de decisões, mostrando a viabilidade deste tipo de solução. Desta forma, percebe-se a necessidade de abordar este assunto e propor uma metodologia que integre as ferramentas necessárias para criar modelos interpretáveis que classifiquem de forma multiclasse LNH e DOE.

Capítulo 3

Metodologia Proposta

A utilização de técnicas de visão computacional e reconhecimento de padrões pode possibilitar a criação de funcionalidades importantes em sistemas CAD, como as etapas de detecção e obtenção de regiões de interesse assim como a representação para um diagnóstico mais efetivo dos pacientes (DOI, 2007).

Grande parte dos trabalhos do estado da arte que tratam destas funcionalidades foram realizados em um contexto cuja a métrica principal era obtida por meio de acurácia ou área sob a curva ROC. Por não terem um foco em explicabilidade, é comum que esses métodos utilizem de técnicas de extração de características e tomada de decisão que sejam pouco interpretáveis por seus usuários (ROBERTO et al., 2017; MENG et al., 2010; SONG et al., 2016; CODELLA et al., 2016; MARTINS et al., 2020; BAI et al., 2019; NASCIMENTO et al., 2015).

Considerando a necessidade de manter sistemas CAD interpretáveis e transparentes aos usuários, neste trabalho é proposta uma metodologia de classificação multiclasse de imagens médicas que integra diferentes técnicas de XAI, oferecendo explicações de diversas partes do sistema preditivo. Esses algoritmos são categorizados em três grupos, que se diferem de acordo com o foco de suas explicações (BANIECKI; BIECEK, 2020): i) métodos de análise de dados são utilizados para investigar a distribuição de diferentes atributos e suas relações entre si e com a variável alvo (classe predita); ii) métodos de explicação global são utilizados para entender o comportamento do método preditivo quando aplicado ao problema proposto, encontrando quais atributos mais influenciam na classificação, e como o valor destes atributos reflete nas decisões do modelo; e iii) métodos de explicações locais são utilizados para prover explicação para uma decisão pontual (um caso específico) do modelo preditivo, possibilitando que um profissional de saúde possa utilizar essas informações como suporte para sua própria decisão sobre o caso.

A Figura 7 apresenta um diagrama geral da metodologia proposta neste trabalho. Ele mostra as etapas necessárias para a construção do modelo preditivo e indica em qual delas cada tipo de explicação atua. Primeiramente, as imagens histológicas são segmentadas, um processo que em que foram demarcados as regiões de núcleo das células.

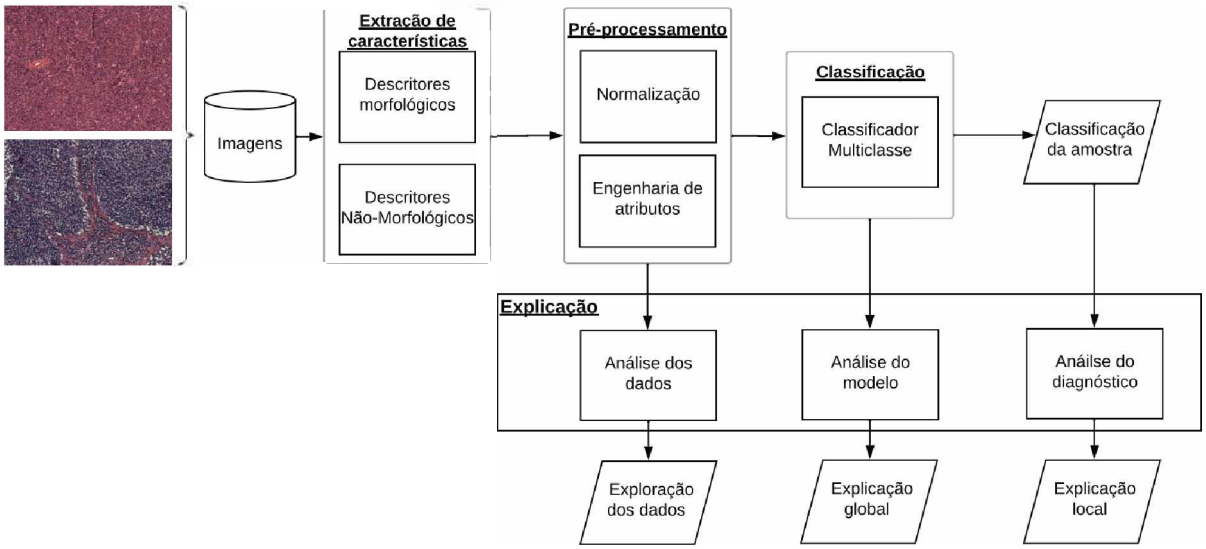


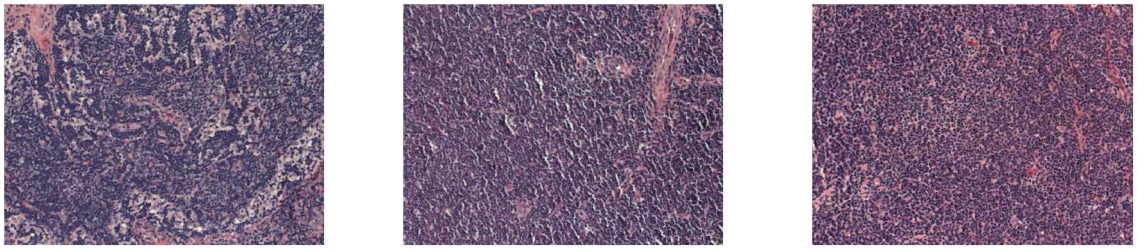
Figura 7 – Visão geral da metodologia proposta para a classificação interpretável de imagens histológicas.

Essas regiões foram utilizadas para extração das características. Nesse trabalho foram extraídos 34 descritores morfológicos e 80 não-morfológicos. Na etapa de engenharia de atributos, foram removidos os atributos com pouca variância ou que tenham alta correlação com outros atributos da base. Em seguida, é realizada a normalização dos vetores de atributos e dos dados, de modo que todos os atributos sejam representados na mesma ordem de grandeza. Por fim, um modelo preditivo baseado em redes neurais multi-camadas é utilizado para classificar a amostra desejada. O diagrama também mostra como cada método de explicação interage com este sistema, indicando que as explicações dos dados são realizadas através de análises sobre a engenharia de atributos, a explicação global dos algoritmos realizadas por meio de uma investigação do modelo preditivo e as explicações locais são feitas analisando cada amostra individualmente. Detalhes sobre o funcionamento de cada uma destas etapas, assim como justificativas para o uso das técnicas serão expostos nas seguintes sessões deste capítulo.

3.1 Lesões Histológicas Investigadas e Obtenção dos Descritores

Para realizar experimentos e avaliar o desempenho dos métodos propostos, foram escolhidas duas bases de imagens histológicas. Uma base de amostras de tecido linfático de pacientes com linfoma, separados em três tipos de linfoma não-hodkin; e uma base de amostras de tecido da língua de 30 ratos com diferentes graus de displasia. Ambos são tecidos histológicos corados com hematoxilina e eosina e avaliados por um especialistas na etapa de separação dos núcleos.

A fim de obter dados que quantifique os núcleos presentes nas imagens histológicas,



(a) FL (b) CLL (c) MCL

Figura 8 – Exemplo de imagem de tecido linfático de cada tipo de linfoma não hodkin: (a) Linfoma Folicular (FL), (b) Leucemia Linfocítica Crônica (CLL) e (c) Linfoma de Células do Manto (MCL).

métodos de segmentação foram empregados para localização e separação de suas estruturas. Nessa etapa foram empregados os métodos propostos por Nascimento et al. (2018) para os núcleos dos tecidos de linfomas e Silva et al. (2019) para os núcleos de displasia de tecido oral. Esses algoritmos foram já avaliados e validados no contexto dos trabalhos já publicados na literatura.

Para quantificar as características dos núcleos celulares por meio de descritores numéricos, foram aplicados métodos para extração de características morfológicas e não-morfológicas, criadas a partir de informações de forma e intensidade de brilho dos canais de cores, conforme proposto pelos autores em Nascimento et al. (2018). Para as diferentes bases de imagens foram empregados os mesmos descritores para representação das propriedades dos núcleos.

3.1.1 Bancos de Imagens Histológicas

A base de dados dos diversos grupos de lesões do linfoma foi apresentada em (SHAMIR et al., 2008). As imagens digitalizadas foram obtidas a partir de 30 amostras histológicas de nódulos de linfomas coradas com H&E, contendo 10 de cada grupo de linfoma (CLL, MCL e FL). A partir dessas amostras, foi obtido um conjunto composto por 375 imagens com resolução de 1388 × 1040 pixels, com quantização de 24 bits em modelo de cores RGB. Foi utilizado um microscópio (Zeiss Axioscope) com ampliação de 20 vezes e uma câmera digital colorida (AXio Cam MR5). Das imagens coletadas, 113 são de linfomas do tipo CLL, 140 são FL e 122 são MCL. Na Figura 8 é apresentada uma imagem digitalizada de cada grupo de lesão investigado de linfoma.

A segunda base histológica foi obtida de imagens da cavidade oral de lesões de displasia. Essa base foi elaborada para estudo de algoritmos de segmentação baseado em abordagens de aprendizagem profunda (SILVA et al., 2019). Diferente da base de linfomas, essa base possui um grupo de imagens de tecido saudável, além de três níveis de gravidade da doença, sendo catalogadas em displasia leve, moderada e severa. Estas imagens foram retiradas de 43 amostras da língua de camundongos. A Figura 9 mostra

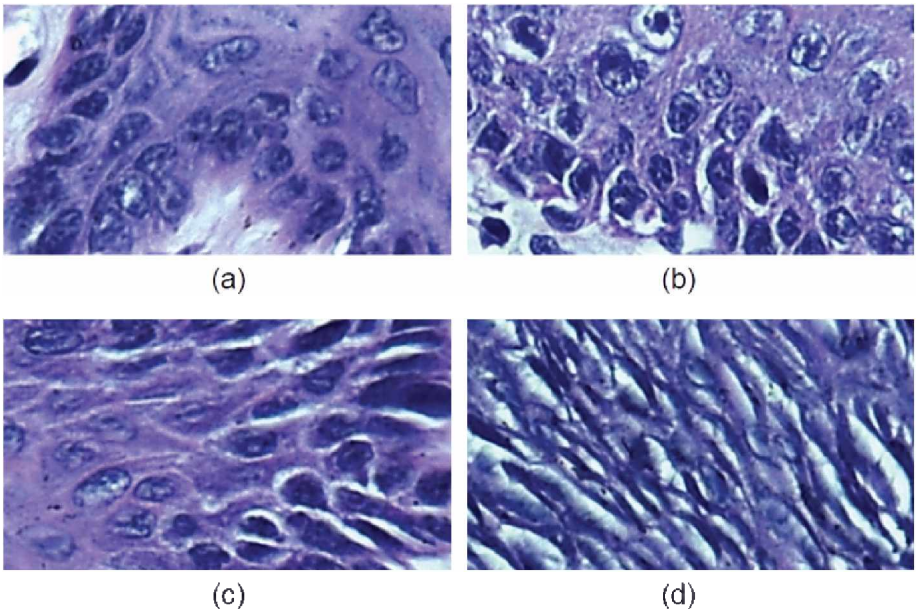


Figura 9 – Exemplos das diferentes classes de imagens histológicas da cavidade oral presentes na base de dados. (a) Tecido saudável, (b) Displasia Leve, (c) Displasia Moderada e d) Displasia Severa. Adaptado de (SILVA et al., 2019)

um exemplo de cada uma dessas classes. Todas as amostras foram coradas com H&E, e imagens foram obtidas por um microscópio óptico Leica DM500, em magnificação de 400× utilizando o esquema de cores RGB, com quantização de 64 bits e resolução de 2048 × 1536 pixels. Das imagens coletadas, 296 regiões de interesse de tamanho 450 × 250 pixels foram recortadas para compor a base de dados em que cada classe foi definida com 74 amostras.

3.1.2 Segmentação dos núcleos nas imagens histológicas

Para cada base de dados, foi utilizado uma metodologia de segmentação específica, que já foi avaliada previamente com os mesmos dados. Para segmentar as imagens de linfoma, foi utilizado uma técnica baseada em filtros de realce, equalização por histograma e preenchimento por inundação (NASCIMENTO et al., 2018), já as imagens de displasia foram segmentadas utilizando técnicas de aprendizado profundo (SILVA et al., 2019).

Para a segmentação das imagens de linfoma foi empregado a técnica apresentada em Nascimento et al. (2018). Esse método usa uma equalização adaptativa de histograma com contraste limitado (CLAHE - *contrast limited adaptive histogram equalisation*) para melhorar o contraste das imagens. Isso é alcançado ajustando a escala dos valores de intensidade de luz com base em seu histograma, diminuindo a homogeneidade de partes da imagem sem aumentar o impacto de ruídos. Após esta etapa, um filtro de realce foi aplicado para remover ruídos. Este filtro muda o valor de um pixel para o menor valor encontrado em sua vizinhança. Nesta metodologia, a vizinhança foi definida como uma

janela de tamanho 3×3, com o pixel alvo localizado no seu centro.

Para criar uma máscara binária que diferencia pixels dos núcleos celulares de pixels do fundo da imagem, foi utilizado um limiar para a intensidade do pixel com valor igual a 90. Pixels com intensidade maior que este limiar são marcados como pertencentes aos núcleos celulares. Para complementar este método, foi utilizado uma técnica de preenchimento com quatro vizinhos conectados para localizar pixels da região de interesse que não foram detectados pelo limiar. Para separar os núcleos presentes dentro da região demarcada, o método de preenchimento é utilizado novamente, porém agora com 8 vizinhos conectados. Após este processo, uma etapa de pós-processamento foi aplicada em que regiões com menos de 30 pixels são consideradas como ruído e excluídas da máscara. Por fim, cada componente conectado recebe um identificador único, para que análises de núcleos individuais possam ser realizadas (NASCIMENTO et al., 2018).

A segmentação das imagens da base de displasia foi realizada com uma metodologia baseada em aprendizado profundo (SILVA et al., 2019). O método constrói uma rede neural convolucional baseada em Mask R-CNN para segmentação das regiões da imagem (HE et al., 2017). A abordagem proposta por Silva et al. (2019) é composta por três etapas, e foi utilizada para obter as máscaras binárias que indicam a localização dos núcleos celulares. A primeira etapa utiliza de uma rede ResNet50 (HE et al., 2016) para identificar objetos presentes na imagem, criando regiões candidatas a serem classificadas como núcleos celulares. A segunda utiliza uma janela deslizante 3×3 pixels para criar um vetor de características que é usado para determinar quais objetos são núcleos celulares, criando caixas delimitadoras para cada um. Na terceira etapa, uma rede totalmente convolucional (LONG; SHELHAMER; DARRELL, 2015) e uma etapa de pós-processamento baseado em operações morfológicas são utilizadas para determinar, para cada região de interesse demarcada, uma mascara binária, indicando a localização do núcleo.

A rede foi pré-treinada com a base da dados de imagens ImageNet, que contém milhões de amostras demarcadas de diversos temas, incluindo imagens histológicas. Uma segunda etapa de treinamento é realizada para refinar os pesos da rede, utilizando dez imagens de cada classe da base de dados de displasia, demarcadas por especialistas.

Essas máscaras passaram, também, por uma etapa de pós-processamento, onde uma operação morfológica de dilatação foi utilizada para completar o contorno dos núcleos e uma operação de preenchimento foi utilizada para eliminar regiões de falsos positivos dentro de núcleos. Um filtro de erosão também foi utilizado para retirar ruídos e retornar os núcleos dilatados ao seu tamanho original. Finalmente, objetos com tamanho menor que 30 *pixels* foram considerados ruídos, e excluídos.

3.1.3 Extração de características

Com base nos núcleos das células definidos, foram extraídos descritores morfológicos e não morfológicos. As características morfológicas foram criadas a partir de informações

geométricas dos núcleos. Para cada núcleo, foram calculadas nove métricas: área, extensão, perímetro, área convexa, solidez, excentricidade, diâmetro equivalente, eixo menor e eixo maior, conforme explorado no estudo de Nascimento et al. (2018). As características dessas métricas são:

Área: essa métrica é calculada pela contagem do número de *pixels* dentro da região demarcada correspondente ao núcleo da célula, destacada em vermelho na Figura 10(a).

Extensão: é a fração de *pixels* dentro da caixa delimitadora da região demarcada que pertencem ao núcleo da célula, como exemplificado na Figura 10(b), onde a região em vermelho representa a área do núcleo e o retângulo em azul representa a área da caixa delimitadora. Esta medida pode ser calculada pela Equação 11:

$$Extensão = \frac{Área}{Área\_da\_caixa\_delimitadora}. \tag{11}$$

Perímetro: corresponde ao comprimento da borda da região demarcada, como ilustrado na Figura 10(c). Essa métrica é calculada a partir da soma das distâncias entre cada par de *pixels* adjacentes.

Área Convexa: é a soma das áreas de cavidades convexas presentes na região demarcada, como esta representada em vermelho na Figura 10(d).

Solidez: é a relação entre os pixels da área do núcleo e aqueles que estão nas áreas convexas, a qual é calculada pela Equação 12:

$$Solidez = \frac{Área}{Área\_Convexa}. \tag{12}$$

Eixos menor e maior: correspondem, respectivamente, aos tamanhos do menor e maior eixos da elipse que tem a mesmo segundo momento central normalizado da região demarcada, como ilustrado pelas retas vermelhas nas Figuras 10(e) (eixo menor) e 10(f) (eixo maior).

Excentricidade: esta métrica pode ser interpretada como a proximidade da elipse que tem o mesmo segundo momento da região demarcada com um círculo. É calculada a partir da Equação 13, onde *df* representa a distância entre os focos da elipse que tem o mesmo segundo momento da região demarcada:

$$Excentricidade = \frac{df}{Eixo\_maior}. \tag{13}$$

Diâmetro equivalente: essa medida mede o diâmetro de um circulo com a mesma área que a região demarcada. Essa métrica pode ser calculada pela Equação 14:

$$Diâmetro\_Equivalente = \sqrt{\frac{4 \cdot Área}{\pi}}. \tag{14}$$

Para realização dos experimentos foram utilizados abreviações das palavras em inglês para representação de cada uma das características morfológicas. O significado de cada uma delas esta apresentado na Tabela 1.

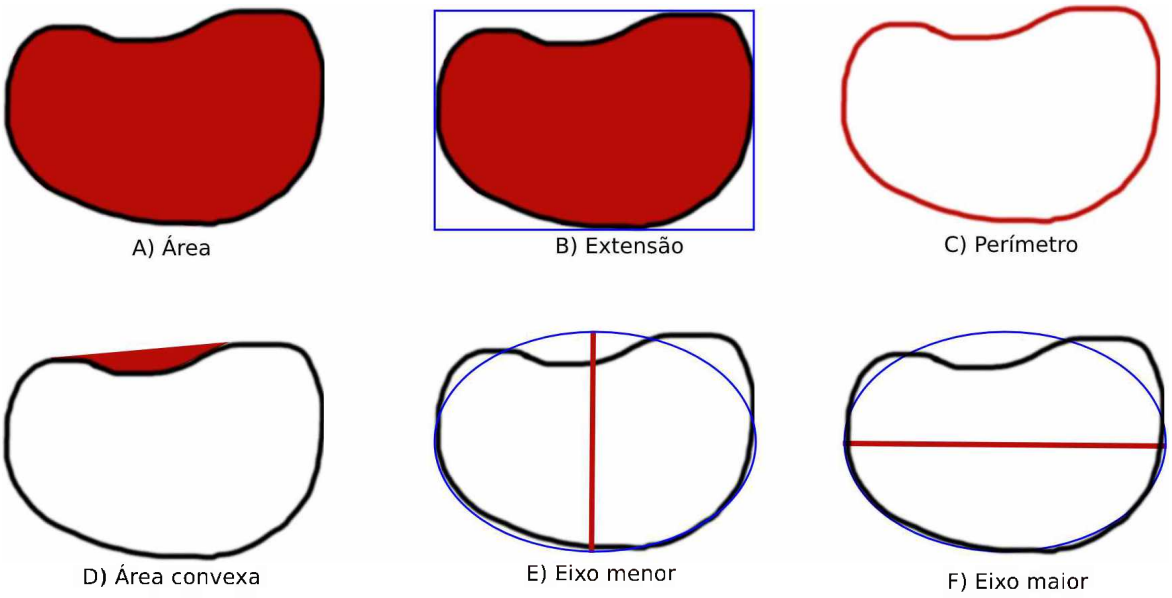


Figura 10 – Representação visual de características morfológicas. (a) Área;(b) Extensão; (c) Perímetro; (d) Área convexa; (e) Eixo menor; (f) Eixo maior

Tabela 1 – Uma relação dos termos utilizados para se referir a cada tipo de característica morfológica durante os experimentos.

Termos utilizados	**Características morfológicas**
Areas	Área
Extents	Extensão
Perimeters	Perímetro
ConvexAreas	Área convexa
Solidity s	Solidez
Eccentricity s	Excentricidade
EquivDiameters	Diâmetro equivalente
MinorAxisLengths	Eixo menor
MajorAxisLengths	Eixo maior

Após a realização destes cálculos para cada núcleo das células, medidas estatísticas são calculadas sobre os valores obtidos em toda a imagem. De cada métrica, são calculados as médias (*Avr*), medianas (*Median*), modas (*Mode*) e desvios padrão (*StdDev*). Ao calcular estas quatro medidas estatísticas de cada uma das nove métricas morfológicas, são gerados 36 descritores morfológicos. Neste trabalho, os descritores são referenciados de acordo com a medida estatística e a métrica utilizadas no seu cálculo. Por exemplo, a média das áreas convexas dos núcleos celulares é referida como *Avr_ConvexAreas*.

As características não morfológicas foram criadas a partir das informações de níveis de intensidade de brilho dos *pixels* nos três canais RGB (verde, vermelho e azul) e em escala

de cinza. De cada um dos quatro canais, foram calculadas as médias (*Avr*), medianas (*Median*), desvios padrão (*StdDev*), valor mínimo (*Min*) e valor máximo (*Max*) de intensidade de brilho dentro de cada núcleo. Utilizamos também *r, g, b* e *gray* para nos referir aos canais de cor vermelho, verde, azul e de tons de cinza, respectivamente. A partir das informações de cada núcleo, também foram calculadas as médias, medianas, desvios padrão e modas de cada canal da imagem. Combinando os quatro cálculos estatísticas utilizados com cinco tipos de medidas não-morfológicas e os quatro canais de cores, foram obtidos 80 descritores. Esses são também referenciados pela combinação dos termos abreviados. Por exemplo, a média do valor de pixel mínimo no canal azul é referenciado como *Avr_min_b*.

3.2 Normalização dos dados

Atributos geralmente têm escalas diferentes entre si. Essa diferença pode fazer com que os atributos de grande escala tenham um impacto maior nas decisões do modelo do que os de menor escala (AGGARWAL, 2015). Uma maneira comumente usada de lidar com esse problema é dimensionar os dados com base em seu desvio padrão médio. Este método é chamado de padronização e pode ser calculado de acordo com a Equação 15:

$$z_j^i = \frac{x_j^i - \mu_j}{\sigma_j}, \tag{15}$$

onde x_j^i e z_j^i são, respectivamente, os valores originais e normalizados do atributo j da amostra i; e μ_j e σ_j são a média e o desvio padrão do atributo j no conjunto de dados de treino. Esse método foi utilizado pois os dados de ambas as bases (linfoma e displasia) seguem uma distribuição normal (Gaussiana), a qual é apropriada para esse tipo de normalização. Para verificar se os atributos seguem essa distribuição, foi utilizado o teste de *Shapiro-Wilk* (SHAPIRO; WILK, 1965), com 95% de confiança.

3.3 Engenharia de Atributos

Uma importante etapa na construção de um modelo preditivo é a análise e remoção de atributos desnecessários. Alguns atributos em uma base de dados podem ser considerados irrelevantes ou redundantes para o problema (AGGARWAL, 2015). Atributos que não tenham variância alguma entre as amostras da base de treino, ou seja, que assumem o mesmo valor em todas as ocasiões, independente de sua classe, podem ser descartados sem prejudicar o processo de classificação, já que estes não provêm informação relevante sobre a amostra (AGGARWAL, 2015). Em uma base de dados também pode haver grupo de atributos altamente correlacionados, ou seja, que apresentam uma relação direta ou inversamente proporcional entre seus valores em grande parte das amostras. Nesses casos,

apenas um dos atributos já pode ser suficiente para representar o comportamento de todo o grupo, enquanto os demais possuem informações redundantes para o problema (BIESIADA; DUCH, 2007). Identificar e remover esses atributos irrelevantes pode acarretar em benefícios, como reduzir a dimensionalidade dos dados, e auxiliar no treinamento do modelo preditivo, uma vez que os atributos removidos poderiam atrapalhar na construção dos hiperplanos que separam as classes.

3.3.1 Remoção de atributos irrelevantes

Nesse trabalho, atributos com nenhuma (zero) variância entre as amostras da base de treino são removidos da base de dados durante a etapa de engenharia de atributos. A variância (S^2) de um atributo é calculado segundo a Equação 16:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}, \tag{16}$$

sendo x_i o valor do atributo na amostra i; $\bar{x}$ a média do valor do atributo na base de treino; e n a quantidade de amostras. Na prática, atributos com variância zero possuem o mesmo valor para todas as amostras do conjunto, o que o torna descartável para a tarefa de diferenciar amostras de classes distintas (AGGARWAL, 2015). Por este motivo, optou-se por criar um filtro em nossa base de dados que retira estes atributos. Apesar da importância da aplicação do filtro de uma forma geral, nos experimentos realizados neste trabalho, não houve nenhum atributo com variância nula identificado.

3.3.2 Filtragem de atributos por correlação de Pearson

O coeficiente de correlação de Pearson (ρ) é uma forma de medir a correlação linear entre dois conjuntos de dados. Para duas populações dadas por A e B, ele pode ser calculado como descrito na Equação 17 (BENESTY; CHEN; HUANG, 2008; RODGERS; NICEWANDER, 1988):

$$\rho_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B}, \tag{17}$$

sendo $cov(A,B)$ a covariância de A e B; e σ_A e σ_B o desvio padrão de A e B, respectivamente. Este coeficiente tem um valor entre -1 e 1. Como pode ser visto na Figura 11, um coeficiente de Pearson próximo a 0 indica pouca correlação entre A e B, enquanto um valor próximo a 1 ou -1 indica alta correlação. Um coeficiente positivo indica que valores de A crescem quando valores de B crescem (relação diretamente proporcional), e um coeficiente negativo indica que valores de A diminuem quando valores de B crescem (relação inversamente proporcional).

Para realizar uma filtragem de atributos redundantes, nosso método utiliza a medida de correlação de Pearson e um limite de correlação alfa (α) para determinar quando os atributos são altamente correlacionados. Os passos para a remoção dos atributos redundantes são descritos no Algoritmo 1.

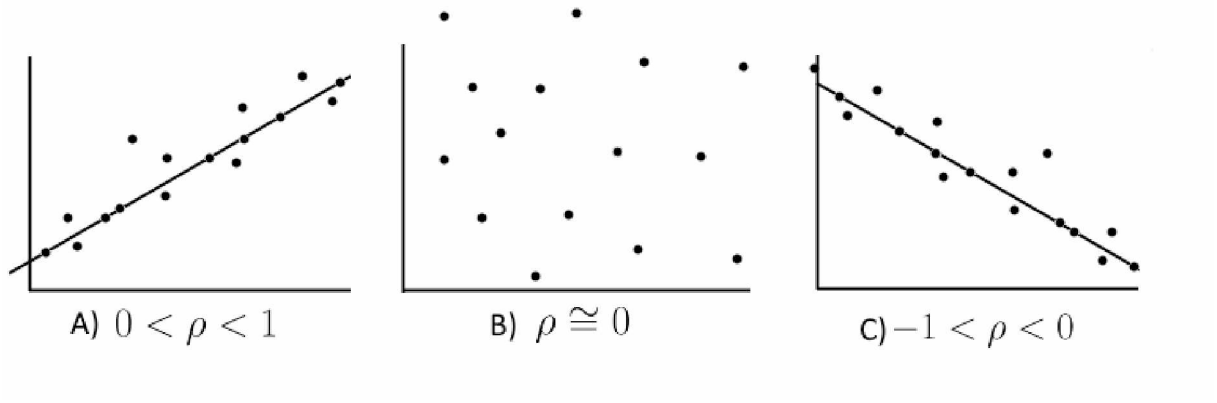


Figura 11 – Demonstração da relação entre dois atributos com diferentes níveis de coeficiente de correlação de Pearson. (a) e (c) Atributos correlacionados e; (b) Atributos pouco ou não correlacionados.

Algoritmo 1 Algoritmo para Filtragem de atributos por correlação de Pearson

```
ρ ← matriz de correlação de Pearson
C ← lista de atributos
C' ← conjunto vazio
while C não estiver vazio do
  D ← conjunto vazio
  remove x qualquer de C e adiciona-o em D
  for x' em C do
    if ρ[x][x'] > α then
      remove x' de C e adiciona-o em D
    end if
  end for

  C' ← elemento de D com maior correlação com a variável alvo
end while
C ← C'
```

O algoritmo recebe como entrada um conjunto C formado por todos os atributos disponíveis na base de dados e o valor de α. A adoção de um valor baixo para esse parâmetro provocará a redução excessiva do conjunto de atributos considerados no treinamento, podendo provocar a remoção de descritores relevantes para a classificação. Por outro lado, um α elevado afeta a efetividade da remoção, mantendo atributos redundantes no conjunto de treinamento. Neste trabalho foi empregado $\alpha = 0{,}99$. Utilizando este parâmetro, o método retirou, em média, 33,5 atributos da base de imagens de linfoma e 14,4 da base de imagens de displasia, durante os experimentos realizados.

O processo inicia com o cálculo da correlação ρ entre todos os pares de atributo da base de dados e entre cada atributo e a classe das amostras. Também é criado um conjunto C', inicialmente vazio. A cada passo iterativo, remove-se um atributo x qualquer e todos aqueles altamente correlacionados com x ($|\rho| > \alpha$) do conjunto C. Dentre os atributos

removidos, escolhe-se aquele com maior correlação com a classe alvo, o qual é incluído no conjunto C'. Esse processo se repete até que o conjunto C esteja vazio. Por fim, o algoritmo retorna o conjunto C', o qual é formado apenas pelos atributos com baixa correlação ($|\rho| \leq \alpha$) entre si, ou seja, sem redundância entre eles.

Vale destacar que, embora seja utilizada a correlação de Pearson na remoção de atributos redundantes, outros métodos também podem ser adotados sem afetar as demais etapas da metodologia.

3.4 Classificação

Um algoritmo de classificação é um método supervisionado que precisa de uma base de dados rotulada para construir uma função que aproxima a relação entre os atributos e o rótulo das amostras, de forma que, dada uma amostra nova, ele consiga estimar sua classe (AGGARWAL, 2015). A metodologia proposta neste trabalho foi construída de tal forma que a técnica utilizada para classificação das amostras pode ser substituída sem a necessidade de grandes adaptações. Para garantir esta flexibilidade, foram utilizados métodos de explicação *post-hoc*, que são agnósticas quanto ao modelo (RIBEIRO; SINGH; GUESTRIN, 2016). Nos experimentos, foram utilizados métodos baseados em ensemble de árvores de decisão (GBDT), máquina de vetores de suporte (SVM), regressão linear (RL) e redes neurais (MLP), os quais foram previamente descritos na Fundamentação Teórica (Seção 2.3).

Para a execução dos experimentos propostos neste trabalho, utilizamos a implementação do GBDT chamada LightGBM, presente no pacote LGBM. Os resultados obtidos pelos autores desta versão indicam que ela consome menos recurso para treinar o modelo do que a GBDT original (KE et al., 2017). Foi utilizada a parametrização padrão fornecida pelo pacote, para que a metodologia seja generalizável para diferentes casos de uso. Estes parâmetros definem a criação de 100 árvores para cada modelo, e o uso de no máximo 31 folhas em cada árvore. Para a construção dos modelos restantes, foram utilizadas implementações disponíveis no pacote *scikit-learn* (PEDREGOSA et al., 2011). Para os modelos baseados em RL e SVM, o modelo de decisão multiclasse foi construído na forma um-contra-todos. Para o SVM, foi utilizado o *kernel* linear, e os parâmetros padrão de sua implementação (*LinearSVC*), que incluem o uso da norma l2 para penalização, a função de perda quadrado da perda de articulação (do inglês *hinge loss*) e um valor de regularização de 1. Também foram utilizados os parâmetros padrão para a criação da MLP, criando uma rede com uma camada escondida com 100 neurônios, utilizando a função de ativação *relu* e o otimizador *adam*, com taxa de aprendizado fixa de 0,001, e um número máximo de 200 épocas. O treinamento é interrompido se a melhora na função de perda é menor que 0,0001 durante 10 épocas.

3.4.1 Método de estimação da confiança das predições

O grau de confiança da classificação pode ser uma importante informação para algumas aplicações, principalmente em um sistema CAD. Um profissional da saúde pode utilizar essa informação para confirmar a decisão do sistema ou verificar a necessidade de uma avaliação humana ou algum outro exame complementar para definir o diagnóstico. Alguns classificadores provêm um valor numérico relativo a cada rótulo, indicando a propensão da amostra pertencer a cada classe, enquanto outros apenas indicam o rótulo predito para cada instancia (AGGARWAL, 2015).

A fim de contornar esse problema, foi desenvolvida uma abordagem capaz de estimar a confiança para as predições multiclasses de qualquer classificador adotado. Essa abordagem é baseada no trabalho de (YOO et al., 2020), que utiliza um classificador multiclasse e um conjunto de classificadores binários, formado a partir de análises um-contra-um e um-contra-todos. Ao fazer uma predição, o sistema compara a decisão do algoritmo multi-classe com os binários, e caso não haja uma decisão unânime, ele considera a classificação inconclusiva e recomenda uma análise por um especialista humano. A nossa proposta desenvolve esta ideia adicionando mais granularidade em sua resposta, categorizando o nível de confiança da predição multiclasse em três grupos: confiável, incerta ou inconclusiva. Também, por não utilizar os modelos um-contra-todos, são necessários menos classificadores binários, diminuindo assim o custo computacional da técnica.

A Figura 12 mostra o funcionamento geral da abordagem proposta para estimar o nível de confiança na decisão do sistema. A ideia básica do método original é mantida, ou seja, determinar a confiança com base na comparação entre as predições dos classificadores multiclasse e binários. Entretanto, neste trabalho são adotados apenas classificadores binários construídos a partir das análises um-contra-um. Para cada par de classes do problema, é criado um classificador binário baseado no mesmo algoritmo de classificação que o modelo multi-classe sendo utilizado. A técnica é aplicada após uma amostra ser classificada pelo classificador multiclasse com o rótulo y. Então, é realizada a predição da mesma amostra utilizando todos os modelos binários treinados para classificar a classe y. Por exemplo, considerando que a base de treinamento possui 4 classes distintas (x, y, z e w), como o classificador multiclasse diagnosticou a amostra como sendo da classe y, a análise considerará apenas os classificadores binários treinados para decidir entre (y, x), (y, z) e (y, w). Se todos esses modelos binários concordarem com a decisão do modelo multiclasse, ou seja, rotularem a amostra como y, a decisão é qualificada como confiável. Se todos os modelos binários discordarem do multiclasse, indicamos que a predição do sistema é inconclusiva. Já se apenas um subgrupo de modelos discordar, concluímos que há uma decisão incerta, e é possível indicar o subgrupo de rótulos que foi predito pelo menos uma vez, a fim de auxiliar na análise do especialista humano.

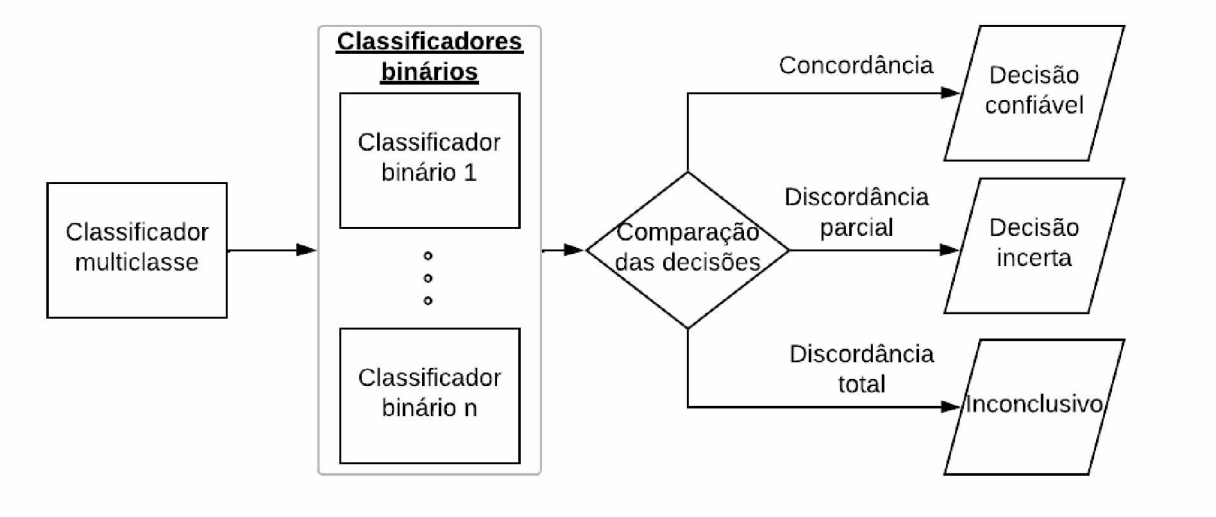


Figura 12 – Método para estimação da confiança das predições multiclasse com base em classificadores binários.

3.5 Explicação

A fim de fornecer informações adicionais para apoiar a compreensão das decisões do classificador, a metodologia deste trabalho prevê três níveis de análises para auxiliar na explicação das decisões do sistema CAD: análise dos dados; do modelo (explicação global); e do diagnóstico (explicação local). Esses níveis são definidos de acordo com o alvo da explicação (BANIECKI; BIECEK, 2020). As três etapas são modelo-agnósticas, ou seja, podem ser aplicadas independente da técnica de classificação utilizada.

Uma importante propriedade de qualquer explicação é ser facilmente interpretada pelo seu publico alvo. Algoritmos de explicação tentam apresentar as informações de forma mais amigável ao usuário. Por exemplo, o uso de regras de associação do tipo SE-ENTÃO facilitam o entendimento pelos seres humanos dos critérios usados nas decisões de um modelo (RUSSELL; NORVIG, 2002). Entretanto, se a quantidade de premissas for elevada, a interpretação da regra se torna difícil. O uso de transformações dos dados também dificulta o entendimento da explicação. Por exemplo, se ao invés de serem usados os valores originais de área ou perímetro das células, fossem considerados os valores transformados pelo processo de normalização dos dados, a explicação poderia ser mais difícil de ser entendida, já que esta diferença de valores se distancia da amostra observada pelo usuário. Por isso, houve uma atenção na escolha de métodos de transformação que pudessem ser revertidos facilmente antes de apresentar as explicações aos usuários. A normalização dos dados é revertida, recuperando os valores originais da amostra por meio de uma adequação da Equação 15, dada por:

$$x_j^i = (z_j^i \sigma_j) + \mu_j, \tag{18}$$

sendo x_j^i e z_j^i, respectivamente, os valores originais e normalizados do atributo j da amostra i; e μ_j e σ_j a média e o desvio padrão do atributo j no conjunto de dados de treino.

3.5.1 Análise de dados

Uma análise de dados geralmente utiliza de técnicas de analise exploratória de dados (EDA, do inglês *Exploratory Data Analysis*), e visa extrair informações sobre as distribuições e as relações entre pares de atributos contidos na base de dados utilizada na construção do modelo (BANIECKI; BIECEK, 2020). Este tipo de explicação pode ajudar os responsáveis pelo desenvolvimento e manutenção da ferramenta a entender melhor os dados disponíveis, podendo contribuir para a construção de um modelo mais apropriado (BANIECKI; BIECEK, 2020). Por exemplo, as informações acerca das relações entre pares de atributos possibilitam aos usuários averiguar se há falsas correlações ou, ainda, algum viés introduzido por uma coleta de dados inadequada. Esta transparência confere maior confiança e segurança ao utilizar o modelo (ARRIETA et al., 2020).

Neste trabalho, utilizamos o coeficiente de correlação de Pearson para realizar uma filtragem na base de dados e retirar atributos altamente correlatos, como descrito na Seção 3.3.2. Esta mesma técnica é utilizada na análise de dados para calcular a matriz de correlação entre os atributos da base de dados. A partir dessa matriz, é possível mostrar os grupos de atributos com alta correlação entre si, como pode ser visto na Tabela 2 e uma lista com os atributos de maior correlação com a classe das amostras, como apresentado na Figura 13. Os valores e grupos apresentados nas figuras presentes nesta seção são apenas exemplos de como esse tipo de informação é apresentado, e foram obtidos utilizando o caso de uso de linfomas.

Para criar os grupos de atributos correlacionados, uma lista dos atributos disponíveis na base de dados é analisada. Para cada atributo, identifica-se todos aqueles que possuem uma alta correlação com ele, de acordo com o limite de correlação adotado ($\alpha = 0{,}97$) e forma-se um grupo de atributos correlatos, e então, os elementos encontrados são retirados da lista. Os atributos que não têm uma correlação alta com nenhum outro, não são exibidos na tabela de visualização do método.

3.5.2 Explicação global baseada na análise do modelo

O foco da explicação global é entender o comportamento de um modelo em um certo conjunto de dados. Ela é comumente utilizada para avaliar se o modelo funciona como o esperado, promovendo confiança na utilização do sistema (BANIECKI; BIECEK, 2020; ARRIETA et al., 2020). Para prover este tipo de explicação, nossa metodologia faz uso do algoritmo SHAP (LUNDBERG; LEE, 2017; LUNDBERG et al., 2020), um método constantemente utilizado na literatura, inclusive em trabalhos relacionados a área medica (WU et al., 2021; ELSHAWI; AL-MALLAH; SAKR, 2019; MA et al., 2020) e que obteve bons resultados em um estudo comparativo com diversos outros métodos de XAI (ELSHAWI et al., 2021). Para gerar uma explicação global do modelo, o método realiza a explicação local das amostras da base de treino, calculando os valores aproximados de

Tabela 2 – Exemplo de uma análise de dados gerada a partir da base de linfoma, mostrando os grupos de atributos correlatos de acordo com o coeficiente de correlação de Pearson (ρ) e considerando um limite de correlação $\alpha = 0{,}97$.

Avr_Min_g Avr_Median_g Avr_Min_gray Avr_Median_gray Median_Min_g Median_Min_gray Mode_Min_g Mode_Avr_g Mode_Median_g		
Avr_Extents Avr_Solidity s	Avr_Median_r Median_Min_r Mode_Median_r	Avr_Max_g Median_Max_gray Mode_Max_g Mode_Max_gray
Avr_Areas Avr_MajorAxisLengths	Median_Perimeters Median_ConvexAreas Median_MajorAxisLengths	Avr_Min_b Avr_Median_b Median_Min_b Median_Median_b Mode_Min_b
Avr_Max_b Mode_Max_b	Median_Max_r Mode_Max_r	Median_StdDev_g Median_StdDev_gray
StdDev_Areas StdDev_Perimeters StdDev_EquivDiameters	StdDev_Extents StdDev_Solidity s	StdDev_Min_g StdDev_Min_gray
StdDev_Max_g StdDev_Max_gray	StdDev_Median_g StdDev_Median_gray	StdDev_StdDev_g StdDev_StdDev_gray
Mode_Min_r Mode_Min_gray	Mode_Avr_gray Mode_Median_gray	

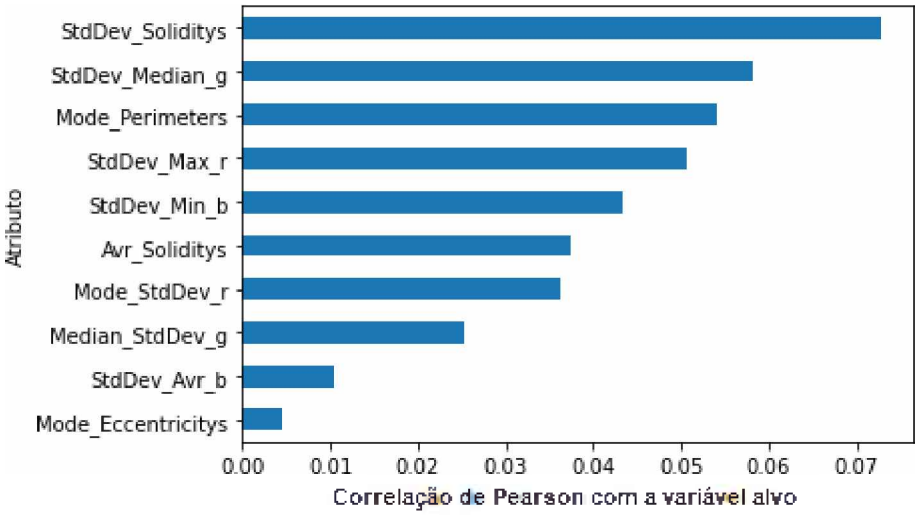


Figura 13 – Exemplo de uma análise de dados gerada a partir da base de linfoma, mostrando uma lista ordenada dos 10 atributos com maior correlação com a variável alvo do problema (atributo rotulado).

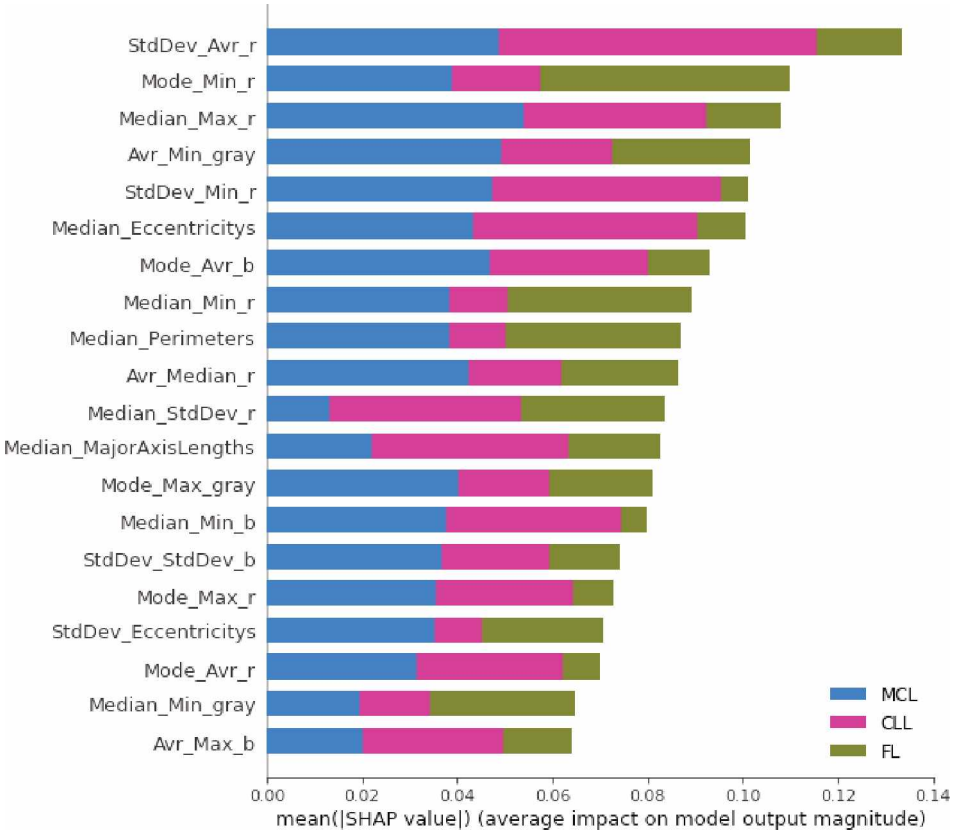


Figura 14 – Exemplo de uma explicação global com a técnica SHAP

shapley de cada atributo. Então, o impacto global de um atributo é definido como a média aritmética desses valores. O resultado é uma lista de atributos ordenada de acordo com o impacto/influência de cada um no comportamento do modelo, conforme o exemplo apresentado na Figura 14. Um valor numérico entre 0 e 1, chamado de *Shapley*, também é atribuído, como uma unidade de grandeza dessa avaliação, o que é mostrado por meio das diferentes corres nas barras do gráfico da Figura 14. Tal informação possibilita uma análise mais detalhada do funcionamento do modelo de predição (LUNDBERG et al., 2020).

Para realizar o cálculo do valor de *Shapley* de cada atributo, o SHAP realiza uma avaliação aproximada do desempenho do modelo sem considerar um grupo de atributos, e então calcula o impacto da adição de um atributo nesta situação. Como não é viável um novo treinamento do modelo a cada iteração e o método deve funcionar pra qualquer algoritmo de predição, não é possível a retirada efetiva de um atributo da amostra. Em vez disso, o atributo assume valores escolhidos de modo a simular o efeito de sua retirada da amostra. Em nossa metodologia, o valor do atributo é substituído por diferentes pontos do espaço amostral de treinamento, os quais são obtidos por um método baseado no algoritmo de agrupamento K-médias (AGGARWAL, 2015), como descrito no Algoritmo 2.

Basicamente, o algoritmo agrupa as amostras em K grupos e usa seus centróides para substituir o valor do atributo desejado. Neste trabalho, o valor de K corresponde a 15%

Algoritmo 2 Algoritmo para aproximação do impacto médio de um atributo em uma amostra com um conjunto de atributos retirados

$A \leftarrow$ amostra
$ATR \leftarrow$ atributo do qual o impacto será calculado
$C \leftarrow$ centróides de k grupos, formados pelo K-médias, dos dados de treino
$M \leftarrow$ quantidade de amostras na base de treino
$SUM \leftarrow 0$
for c em C **do**
  $A' \leftarrow A$, porém com atributos substituídos pelos de c
  $PRED \leftarrow$ predição de A'
  $PRED' \leftarrow$ predição de A' com atributo ATR adicionado
  $IMPACTO \leftarrow PRED' - PRED$
  $m \leftarrow$ quantidade de amostras no grupo relativo a centroide c
  $SUM \leftarrow SUM + IMPACTO * m$
end for
$IMPACTO\_MEDIO \leftarrow SUM/M$

dos dados da base de treinamento, como recomendado pela documentação do pacote que implementa o método de explicação. O SHAP é executado para cada centróide gerado e o valor de *Shapley* do atributo é calculado a partir da média ponderada das K execuções. O número de amostras em cada grupo define o peso associado ao seu centróide na média ponderada (LUNDBERG; LEE, 2017; LUNDBERG et al., 2020).

Existem diferentes heurísticas para o cálculo dos valores de *Shapley*. Algumas fazem o uso de características inerentes a certos tipos de modelo de aprendizado de máquina para aumentar sua velocidade de execução, enquanto outras se mantém agnósticas ao modelo de decisão (LUNDBERG; LEE, 2017). Neste trabalho, é apresentada uma metodologia capaz de lidar com diferentes algoritmos de classificação multiclasse. Então foi adotado o algoritmo Kernel SHAP (LUNDBERG; LEE, 2017), baseado no método de explicação local LIME (RIBEIRO; SINGH; GUESTRIN, 2016), que utiliza uma regressão linear ponderada para aproximar os valores de *Shapley*.

3.5.3 Explicação local baseada na análise da predição

Enquanto explicações globais focam em elucidar algum aspecto do funcionamento geral do modelo de aprendizado de máquina, explicações locais tentam esclarecer o processo utilizado pelo modelo preditivo na decisão sobre uma amostra específica (ADADI; BERRADA, 2018). Há diferentes maneiras de se construir uma explicação (ARRIETA et al., 2020): a) a abordagem SHAP (LUNDBERG; LEE, 2017) calcula a influência de cada atributo na decisão, de forma que estes valores também podem ser interpretados como pesos de uma regressão linear para a área próxima a amostra; b) na abordagem âncora (Anchors) (RIBEIRO; SINGH; GUESTRIN, 2018) também há a criação de um modelo de decisão simplificado que funciona nas proximidades da amostra, porém ela é composta

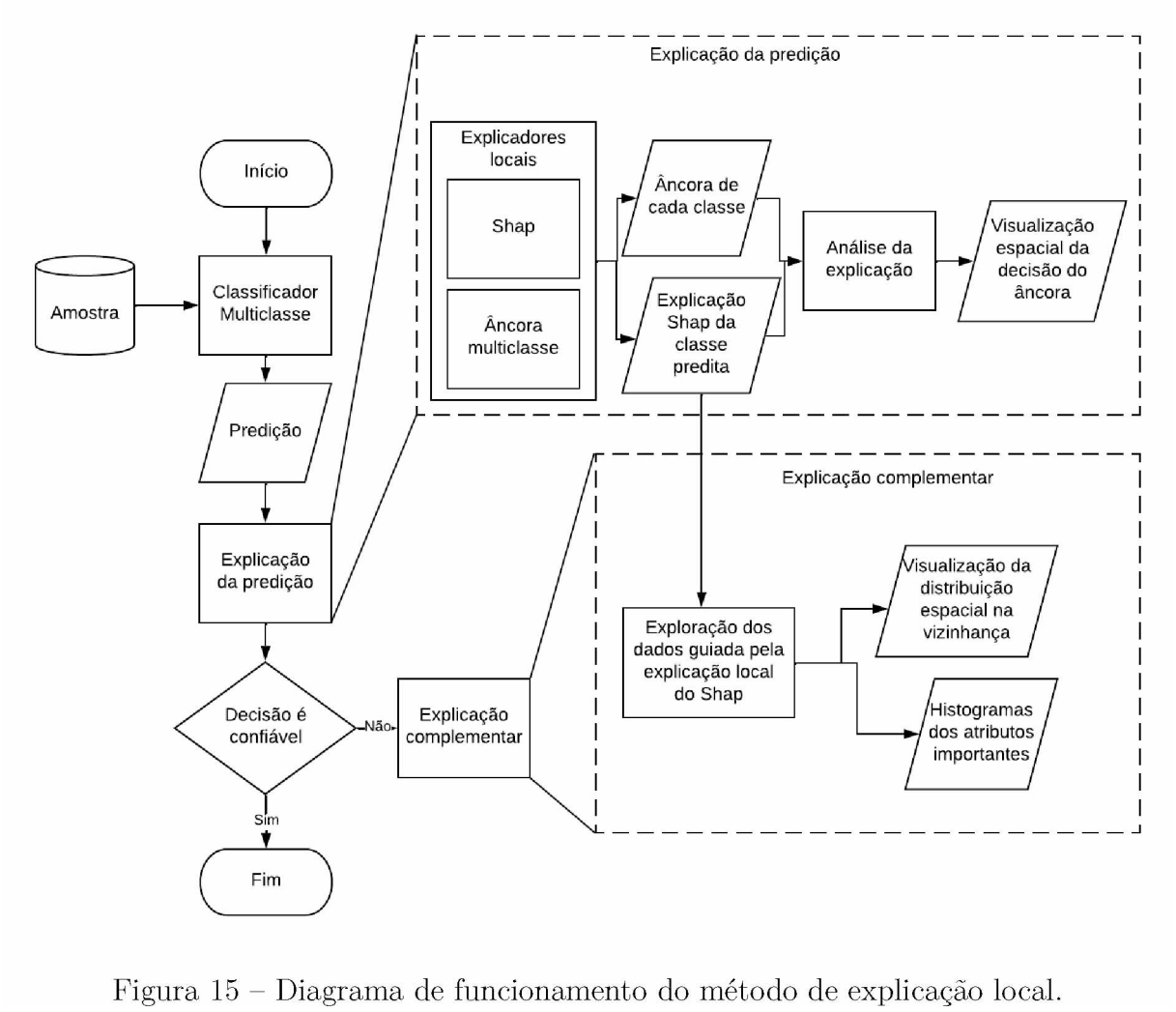


Figura 15 – Diagrama de funcionamento do método de explicação local.

por uma regra de decisão, utilizando poucos atributos.

A metodologia para explicação local deste trabalho combina o uso dos algoritmos SHAP e âncora a fim de explorar suas características particulares para oferecer mais informação ao usuário. Para complementar as explicações geradas por essas técnicas, também foram construídas outras formas de visualização dos dados baseadas em histogramas e em gráficos com projeções tridimensionais de regiões do espaço amostral. Na construção dessas visualizações, as explicações locais são utilizadas para selecionar os atributos a serem considerados, restringindo a dimensionalidade dos dados de modo a possibilitar sua exibição para o especialista humano.

A escolha de quais explicações serão apresentadas ao usuário é definida de acordo com o grau de confiança da decisão, ou seja, o sistema provê diferentes explicações com base na chance do classificador ter feito uma predição correta. O fluxo de execução desse processo é apresentado na Figura 15.

Após a classificação de uma amostra, o sistema CAD toma uma decisão baseada na estimativa do grau de confiança da decisão. O cálculo dessa estimativa depende do classificador utilizado. Alguns métodos já fornecem esse tipo de informação, como é o caso

de uma rede neural multicamadas que utiliza como função de perda a entropia cruzada (BISHOP, 2006; RUSSELL; NORVIG, 2002). Porém, há métodos que apenas indicam o rótulo predito, então foi adotada uma forma de categorizar a classificação em três níveis de confiança: confiável; incerta; e inconclusiva. Este procedimento foi detalhado na Seção 3.4.1.

Em casos em que a predição é categorizada como confiável, o sistema oferece apenas explicações locais, utilizando o SHAP, o *âncora* e uma proposta de visualização dessa explicação, denominada visão espacial da explicação do âncora, a qual apresenta a distribuição espacial das amostras de treinamento e daquela em análise, bem como a região de cobertura da regra criada pelo método âncora a partir do classificador multiclasse. Em casos em que a predição é categorizada como incerta ou inconclusiva, além dessas mesmas explicações locais, é também apresentada uma exploração dos dados guiada pela explicação local. Quando a predição é incerta, há também a indicação de quais classes foram descartadas durante a análise dos modelos binários, oferecendo assim, uma informação adicional que não é possível para predições inconclusivas.

Maiores detalhes sobre cada um dos artefatos disponibilizados nas explicações locais são apresentados a seguir.

3.5.3.1 Explicação local baseada no SHAP

A fim de mostrar como o modelo efetuou a classificação de uma amostra em particular, o método SHAP gera um gráfico de forças (Figura 16(a)), o qual mostra os atributos que mais influenciaram na decisão, bem como se eles afetaram positiva ou negativamente na escolha da classe (LUNDBERG; LEE, 2017). A largura das barras no gráfico indicam o impacto dos atributos, ou seja, atributos mais relevantes possuem barras mais largas. As barras em vermelho mostram os atributos que influenciam o modelo a decidir a favor da classe predita, enquanto as barras em azul representam os atributos que contribuem para uma decisão contrária à classe.

3.5.3.2 Explicação local baseada em âncoras

O método âncora (*anchors*) foi projetado para explicar predições de classificadores binários, pois o tipo de explicação gerado mostra as condições presentes na amostra que a qualifica como pertencente ou não a apenas uma classe. Portanto, é necessária uma estratégia para adaptá-lo para lidar com três ou mais classes. Para isso, foi desenvolvido uma abordagem que manipula qualquer modelo multiclasse treinado, tratando-o como uma caixa preta, e o converte em um classificador binário, no estilo um-contra-todos, para uma classe específica. Essa abordagem é aplicada para cada classe presente no modelo, gerando assim, um conjunto de classificadores binários. Ao explicar cada um deles, é possível entender os motivos da amostra pertencer ou não a cada uma das classes.

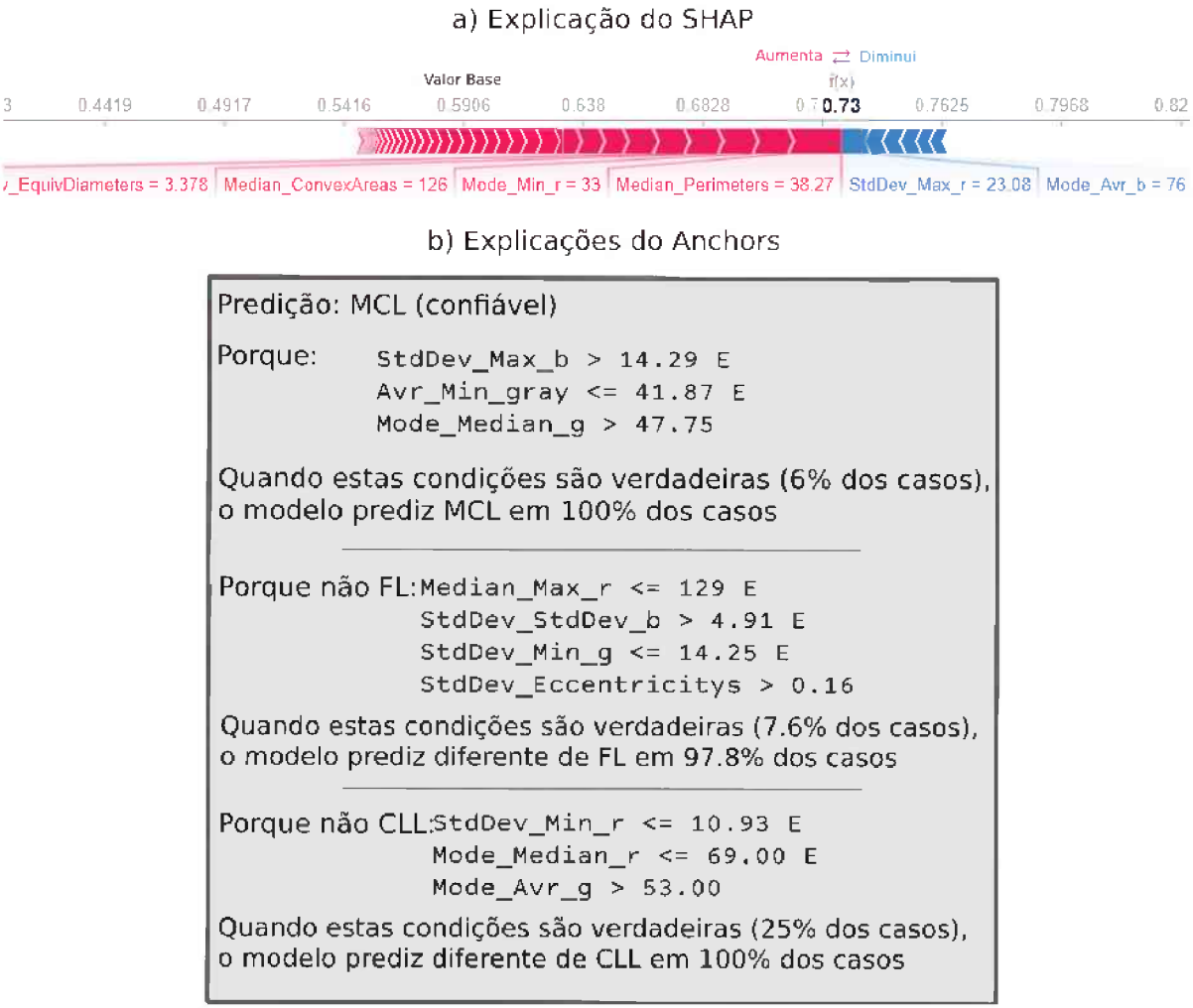


Figura 16 – Exemplo de parte da explicação local proposta na metodologia, contendo a) um gráfico de forças criado pelo algoritmo SHAP; e b) uma visualização das ancoras criadas a partir de cada classe pertencentes ao problema.

Esta técnica foi baseada na metodologia apresentada em (YOO et al., 2020) para explicar as predições de um modelo multiclasse, desenvolvido para recomendar o melhor tratamento para pacientes com problemas de visão. Para fornecer explicações binárias em cenários multiclasse, os autores adotam uma solução baseada no treinamento de classificadores binários, nas formas um-contra-todos e um-contra-um, e na explicações locais para as predições desses modelos. Ao contrário da proposta original, em nossa abordagem, não há o treinamento de outros modelos, o que aumenta a fidelidade das explicações, uma vez que elas são geradas a partir do próprio modelo utilizado na predição, e reduz consideravelmente o custo computacional, tanto de memória, quanto de processamento, já que não é necessário guardar nem treinar mais que um modelo.

O modelo proposto nesse trabalho gera uma explicação binária para cada rótulo e depois os combina para explicar a classificação multiclasse. Explicações binárias são geradas com base em vários mapeadores, como mostrado na Figura 17. Cada mapeador trata uma única classe, e mapeia as amostras entre dois rótulos binarizados: um que

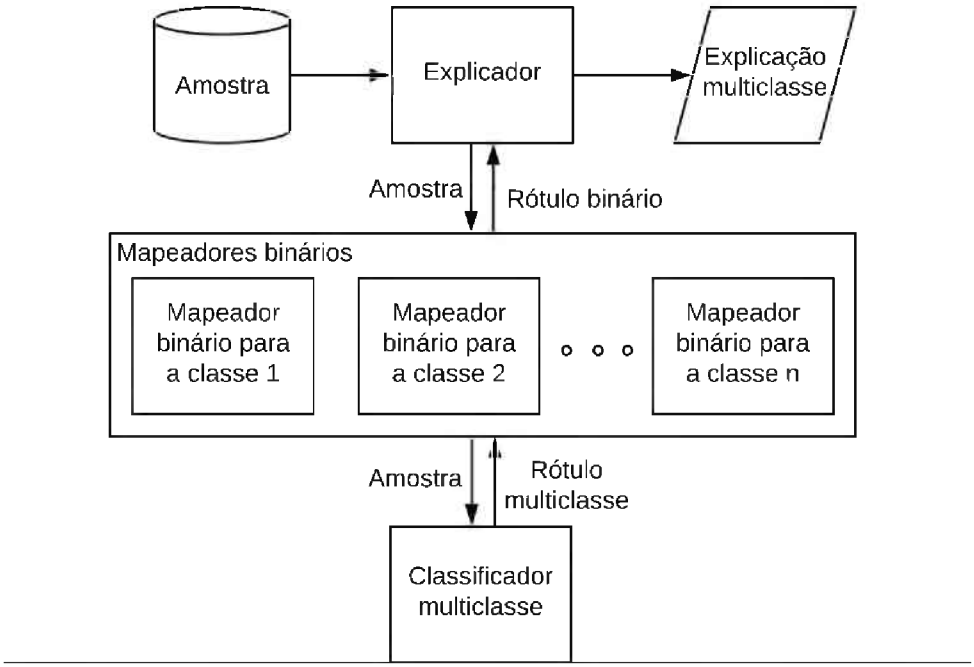


Figura 17 – Arquitetura do método de explicação das predições multiclasse com base em múltiplos mapeamentos binários.

representa a classe original desejada (rótulo 1); e outro que agrupa todas as demais (rótulo 0). Por exemplo, se uma amostra z pertence à mesma classe endereçada pelo mapeador, ela classifica z como rótulo 1. Caso contrário, z é rotulado como 0. O método usa esses mapeadores binários para explicar as amostras multiclasse, criando uma explicação para cada rótulo. Um exemplo de explicação local do âncora baseada nessa estratégia é apresentado na Figura 16(b). Como pode ser observado, o sistema apresenta uma explicação do motivo pelo qual o modelo classificou a amostra como a classe prevista e porque os demais rótulos não foram escolhidos.

Além da condição criada a partir da conjunção de predicados, a explicação também é formada pelo grau de confiança da predição (valor categórico, estimado pelo nosso método); pela cobertura da regra, que corresponde ao percentual de amostras na base de treinamento que satisfazem as condições da regra; e pela acurácia da âncora, ou seja, a probabilidade do preditor fazer a mesma predição, dado que as condições sejam verdadeiras.

3.5.3.3 Visualização espacial da explicação do âncora

A explicação local criada pelo SHAP deixa claro quais são os atributos que mais contribuem para a classificação da amostra e o método âncora cria um processo simples de decisão que imita o comportamento do modelo preditivo nas proximidades da amostra. Porém, estas explicações não se propõem a explicar o motivo destes atributos serem importantes, nem a mostrar quais dados foram levados em conta pelo modelo para tirar

estas conclusões. Neste trabalho, é proposta uma forma de visualização dos dados que busca fornecer esse tipo de informação por meio de uma projeção tridimensional de parte do espaço amostral. O gráfico 3D resultante pode ser rotacionado, de forma interativa, para uma melhor visualização pelo usuário.

Uma explicação gerada pelo método âncora contém uma regra do tipo SE-ENTÃO que separa o espaço amostral em duas partes (classificação binária). A região delimitada pelas premissas (condições) da regra formam um hipercubo em que cada dimensão representa um dos atributos utilizados. As amostras que estão dentro desta área são frequentemente classificadas pelo modelo preditivo com um certo rótulo. Na visualização proposta, o hipercubo é formado a partir dos três atributos mais relevantes da regra âncora criada para a classe predita pelo modelo multiclasse. Os vértices do hipercubo são definidos pelo valor utilizado na premissa condicional de cada atributo e pelo valor limítrofe (máximo ou mínimo, a depender do operador relacional da premissa), encontrado na base de dados de treinamento. Por exemplo, se a regra define que o atributo Y deve ser maior (ou maior e igual) que o valor X, então os vértices de Y serão formados por X e o valor máximo desse atributo na base. Caso a regra âncora possua mais de três atributos, os três primeiros são usados na visualização, uma vez que são os mais influentes na decisão, e os demais são descartados. Por outro lado, se a regra possui um número menor de atributos, a projeção tridimensional é assegurada pela seleção de atributos adicionais a partir do ranqueamento de relevância gerado pelo SHAP. Neste caso, como não existe uma premissa condicional para os atributos adicionados, adotou-se os valores de máximo e mínimo desses atributos na base de treinamento para formar os vértices do hipercubo.

A Figura 18(a) mostra um exemplo desse tipo de visualização, construída a partir de uma explicação que usava a regra de decisão "$StdDev\_Avr\_r > 13,14$ E $StdDev\_MajorAxisLengths <= 3,85$ E $StdDev\_Median\_gray > 13,88$". Como a regra utiliza três atributos, todos são considerados. No gráfico, também são representadas as posições das amostras mais próximas do objeto sendo classificado. Foram utilizados símbolos e cores diferentes para representar a classe de cada amostra apresentada: círculos laranja para as amostras da classe CLL; triângulos verdes para a classe FL; e cruzes azuis para a classe MCL. A amostra em análise (cujo o diagnóstico está sendo explicado) é representada por uma estrela vermelha de modo a diferenciá-la das demais.

Para auxiliar na análise da visão espacial, o usuário tem acesso aos dados da amostra classificada (alvo) e das três instâncias de cada classe que estão mais próximas dela, como ilustrado no exemplo da Figura 18(b). Para cada instância são exibidos os valores originais (sem normalização) dos atributos usados na explicação do âncora, bem como sua distância para a amostra alvo, calculada a partir dos valores normalizados desses atributos, e a imagem da lesão. Dessa forma, o método permite que o usuário realize uma análise comparativa, no contexto da explicação local, entre a amostra e as instâncias mais próximas de cada classe. Tal análise pode auxiliar no entendimento acerca da decisão do

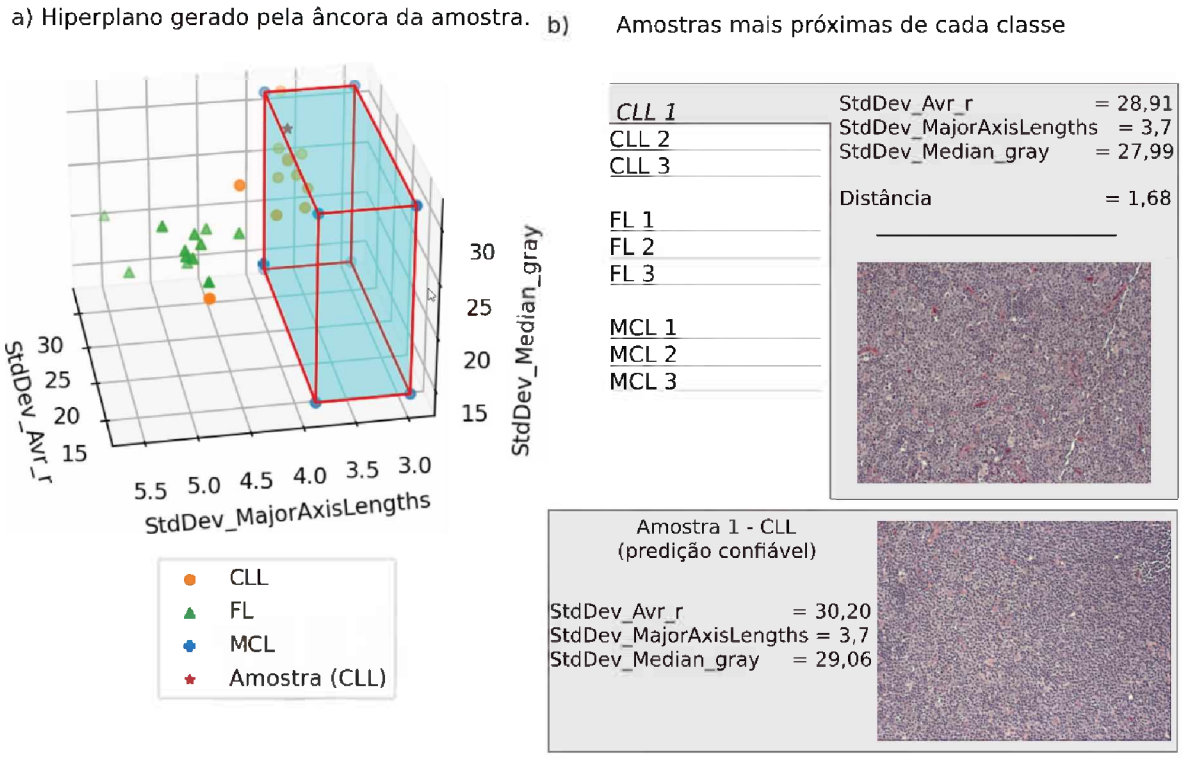


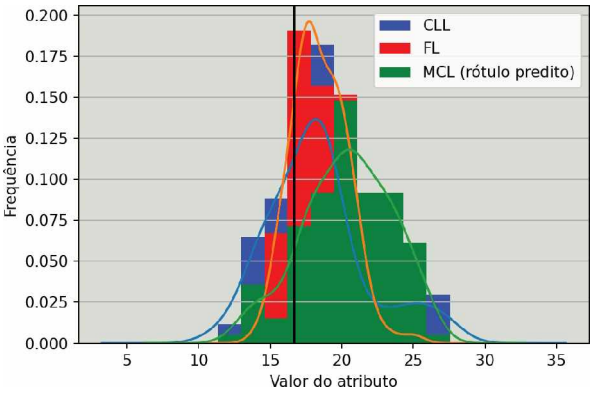
Figura 18 – Visualizações geradas a partir da explicação local do método âncora para uma amostra da base de linfoma: (a) visão espacial da explicação do âncora; e (b) dados das três instâncias mais próximas de cada classe.

modelo e, consequentemente, aumentar a confiança no diagnóstico.

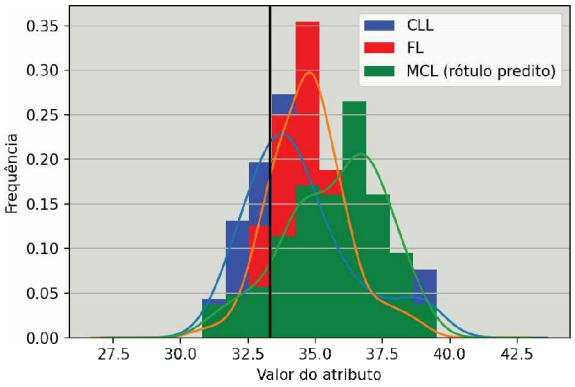
3.5.3.4 Exploração dos dados guiada pela explicação local

Caso a predição de uma amostra seja rotulada como incerta ou inconclusiva, a confiança na decisão do modelo e nas explicações locais geradas a partir dela fica comprometida. Nesse cenário, o sistema prevê o fornecimento de informações adicionais que visam dar suporte ao diagnóstico da lesão pelo especialista humano. A abordagem emprega as explicações locais dos métodos SHAP para determinar os atributos mais relevantes para a predição da amostra, utilizando-os para mostrar informações acerca da distribuição das instâncias usadas no treinamento do modelo.

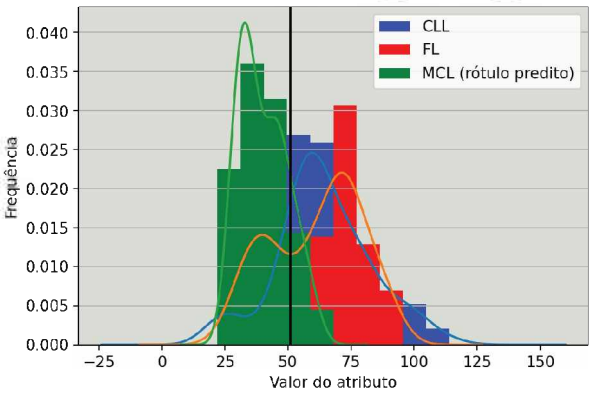
Uma das formas propostas para apresentar informações adicionais ao usuário é a criação de um gráfico com histogramas da distribuição dos valores de um atributo entre as classes do problema. Nesse processo, o SHAP é usado para selecionar os três atributos mais relevantes na predição da amostra investigada. Cada atributo é usado na construção de um gráfico, cujo os histogramas mostram as frequências em que cada valor desse atributo aparece nas diferentes classes, considerando as instâncias do conjunto de treinamento. Em outras palavras, mostra quantas instâncias de cada classe possuem aquele valor de atributo. A Figura 19 mostra os histogramas gerados a partir de uma amostra da base do linfoma. Este exemplo não foi construído a partir da mesma predição utilizada



(a) histograma - median_StdDev_r



(b) histograma - Median_Perimeters



(c) histograma - Mode_Min_r

Figura 19 – Histogramas com a distribuição da quantidade de instâncias de cada classe pelos valores dos atributos mais relevantes, de acordo com a explicação local do SHAP.

nos exemplos anteriores, então os atributos mais importantes são diferentes. Cada classe é representada por uma cor distinta: azul para o CLL; vermelho para o FL; e verde para o MCL, que também é a classe predita para a amostra analisada. O valor do atributo para a amostra classificada é destacado em preto, permitindo ao usuário comparar os valores da amostra com as distribuições de cada classe.

Outra forma de prover informação sobre os dados disponíveis é mostrar a distribuição espacial das instâncias do conjunto de treinamento que se encontram na vizinhança da amostra classificada. Isso pode ser realizado através um gráfico tridimensional cujo as dimensões são formadas pelos três atributos mais importantes para a predição da amostra, selecionados a partir da explicação do SHAP. Esses atributos também são empregados no cálculo das distâncias entre as amostras. Como uma visualização de todas as amostras da base de dados de treino seria de difícil compreensão, são apresentadas apenas as três amostras mais próximas de cada classe. A Figura 20 mostra um exemplo deste método para um caso em que a confiança da decisão foi classificada como incerta. Além da visualização espacial presente na Figura 20(a), também são apresentados, como exemplificado na Figura 20(b), os dados referentes à amostra classificada e às três mais próximas de cada

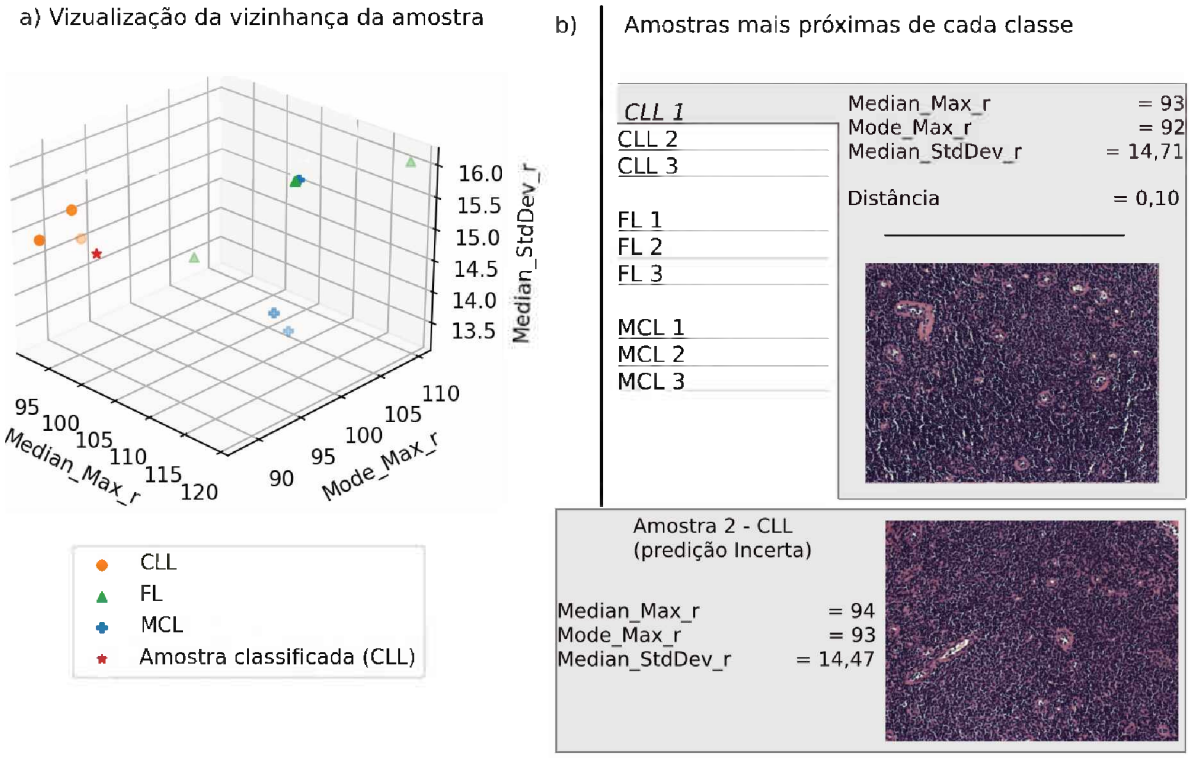


Figura 20 – Visualizações geradas a partir distribuição espacial das instâncias do conjunto de treinamento da base de linfoma: (a) visão espacial da vizinhança da amostra classificada; e (b) dados das três instâncias mais próximas de cada classe.

classe. Os valores dos atributos mais impactantes são exibidos, bem como as distâncias das amostras vizinhas à amostra alvo e as imagens histológicas de cada caso.

Capítulo 4

Experimentos e Análise dos Resultados

Neste capítulo, são apresentados os experimentos realizados para verificar a eficiência da metodologia proposta para a classificação multiclasse de imagens histológicas, e também uma análise empírica da metodologia de explicação proposta. Primeiramente, o método preditivo foi avaliado utilizando a base de dados de imagens de linfomas. Então, utilizando estes mesmos dados de treinamento e teste, é apresentada uma demonstração de como cada parte da metodologia de explicação funciona, e uma discussão de como elas podem ser utilizadas. Por último, a metodologia é novamente avaliada sobre a segunda base de imagens, imagens histológicas de displasia da cavidade oral, com diferentes níveis de severidade.

4.1 Método para a Avaliação

Para validar a metodologia de classificação multiclasse proposta, inicialmente verificou-se a efetividade da etapa de pré-processamento em eliminar atributos que não contribuem para a classificação. Em seguida, avaliou-se o desempenho do método em classificar novas amostras, ou seja, que não foram utilizadas na construção do modelo. Esta segunda avaliação foi construída através de duas métricas comumente utilizadas para avaliar modelos de classificação: a acurácia e o *F-score* (AGGARWAL, 2015; BISHOP, 2006). A acurácia, que também é chamada de taxa de acerto, é calculada segundo a Equação 19:

$$Acurácia = \frac{NCorretas}{NPreditas}, \tag{19}$$

sendo "N Corretas" o número de amostras classificadas corretamente pelo modelo, e "N Preditas" o número total de amostras preditas. O *F-score* é avaliado separadamente para cada classe do problema, e então, a média aritmética desses valores é calculada. Esta medida pode ser calculada através da Equação 20:

$$Fscore(Y_i) = 2 * \frac{precisão * revocação}{precisão + revocação}, \tag{20}$$

sendo Y_i uma das classes do problema; precisão a fração das amostras classificadas como Y_i que realmente pertencem a essa classe (Equação 21); e revocação a fração de amostras da classe Y_i que foram corretamente classificadas (Equação 22). Essas duas métricas (precisão e revocação) são calculadas considerando uma das classes como positiva, e o restante como negativas. Elas são calculadas com base na quantidade de exemplos positivos classificadas corretamente (Verdadeiras positivas - VP); na quantidade de exemplos negativos classificados como positivos (falso positivo - FP); e na quantidade de exemplos positivos classificados erroneamente como negativos (falso negativo - FN).

$$precisão = \frac{VP}{VP + FP} \tag{21}$$

$$revocação = \frac{VP}{VP + FN} \tag{22}$$

As métricas avaliadas nos experimentos foram calculadas considerando 50 execuções do algoritmo, utilizando a técnica de validação cruzada estratificada repetida, com cinco partições dos dados e dez repetições. Este método é indicado para pequenos conjuntos de dados e apresenta menor variação nos resultados (WONG; YEH, 2019; RODRIGUEZ; PEREZ; LOZANO, 2009).

A validação cruzada consiste em separar os dados em k grupos, chamados de partições, criados através de uma seleção aleatória. Entretanto, na abordagem estratificada, as partições devem conter a mesma proporção de amostras de cada classe observada no conjunto completo. Então, o método proposto é avaliado k vezes, utilizando uma das partições para teste e o restante para o treinamento do modelo (AGGARWAL, 2015). Foram realizadas 10 repetições da validação cruzada para cada método de classificação a fim de aumentar o número de avaliações, diminuindo o efeito da variação na média dos resultados (WONG; YEH, 2019).

A fim de se ter uma comparação mais precisa entre as diferentes abordagens avaliadas, foi empregado teste de hipótese na análise dos resultados, mais especificamente o teste-t de *Student* de rastro bilateral (RICE, 2006). Esse teste é usado para verificar a hipótese nula (H_0) de que duas populações (neste caso, avaliações realizadas utilizando validação cruzada) tenham a mesma média, assumindo que os dados seguem uma distribuição normal e considerando um intervalo de confiança de 95% (valor-p = 0,05). Assim, se o teste de hipótese resultar em um valor-p menor que 0,05, a hipótese nula é rejeitada, concluindo que existe diferença estatisticamente significativa entre as abordagens. Caso contrário, tem-se que a diferença na avaliação pode ter sido resultante da variação aleatória inerente aos métodos estocásticos.

4.2 Avaliação da Metodologia na Classificação de Linfomas

Nesta primeira parte do capítulo, são apresentados experimentos, resultados e análises que foram construídos utilizando a base de dados de LNH como caso de uso. Estes experimentos iniciais foram utilizados para definir parâmetros e guiar a escolha das técnicas utilizadas na metodologia. O problema de classificação de LNH foi utilizado para este fim pois, em relação a base de dados de displasias, possui mais trabalhos correlatos e resultados publicados, que podem ser utilizados para comparação. Além disso, ele pode ser considerado um problema mais simples, já que possui apenas três classes, o que o faz mais apropriado a ser utilizado como um primeiro objetivo.

4.2.1 Análise Comparativa entre os Classificadores Multiclasse

Este experimento compara o desempenho de quatro algoritmos de aprendizado de máquina supervisionado (GBDT, SVM, regressão linear e MLP) na tarefa de classificação multiclasse de LNH. Para a preparação dos dados, foi empregada a base de treino para realizar a filtragem de atributos sem variância, como descrita na seção de metodologia proposta. Neste caso de uso, não houve nenhuma ocorrência de atributo com variância igual a zero, então os descritores foram mantidos com o total de 114 após a filtragem. O processo de padronização dos dados foi realizado, conforme descrito na seção de metodologia proposta (Seção 3.2). Isso é necessário pois alguns dos métodos de classificação investigados, como a regressão linear e o MLP, podem ter resultados comprometidos se utilizados com atributos de ordens de grandeza distintas (AGGARWAL, 2015).

A Tabela 3 mostra os resultados deste experimento. A coluna Acurácia mostra a acurácia média e seu desvio padrão, e a coluna *F-score* mostra o valor médio alcançado em cada modelo para essa métrica. Estes valores foram calculados a partir das 50 execuções do algoritmo, que foram realizados conforme a validação cruzada apresentada na Seção 4.1. A coluna de nome *p-value* indica o resultado do teste de hipótese considerando a acurácia média entre a regressão linear (maior resultado) e cada um dos outros algoritmos. O classificador baseado em regressão linear obteve os melhores resultados (acurácia média de 95,6% e *F-score* de 0,955), com significância estatística em relação aos outros métodos, considerando o valor de corte de 0,05 adotado nos testes deste trabalho, com 2.6% a mais em sua taxa de acerto média do que o segundo melhor avaliado, MLP.

4.2.2 Impacto da Engenharia de Atributos na Classificação

Utilizando a correlação de Pearson, foi possível avaliar quais atributos são mais relacionados entre si. Assumindo que dois atributos altamente correlacionados contém pouca informação a mais do que um deles separado, foi considerado a a hipótese que atributos

Tabela 3 – Desempenho de diferentes métodos de classificação e resultado do teste de hipótese da diferença da acurácia média com a regressão linear (maior resultado) para cada um dos outros algoritmos.

Modelos	Acurácia	*F-score*	*p-value*
LGBM	$92,3\% \pm 0,2$	0,920	7.23E-11
Regressão Linear	**95,6% ± 0,2**	**0,955**	**-**
SVM	$91,0\% \pm 0,3$	0,909	4.57E-15
MLP	$93,4\% \pm 0,2$	0,933	1.25E-06

altamente correlacionados poderiam ser retirados da base de dados sem que haja grande impacto na qualidade do classificador.

Para testar essa hipótese, foi avaliado o desempenho do sistema em relação ao uso ou não da redução na quantidade de atributos e o desempenho dos classificadores. Com isso, foi avaliada a capacidade de separação das classes nos dados dos casos com a redução da dimensão dos dados. Neste experimento, foi adotado como altamente correlacionados atributos com uma correlação de Pearson maior ou igual a 0,99. Este parâmetro foi determinado após experimentos preliminares, que mostraram que com este valor, o método filtra uma quantidade significativa de atributos sem grande impacto no desempenho dos classificadores.

A Tabela 4 mostra os resultados obtidos pelo modelo de classificação, com e sem o processo de filtragem dos atributos. As colunas "Nº atr." mostram a quantidade de atributos utilizados para treinar o modelo preditivo. O método de filtragem retirou, em média, 33,5 atributos, que corresponde a uma redução de quase 30%. A coluna Acurácia indica a média e desvio padrão dessa métrica, e *F-score* indica a média da métrica em cada modelo avaliado. O único preditor significantemente impactado pelo uso do filtro de atributos baseado na correlação de Pearson foi o algoritmo baseado em regressão linear. Considerando a aplicação do filtro, três algoritmos tiveram resultados sem diferença estatística, apresentando desempenho próximo com os melhores resultados em relação a classificação: os métodos baseados em LGBM e MLP, que obtiveram 92,8%, e o regressor linear, que obteve 92,7% de acurácia média. Com base nesses resultados, os próximos experimentos foram analisados apenas com um dos algoritmos que proporcionaram os resultados mais promissores. Nesse caso foi selecionado o algoritmo MLP, com a utilização do filtro de atributos. Apesar de ter sido possível construir um modelo mais acurado sem utilizar o filtro de atributos (acurácia média de 95,6%), foi priorizado o ganho em interpretabilidade e simplicidade do modelo proveniente da redução do número de atributos. Dentre os três métodos com desempenhos próximos, foi escolhido o algoritmo MLP, pois, por ter a possibilidade de calcular a probabilidade da classificação, permite o cálculo de métricas de avaliação que necessitam desse valor, como a perda de entropia cruzada, e a área sobre a curva ROC, que foi utilizada para comparar a metodologia proposta com resultados de trabalhos relacionados.

Tabela 4 – Impacto da filtragem de atributos baseada na correlação de Pearson no desempenho de diferentes métodos de classificação.

	Sem Filtro			Correlação de Pearson		
Modelos	Nº atr.	Acurácia	F-score	Nº atr.	Acurácia	F-score
LGBM	114	92,3% ± 0,02	0,920	80,6 (~70%)	**92,8% ± 0,03**	**0,926**
Regr. linear	114	95,6% ± 0,02	0,955	80,5 (~70%)	**92,7% ± 0,03**	**0,926**
SVM	114	91,0% ± 0,03	0,909	80,5 (~70%)	90,5% ± 0,04	0,904
MLP	114	93,4% ± 0,02	0,933	80,3 (~70%)	**92,8% ± 0,03**	**0,927**

4.2.3 Comparação com Trabalhos do Estado da Arte

A Tabela 5 mostra o desempenho em relação as métricas acurácia (ACC) e área sob a curva ROC (AUC ROC) de trabalhos que investigaram a classificação multiclasses das lesões NHL. Para comparação, são apresentados os dados do nosso classificador baseado em MLP, e que utiliza uma filtragem de 70% dos atributos, pois foi o modelo empregado em todos os experimentos de explicação. A coluna "Ref" mostra as informações da citação do artigo utilizado. O número de atributos e métodos de extração estão nas colunas "Qtd." e "Extração de atributos", respectivamente.

O modelo de classificação proposto utiliza menos atributos que os métodos de Nascimento et al. (2015), Codella et al. (2016), Song et al. (2016), Bai et al. (2019) e Roberto et al. (2021). A metodologia proposta utiliza descritores formados por informações mais interpretáveis que os outros trabalhos, e alcançou métricas melhores que as abordagens de Roberto et al. (2017) e Martins et al. (2020).

Tabela 5 – Comparação do método de classificação multiclasse de LNH proposto com trabalhos relacionados.

Ref.	Extração de atributos	Qtd.	Classificador	ACC	AUC ROC
(MENG et al., 2010)	Cor, histograma, textura, wavelet e informação de padrões binários	50	C-RSPM	92,7%	
(NASCIMENTO et al., 2015)	Transformada de ondas estacionárias	34.236	SVM	100%	
(CODELLA et al., 2016)	Histograma; LBP; gist; curvelet; correlograma de cores e momentos; wavelet	200	SVM ensemble	95,5%	
(SONG et al., 2016)	IFV; LBP; HOG;GIST e CENTRIST	180	SVM	96,8%	
(ROBERTO et al., 2017)	Atributos baseados em teoria de percolação	15	DECORATE	92,0%	0,943
(BAI et al., 2019)	Atributos de textura estática; Atributos de cor; modelo pré-treinado GoogLeNet	550	RF + CNN	99,1%	0,998
(MARTINS et al., 2020)	Características de geometria fractal	18	HPG4	91,4%	
(ROBERTO et al., 2021)	Características de geometria fractal	100	CNN	95,55%	
Método Proposto	Descritores morfológicos e não-morfológicos	80	MLP	92,8%	0,988

4.2.4 Avaliação do Método de Estimação da Confiança

Neste trabalho, foi proposta uma metodologia para estimar a confiabilidade das predições de um classificador multiclasse. Para fazer isso, modelos auxiliares são treinados, utilizando o mesmo algoritmo do classificador, porém com problemas mais simples, criados com apenas duas das classes existentes nos dados. Utilizando as predições dos modelos binários, e comparando-as com a predição do modelo multiclasse, o método categoriza a predição em um de três grupos: confiáveis; incertas; e inconclusivas. Nesse contexto, a hipótese é de que predições categorizadas como confiáveis tem maior chance de estarem corretas em relação aos grupos categorizados como incertos. Por sua vez, predições categorizadas como incertas, por haver divergência apenas em alguns dos classificadores, tem maiores chances de estarem corretas do que as predições consideradas inconclusivas, onde todos os classificadores binários divergem da classificação multiclasse.

Para avaliar essa hipótese, foi realizado um experimento utilizando o classificador baseado em MLP, sem a utilização do filtro de correlação de Pearson e com os atributos normalizados. Como pode ser observado na Tabela 6, o classificador multiclasse obteve em média uma acurácia de 93,7%. Os modelos auxiliares binários, formados pelos grupos de classes CLL e FL; CLL e MCL; e FL e MCL, tiveram taxas de acerto média de, respectivamente, 95,7%, 92,9% e 97,0%. Nesses experimentos, cada amostra dos conjuntos de treino foi categorizada pelo método de estimação de confiança, e então, para as amostras de cada categoria, foi calculada a acurácia média do modelo multiclasse e a probabilidade de classificação média, obtida pelo próprio modelo de classificação MLP. Como pode ser visto na coluna "Acurácia" da Tabela 6, a taxa de acerto das amostras do grupo confiável (95,5%) foi maior que a média geral do modelo (93,7%). Considerando apenas o grupo "incerta", o modelo teve desempenho de 63,6% e para o grupo "inconclusiva" esse valor foi ainda pior, com uma taxa de acurácia de 43,7%. Isso mostra que a metodologia utilizada contribui na etapa de predição entre as amostras, estimando sua confiabilidade de forma acurada. A coluna "% de Amostras" mostra a fração das amostras de teste que foram categorizadas em cada grupo, em que é possível observar que a maioria das predições (93% de todas avaliadas) são confiáveis e têm alta taxa de acerto (95%), e predições inconclusivas são raras (0,4% das avaliadas). A coluna "Probabilidade MLP" mostra a probabilidade de predição média de cada grupo, de acordo com o classificador treinado. É possível perceber que as amostras das categorias menos confiáveis também tiveram uma probabilidade de classificação inferior numa comparação com aos dados confiáveis.

O grupo das predições incertas é formado por amostras que tiveram predições inconsistentes em pelo menos um, mas não em todos os modelos binários analisados. Para esse grupo, a hipótese adotada foi que por meio dos modelos discordantes, é possível definir um subgrupo de rótulos que tenham maior probabilidade de pertencer a classe da amostra. Esse grupo é composto pelos rótulos que aparecem pelo menos uma vez no conjunto de predições binárias e multiclasse. Para avaliar esta hipótese, pode-se contar a taxa de

Tabela 6 – Avaliação do método de estimação de confiabilidade de predições multiclasse.

Categoria	% de Amostras	Acurácia	Probabilidade MLP
Confiável	93,7%	95,5%	0,964
Incerta	5,8%	63,6%	0,792
Inconclusiva	0,4%	43,7%	0,665

amostras de teste da categoria incerta que pertence a uma das classes contidas nos seus respectivos subgrupos. Nas investigações essa taxa foi de 93,8%, o que indica que em apenas 6,2% dos casos, a classe correta estava fora do subgrupo.

4.2.5 Exploração dos Dados

Para realizar uma avaliação empírica dos métodos de explicação propostos, a base foi dividida em conjuntos de treinamento e de teste, sem a utilização da técnica de validação cruzada. Essa escolha foi feita para uma análise qualitativa das explicações geradas, com discussões a cerca dos resultados observados com o caso de uso, para cada tipo de explicação. Para isso, os experimentos seguintes que propõem este tipo de análise sobre as explicações de dados, global e local foram todos realizados utilizando os mesmos dados de treinamento e teste, e o mesmo modelo de classificação. O conjunto de treino foi formado por 80% dos dados, escolhidos aleatoriamente de forma que a proporção entre classes fosse mantida, e 20% dos dados foi destinado para os testes.

Na análise que visa a explicação dos dados, foi adotado um limite de correlação de 0,97 para delimitar grupos de atributos. Essa estratégia considera que quando dois atributos tem correlação maior ou igual a 0,97, eles formam um grupo. Este parâmetro pode ser alterado para ajustar a quantidade de grupos gerados e seus tamanhos. Esse parâmetro foi adotado por um estudo empírico em que considerou a investigação de grupos com até cinco atributos, pois são mais simples de serem interpretados. É importante que este parâmetro seja menor que o valor utilizado para filtragem de atributos, pois, caso contrário, nenhum grupo poderia ser constituído, uma vez que atributos com alta correção foram removidos inicialmente nos experimentos.

A Tabela 7 apresenta os descritores agrupados pelo método. Foram formados 16 grupos, que juntos contabilizam 37 atributos, dos 80 presentes no experimento. Esta abordagem pode ser utilizada como ferramenta para descobrir possíveis relações entre as variáveis do problema. Há, por exemplo, um grupo formado pelo desvio padrão dos valores de área e o desvio padrão dos valores de diâmetro equivalente. Ao olhar para a natureza desses dados, pode-se observar que existe também uma relação causal entre esses descritores, já que a variação no diâmetro de um círculo altera diretamente sua área. Esta observação pode nos ajudar a entender como um especialista pode utilizar estes dados para descobrir relações novas existentes nos dados. Pode-se também, por exemplo, observar

Tabela 7 – Grupos de descritores com alta correlação de Pearson entre si (> 0,97).

Avr_Median_r Median_Median_r	Avr_Max_g Median_Max_gray	Avr_Min_g Avr_Median_g Avr_Min_gray Avr_Median_gray Median_Min_g
Avr_Min_b Avr_Median_b Median_Min_b	Median_Perimeters Median_ConvexAreas Median_MajorAxisLengths	Median_Median_b Mode_Avr_b
Median_Min_r Mode_Min_r	Median_Max_r Mode_Max_r	StdDev_Perimeters StdDev_MajorAxisLengths
Median_Min_gray Mode_Min_g Mode_Min_gray Mode_Median_gray	StdDev_Areas StdDev_EquivDiameters	StdDev_Min_g StdDev_Min_gray
StdDev_Median_g StdDev_Median_gray	StdDev_StdDev_g StdDev_StdDev_gray	StdDev_Max_g StdDev_Max_gray
Mode_Avr_g Mode_Median_g		

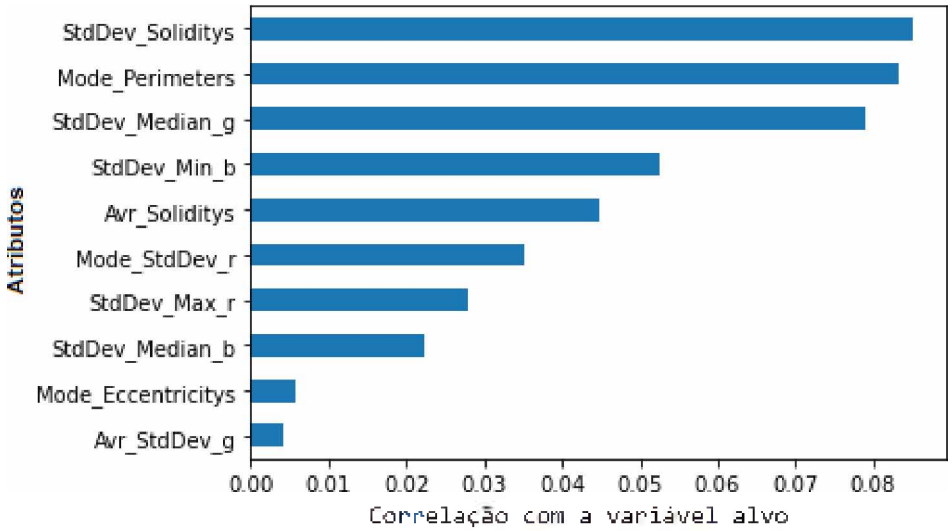


Figura 21 – Atributos com maior correlação com a classe do problema na base de linfomas.

diversos atributos do canal de cor verde e cinza correlacionados, como na segunda posição da primeira linha da tabela, que indica uma correlação entre a média dos valores máximos de brilho no canal de cor verde (Avr_Max_g) e a mediana dos valores máximos de brilho no canal de cor cinza ($Median\_Max\_gray$). Para esta correlação, não há uma explicação tão clara quanto a anterior para justificar a relação entre essas cores. Porém, ao apontar a existência dessas correlações nos dados, o método possibilita uma futura investigação de suas possíveis causas.

Outra informação importante que pode ser obtida com a correlação de Pearson é a

correlação entre cada atributo e a variável alvo do problema. Na Figura 21 são apresentadas as correlações para as classes de linfoma. O atributo com maior correlação foi o desvio padrão dos valores de solidez, com valor próximo a 0,085. Este resultado mostra que, quando separados, estes atributos têm pouca informação sobre a classe da amostra e apenas essas informações podem não contribuir para conseguir relevantes resultados no classificador. Observa-se também que nenhum atributo desta lista foi extraído do canal de cor cinza e quatro dos dez atributos mais correlacionados são morfológicos: o desvio padrão dos valores de solidez (*StdDev_Solidity*); a moda dos perímetros (*Mode_Perimeters*); a média dos valores de solidez (*Avr_Solidity*); e a moda das Excentricidades (*Mode_Eccentricitys*).

4.2.6 Análise do Modelo Baseada em Explicações Globais

A explicação do SHAP fornece uma maneira de definir o impacto de cada atributo nas decisões do modelo, permitindo que os usuários analisem quais informações são mais consideradas pelo classificador e comparem com seus conhecimentos anteriores sobre o problema. Essas informações podem ser usadas, por exemplo, para avaliar a confiabilidade do método. Neste experimento, pretende-se demostrar como essa técnica pode ser usada para interpretar um modelo de classificação de lesão de linfoma, identificando as características em cada conjunto de descritores que mais influenciam nas decisões do modelo.

Para o modelo avaliado no caso de uso de classificação de imagens histológicas de linfomas, o atributo mais importante, segundo a explicação do SHAP, foi obtido pelo desvio padrão dos valores médios de intensidade de brilho nos núcleos celulares no canal de cor vermelha (*StdDev_Avr_r*). Além disso, metade dos descritores não morfológicos dos 20 atributos mais importantes para o modelo são construídos a partir do canal de cor vermelha, o que pode indicar que essa cor é relevante para a decisão do classificador. Juntamente com o conhecimento *a priori* sobre a doença, o especialista pode utilizar essas informações para auxiliar em seu julgamento sobre a confiabilidade do método de classificação do CAD, analisando se esse tipo de informação é de fato uma boa medida para separar as classes, ou se o sistema pode ter sido influenciado por um padrão enganoso no conjunto de treino.

Como o SHAP ordena os atributos em ordem de importância para o modelo, foi levantada a hipótese de que seria possível treinar um segundo modelo preditivo utilizando apenas um subgrupo dos atributos disponíveis, selecionando aqueles mais impactantes segundo o SHAP. E que este segundo modelo teria seu desempenho pouco afetado pela diminuição na quantidade de atributos. Esta hipótese foi construída comparando o método baseado no SHAP com algoritmos de seleção de atributos, mais especificamente, aqueles que provêm uma nota para cada atributo, ou alguma forma de ordená-los, de modo que possa ser selecionada exatamente a quantidade de atributos desejada (AGGARWAL,

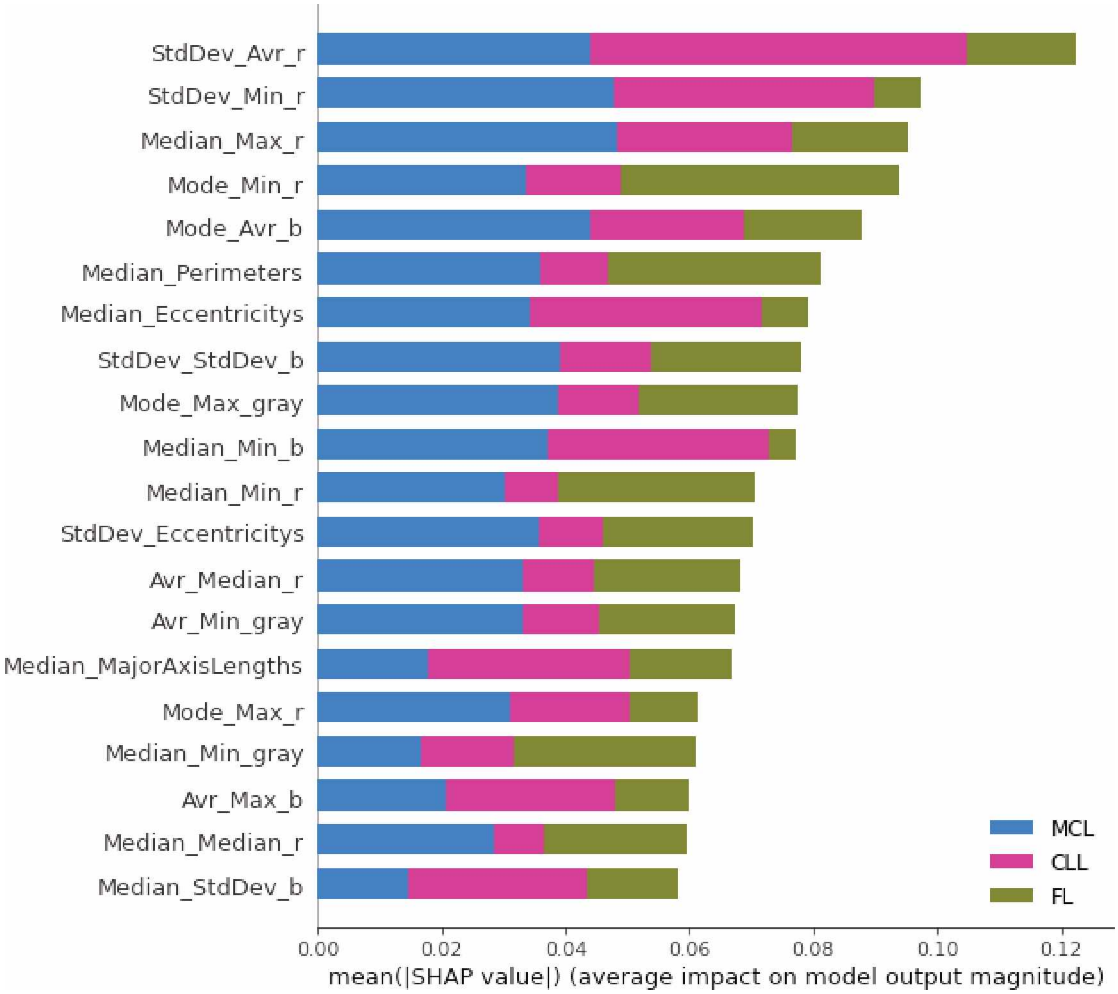


Figura 22 – Impacto médio dos 20 atributos mais impactantes no modelo de predição, especificado por cada classe predita, no caso de uso de classificação de imagens histológicas de linfomas.

2015). Nesse experimento, foram adotados outros dois métodos de seleção que fazem esta forma de ordenação dos atributos: um algoritmo baseado em análise de componentes principais (do inglês: *Principal Component Analysis* - PCA), e outro baseado em análise de variância (do inglês: *analysis of variance* - ANOVA).

O PCA cria novos atributos a partir da projeção dos dados em novas variáveis ortogonais, chamadas de componentes principais, de forma a maximizar a variância (ABDI; WILLIAMS, 2010). Uma alternativa de uso do PCA para a redução de dimensionalidade, e que foi utilizada neste experimento, é selecionar as características mais relevantes da base de dados, escolhendo aquelas que tenham mais variância residual nos componentes principais criados pelo método (SONG; GUO; MEI, 2010). O ANOVA compara a variância de um atributo com a variância da variável alvo do problema, calculando um valor F que representa sua proximidade (MILLER JUNIOR, 1997). A técnica de seleção de atributos baseada no ANOVA utilizada neste experimento seleciona os atributos com maior valor F (DING et al., 2014).

Tabela 8 – Acurácia média de modelos baseado em MLP no caso de uso de linfomas utilizando diferentes seletores de atributos.

Atributos	PCA	ANOVA	SHAP
80% (91)	0.942 ±0.03	0.936 ±0.03	0.941 ±0.03
60% (68)	0.925 ±0.03	0.924 ±0.02	0.924 ±0.03
40% (45)	0.902 ±0.03	0.942 ±0.02	0.935 ±0.02

Para validar essa hipótese, foram avaliados modelos baseados em MLP utilizando os três métodos de seleção e selecionando, com cada um deles, três diferentes quantidades de atributos (80%, 60% e 40%). A Tabela 8 mostra a acurácia e o desvio padrão médios de cada uma dessas combinações. A coluna Atributos indica a quantidade de atributos selecionados pelo método, em porcentagem do número total, e número absoluto. É possível perceber que selecionar atributos por meio do valor de importância calculado pelo SHAP se mostrou uma alternativa viável para o caso de uso proposto, com resultados similares aos outros dois métodos avaliados. Por ter um tempo de execução elevado e resultados equiparados aos outros métodos, não foi encontrada qualquer evidência que justifique a adoção do SHAP como método de seleção de atributos. Entretanto, os resultados observados mostram a eficiência do método explicativo em apontar atributos importantes.

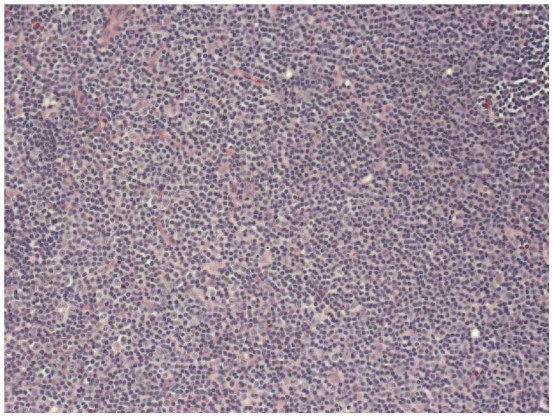
4.2.7 Análise de Predições Baseada na Explicação Local

Este experimento visa mostrar como os métodos de interpretação *Anchors* e SHAP podem ser utilizados para explicar a decisão do modelo para amostras específicas (explicação local).

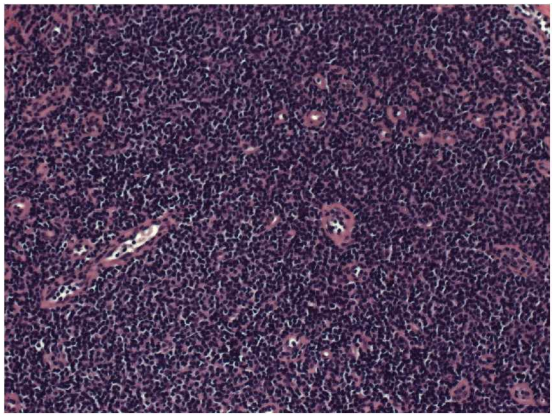
O diagnóstico médico costuma ser um problema muito difícil de generalizar, o que significa que cada caso pode ser tratado como um problema único. Essa característica torna a interpretabilidade ainda mais necessária para um sistema CAD, fornecendo informações para que o especialista analise o caso e tome a decisão final por si mesmo. As âncoras podem ser usadas para explicar classificação de uma amostra por meio de uma regra de decisão simples, fiel às decisões do modelo, enquanto a explicação do SHAP indica o impacto dos atributos mais importantes para essa decisão.

Com o objetivo de prover uma análise empírica da metodologia de explicação local, foram usadas duas amostras da base de linfomas. Uma delas, chamada aqui de amostra 1, teve sua predição categorizada como confiável, e a outra, chamada de amostra 2, teve sua predição categorizada como incerta. A Figura 23 mostra a imagem histológica de ambas as amostras. A Figura 23(a) é a amostra 1, escolhida com a predição categorizada como confiável, enquanto a Figura 23(b) é a amostra 2, escolhida com a predição categorizada como incerta.

A amostra 2 foi predita como CLL, porém de forma incerta. Isso quer dizer que a predição de um dos modelos binários foi contraditória a classificação multiclasse. Neste



(a) Linfoma do tipo CLL, classificado corretamente pelo modelo e categorizado como confiável.



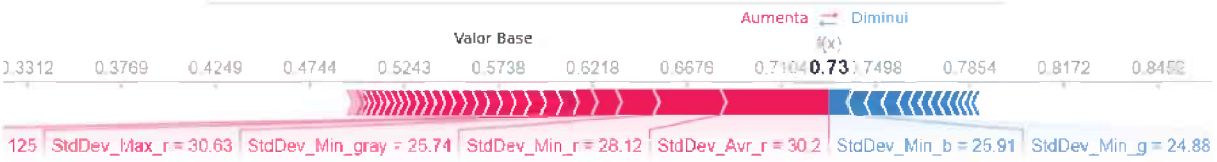
(b) Linfoma do tipo CLL, classificado corretamente pelo modelo, mas categorizado como incerto.

Figura 23 – Imagens histológicas de tecidos afetados por linfoma.

caso, o modelo binário que trata as classes CLL e FL classificou a amostra como FL. Desta forma, a metodologia proposta indica que a predição é CLL, mas que é uma predição incerta, e que há possibilidade da amostra pertencer a classe FL. Para a amostra 1, como ela teve sua predição categorizada como confiável, o método proposto somente indica que a classe predita é a CLL. Em ambos os casos a metodologia apresenta as explicações locais pelo método SHAP e *Anchors*. Na Figura 24 são mostradas essas duas formas de explicação para a amostra 1, enquanto na Figura 25 são apresentadas essas explicações locais para a predição da amostra 2. As âncoras foram criadas com um limite de *0,95*, que produz regras com pelo menos esse valor de precisão.

A explicação local do método SHAP consiste em um gráfico de forças, no qual os atributos da amostra que contribuíram a favor da classe predita são representados em vermelho, enquanto os atributos que contribuíram contra a classe predita são representados em azul. O gráfico também mostra o valor base do classificador, o qual representa um ponto de divisão entre duas classes. Se uma amostra tem o valor de $f(x)$ (soma dos impactos de cada atributo) maior que o valor base, então ela é classificada como a classe positiva da explicação. Nas Figuras 24(a) e 25(a), a classe positiva é o CLL, enquanto a negativa representa as demais classes (classificação um-contra-todos). O valor base destas explicações foi 0,5738, e o valor das amostras 1 e 2 foram, respectivamente, 0,73 e 0,72. Na Figura 24(a) observa-se que os atributos que mais contribuiram para a classificação da amostra 1 ao grupo CLL foram os atributos desvio padrão do valor médio do canal vermelho ($StdDev\_Avr_r$) e o desvio padrão do valor mínimo de intensidade de brilho no canal de cor vermelha ($StdDev\_Min\_r$). Já as características da imagem que mais contribuíram contra esta classificação foram o desvio padrão do valor médio de brilho no canal de cor azul (StdDev_Min_b) e o desvio padrão do valor de brilho mínimo no canal de cor verde (StdDev_Min_g). Na Figura 25(a), percebe-se que os atributos mais relevantes para a classe CLL são a mediana do valor máximo de brilho no canal de cor

a) Explicação do SHAP



b) Explicações do *Anchors*

Predição: CLL (confiável)

Porque: `StdDev_Avr_r > 13,14 E`
`StdDev_MajorAxisLengths <= 3,85 E`
`StdDev_Median_gray > 13,88`

Quando estas condições são verdadeiras (4,5% dos casos), o modelo prediz CLL em 100% dos casos

---

Porque não FL: `Mode_Min_g > 59,00 E`
`Median_Eccentricitys <= 0,62 E`
`StdDev_Max_g > 27,02`

Quando estas condições são verdadeiras (1,8% dos casos), o modelo prediz diferente de FL em 100% dos casos

---

Porque não MCL: `StdDev_Avr_r > 10,28 E`
`Mode_Min_r > 69,00`

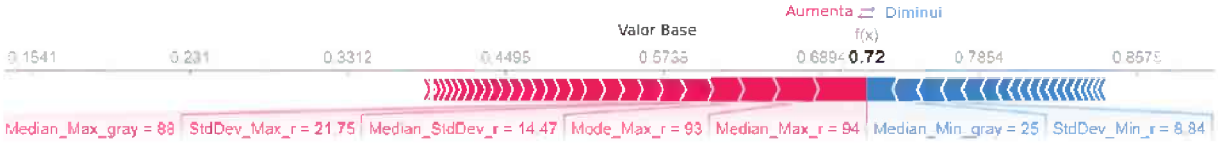
Quando estas condições são verdadeiras (15,9% dos casos), o modelo prediz diferente de MCL em 100% dos casos

Figura 24 – Explicações locais baseadas nos métodos SHAP e Anchors para a amostra 1 da base de linfomas.

vermelha (*Median_Max_r*) e a moda do valor máximo de brilho no canal de cor vermelha (*Mode_Max_r*). Também é interessante notar que esses dois atributos estão no mesmo grupo da explicação de dados, ou seja, são fortemente correlacionados. Já os atributos mais importantes contra essa classificação foram a mediana dos valores de brilho mínimo do canal de cor cinza (Median_Min_gray) e o desvio padrão dos valores mínimos de brilho no canal de cor vermelha (StdDev_Min_r).

As âncoras não necessariamente utilizam os atributos de maior impacto no modelo, mas escolhem um pequeno conjunto de atributos que, quando combinados, podem ser usados para fazer uma função de decisão que tenha um comportamento semelhante ao modelo nas proximidades da amostra. Isso permite ao usuário verificar as características essenciais que o modelo empregou para tomar a decisão, e compará-las com a forma de diagnóstico feita por especialistas humanos. Isso pode ajudar na análise da confiabilidade do método ou até mesmo na obtenção de novas informações sobre o problema. Então, aqui é apresentado uma explicação *Anchors* para cada classe do problema. Também são fornecidas a cobertura e a precisão de cada regra âncora, que mostram, respectivamente,

a) Explicação do SHAP



b) Explicações do *Anchors*

Predição: CLL (Incerto)

Porque: `StdDev_Avr_r > 10,28 E`
`Median_MajorAxisLengths <= 12,71 E`
`StdDev_Max_b > 20,10 E`
`Mode_MajorAxisLengths <= 6,43`

Quando estas condições são verdadeiras (1,3% dos casos), o modelo prediz CLL em 100% dos casos

---

Porque não FL: `Avr_MinorAxisLengths <= 10,17 E`
`StdDev_StdDev_b > 4,91 E`
`Median_MajorAxisLengths <= 12,71 E`
`Avr_Min_b <= 61,00`

Quando estas condições são verdadeiras (4,5% dos casos), o modelo prediz diferente de FL em 99,3% dos casos

---

Porque não MCL: `StdDev_Avr_r > 10,28 E`
`StdDev_StdDev_r <= 4,83 E`
`StdDev_Max_gray > 22,87`

Quando estas condições são verdadeiras (3,1% dos casos), o modelo prediz diferente de MCL em 100% dos casos

Figura 25 – Explicações locais baseadas nos métodos SHAP e Anchors para a amostra 2 da base de linfomas.

com que frequência essa regra pode ser usada e quão próxima ela se aproxima do comportamento real do modelo.

Com a primeira âncora da Figura 24(b), o usuário é capaz de perceber que, na amostra 1, há uma certa variação do valor médio da cor vermelha ($StdDev\_Avr\_r > 13,14$), a variação no tamanho do eixo menor é baixa ($StdDev\_MajorAxisLengths \leq 3,85$) e a variação do nível de brilho cinza mediano é alto ($StdDev\_Median\_gray > 13,88$). Essas características e seus respectivos valores limites são suficientes para que o modelo classifique uma amostra como CLL em 100% dos casos. Já com a primeira âncora da Figura 25(b), o usuário pode perceber que, na amostra 2, há uma certa variação do valor médio da cor vermelha ($StdDev\_Avr\_r > 10,28$), o comprimento mediano do eixo maior é baixo ($Median\_MajorAxisLengths \leq 12,71$), a variação do valor máximo de brilho no canal de cor azul é alto ($StdDev\_Max\_b$) e o valor mais comum do comprimento do eixo maior é baixo ($Median\_MajorAxisLengths \leq 6,43$). Para qualquer amostra, ter essas mesmas características significa que o modelo irá classificá-la como CLL. Essas informações permitem que o especialista decida se essas classificações são confiáveis ou se

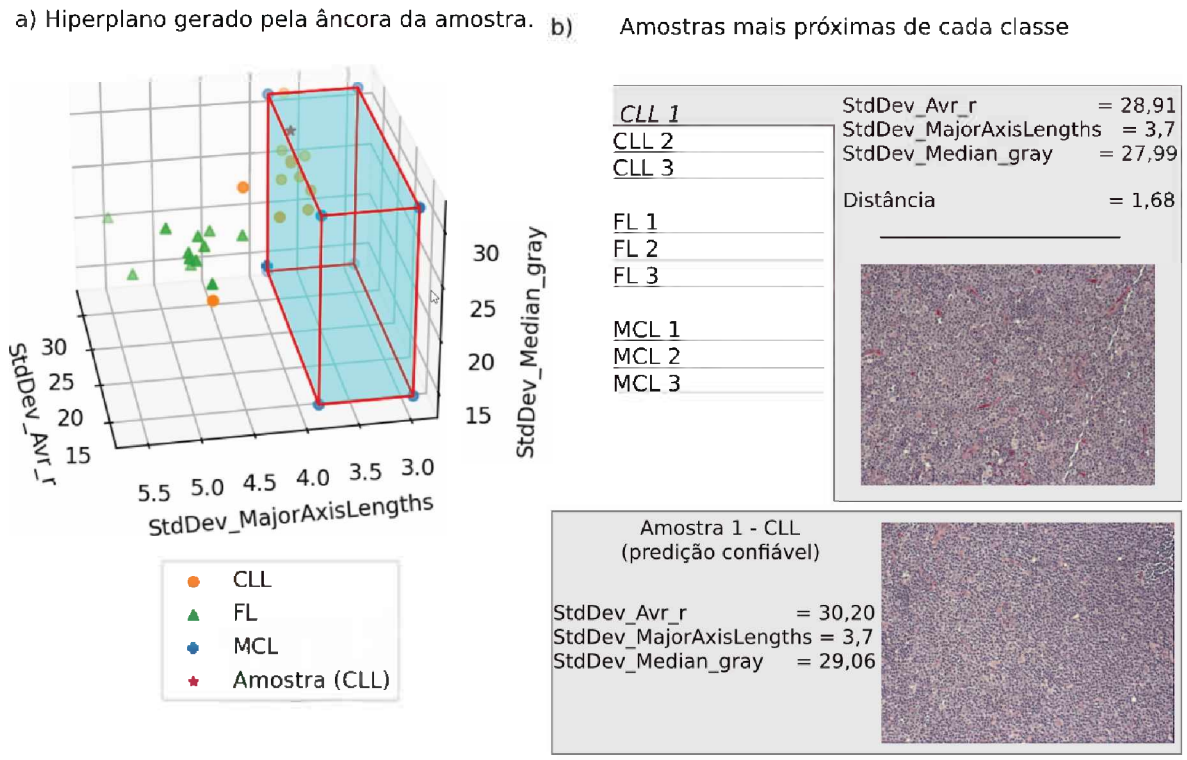


Figura 26 – Visualizações geradas a partir da explicação local do método âncora para a amostra 1 (confiável) do caso de uso de linfoma: (a) visão espacial da explicação do âncora; e (b) dados das três instâncias mais próximas de cada classe e da amostra explicada.

novas análises devem ser realizadas.

Para ajudar o usuário a visualizar como essas âncoras interagem com outras amostras da base de dados, uma representação espacial é apresentada, mostrando os limites espaciais formados pelas condições da âncora e as amostras mais próximas, representadas em cores diferentes para cada classe. Nas Figuras 26 e 27 são apresentadas as representações para as amostras 1 e 2. Para este experimento, foram usados os 25 exemplos mais próximos da amostra. Esse valor foi determinado empiricamente, de modo a possibilitar a visualização e comparação com a vizinhança. Experimentos preliminares mostraram que um valor muito elevado compromete a observação do espaço amostral, devido a sobreposição de pontos. Pela Figura 26(a), é possível observar que a amostra 1, de predição confiável, está localizada em uma região de alta concentração de amostras da mesma classe (CLL), enquanto amostras da classe MCL não aparecem no gráfico. Analisando a Figura 27(a), nota-se que a amostra 2, de predição incerta, está localizada próxima a amostras de outras classes, em uma região de mais difícil separação. As Figuras 26(b) e 27(b) mostram as três instâncias de cada classe de linfoma que estão mais próximas das amostras 1 e 2, respectivamente. Para cada instância apresentada, é possível analisar os valores dos atributos que aparecem regra âncora correspondente; sua imagem histológica; e sua distância até a amostra investigada. Estas informações também são apresentadas

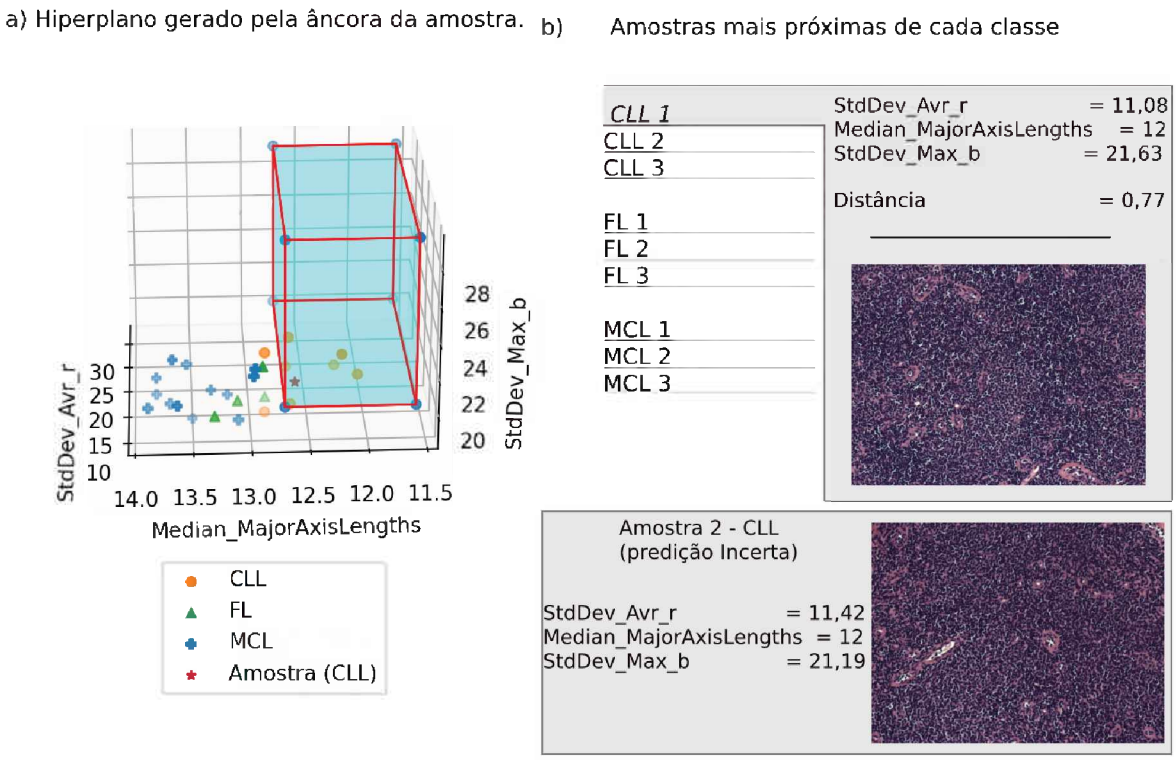
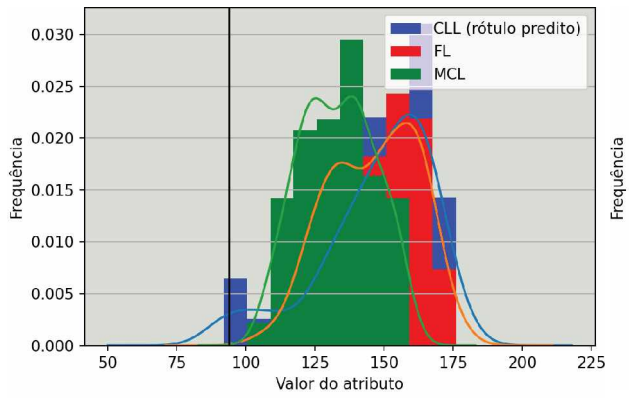


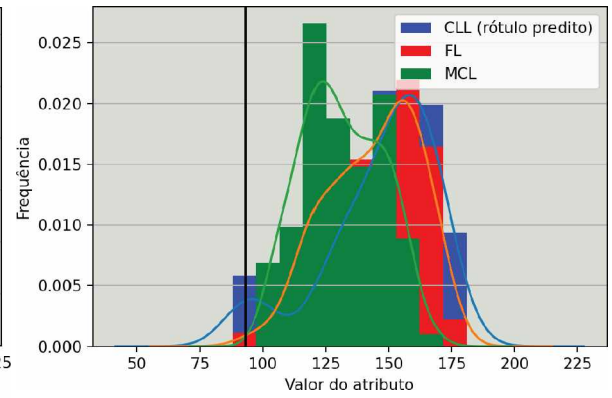
Figura 27 – Visualizações geradas a partir da explicação local do método âncora para a amostra 2 (incerta) do caso de uso de linfoma: (a) visão espacial da explicação do âncora; e (b) dados das três instâncias mais próximas de cada classe e da amostra explicada.

para a amostra em análise e podem ajudar o especialista a entender, por comparação, as as semelhanças e diferenças existentes entre as amostras de cada classe nesta vizinhança, e com isso, saber mais detalhes do porque as amostras foram classificadas como CLL.

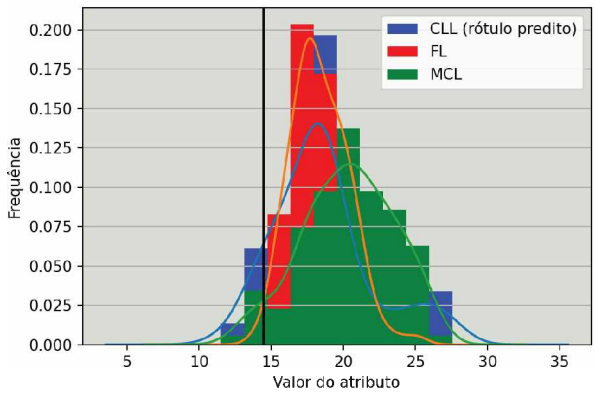
Como a predição da amostra 2 é incerta, uma exploração dos dados guiada pela explicação local é fornecida ao usuário. A Figura 28 mostra os histogramas com a distribuição da quantidade de instâncias de cada classe pelos valores dos atributos mais relevantes, de acordo com a explicação local do SHAP. Em cada histograma, uma linha vertical representada na cor preta marca o valor do atributo na amostra investigada. Na Figura 28(a), pode-se observar que a classe CLL é a única que possui uma quantidade considerável de instâncias com a mediana dos valores máximos de brilho no canal de cor vermelha ($Median\_Max\_r$) próximos ao valor da amostra explicada. Na Figura 28(b), observa-se que há instâncias das classes CLL e FL com a moda dos valores máximos de brilho no canal de cor vermelha ($Mode\_Max\_r$) próximas ao valor da amostra explicada, porém uma quantidade maior dessas instâncias pertence a classe CLL. A Figura 28(c) mostra a distribuição das medianas dos desvios padrões de brilho no canal de cor vermelha ($Median\_StdDev\_r$), indicando que a quantidade de instâncias da classe MCL que possuem esse valor próximo a amostra explicada é menor que a quantidade de instâncias da classe CLL com o mesmo valor. Caso a predição do modelo não seja confiável o suficiente



(a) Histograma - Median_Max_r



(b) Histograma - Mode_Max_r



(c) Histograma - Median_StdDev_r

Figura 28 – Histogramas com a distribuição da quantidade de instâncias de cada classe pelos valores dos atributos mais relevantes, de acordo com a explicação local do SHAP. O valor da amostra sendo explicada é marcado pela linha vertical preta.

para o usuário, por ser categorizada como incerta, estas informações podem ajudá-lo a tomar uma decisão final sobre a classe da doença.

A segunda parte da explicação dos dados guiada pela explicação local complementa os histogramas, mostrando uma visualização espacial das três amostras mais próximas de cada classe, em que os eixos exibidos correspondem aos três atributos de maior impacto na predição da amostra. As imagens relativas as instâncias vizinhas são exibidas, juntamente com os valores dos atributos mais relevantes, e a fim de comparação, a amostra explicada é apresentada com seus respectivos valores para os atributos em questão. A Figura 29 mostra os resultados gerados para a amostra 2, utilizando os mesmos atributos apresentados nos histogramas (Median_Max_r, Mode_Max_r e Median_StdDev_r). Pode-se perceber que a diferença entre os valores da amostra explicada e da instância da classe CLL mais próxima é pequena, e ao comparar as imagens relativas a cada amostra, tem-se observado características semelhantes nessa avaliação. Baseado na experiência do especialista, estas informações podem ser utilizadas para aumentar a confiança na decisão do classificador, e a comparação entre os casos próximos pode ser utilizada pelo especialista para prover explicações adicionais.

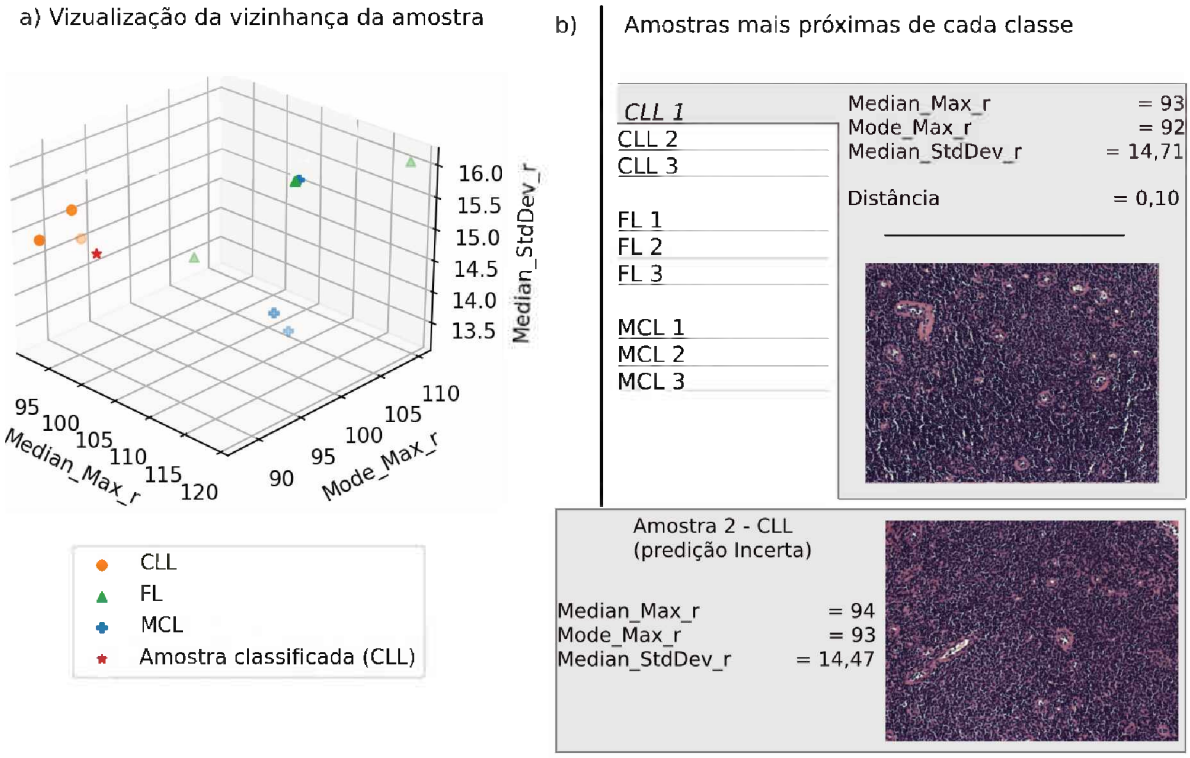


Figura 29 – Visualizações geradas a partir da explicação local do método SHAP para a amostra 2 (incerta) do caso de uso de linfoma: (a) Visão espacial das instâncias mais próximas de cada classe; e (b) dados das três instâncias mais próximas de cada classe e da amostra explicada.

4.3 Validação da Metodologia na Classificação de Displasias

Nesta segunda parte do capítulo, os experimentos realizados para o caso de uso de linfomas serão repetidos, com exceção daqueles utilizados para escolha de parâmetros e técnicas. O objetivo desses experimentos é avaliar se a metodologia proposta pode ser generalizada para um problema diferente, sem a necessidade de grandes alterações. Por isso, os parâmetros e métodos de avaliação empregados foram os mesmos.

Nas 50 execuções, utilizadas para a avaliação do desempenho do classificador (10 execuções da validação cruzada estratificada com 5 partições), o método proposto teve uma acurácia média de 92,1%, com desvio padrão de 2% e um *F-score* de 0,920, utilizando como algoritmo de classificação a MLP e aplicando o filtro de correlação de Pearson. Este desempenho foi próximo do alcançado na classificação de linfomas (taxa de acerto de 0,928). O filtro baseado na correlação de Pearson retirou em média 14,4 atributos da base de dados, o que corresponde a uma redução de aproximadamente 13%, menos da metade do valor retirado no caso de linfomas (30%).

Na Tabela 9 é apresentada uma comparação do desempenho do classificador com abordagens encontradas na literatura que também classificam lesões orais. Em comparação

com a tabela referente aos trabalhos correlacionados ao LNH, foi incluída a coluna "Tipo de lesão", que indica qual o tipo de lesão oral a metodologia se propõe a classificar. Ela foi adicionada pois, como há poucos trabalhos no estado da arte que tratam da classificação de DOE, optou-se por comparar a metodologia com trabalhos que classificam outros tipos de lesões orais. Também foi adicionada a coluna "Tipo Class.", que indica o tipo de problema de classificação tratado pelo trabalho. Os métodos encontrados no estado da arte realizam apenas classificações entre duas classes de displasia, e o método proposto atribui às amostras uma entre quatro possíveis classes. Além disso, essa tabela não indica o valor de área sobre a curva *ROC* dos métodos, pois esse valor não foi disponibilizado por nenhum dos trabalhos citados. O valor dessa métrica para o classificador proposto foi de 0.990.

O método proposto obteve uma métrica de desempenho melhor que o método de Baik et al. (2014), e valores próximos das outras abordagens propostas para classificação de DOE (ADEL et al., 2018; SILVA et al., 2022). Os descritores extraídos das imagens são de mais fácil interpretação que os utilizados nas abordagens de Baik et al. (2014); Adel et al. (2018); e (SILVA et al., 2022).

Utilizando a mesma metodologia adotada na análise dos métodos de explicação, na qual a base de dados é dividida em um conjuntos de treinamento e teste, sem empregar a validação cruzada, o modelo de classificação alcançou uma taxa de acerto média de 90%. Essa diferença no valor da acurácia acontece pela variação inerente aos métodos estocásticos utilizados, e pela diferença dos conjuntos de treinamento e teste. Durante a construção do modelo, nenhum atributo da base teve sua variância igual a zero. No entanto, o filtro baseado na correlação de Pearson permitiu remover 15 atributos da base de dados.

Tabela 9 – Comparação do método de classificação multiclasse de DOE proposto com trabalhos relacionados.

Ref.	Tipo de lesão	Extração de atributos	Qtd.	Classificador	Tipo Class.	ACC
(KRISHNAN et al., 2012)	Fibrose submucosa oral	Descritores morfológicos e de textura no canal de cor cinza	18	SVM	Binária	99,66%
(BAIK et al., 2014)	Lesões orais pré malignas	Descritores morfológicos, quantidade de DNA, distribuição de cromatina	110	Random Forest	Binária	80%
(ADEL et al., 2018)	DOE	ORB, SIFT	16	SVM	Binária	92,8%
(SILVA et al., 2022)	DOE	Descritores morfológicos, entropia, índice de Moran	23	classificador polinomial	Binária	92,4%
Método Proposto	DOE	Descritores morfológicos e não-morfológicos	99	MLP	Multiclasse	92,1%

4.3.1 Avaliação do Método de Estimação da Confiança

Os resultados dos experimentos com o método de estimação da confiança da classificação aplicado ao problema de displasia foram bem diferentes de quando aplicado ao problema de classificação de linfomas. Como pode ser visto na Tabela 10, uma quantidade maior de amostras foi categorizada como confiável, e nenhuma amostra foi categorizada como inconclusiva.

Tabela 10 – Avaliação do método de estimação de confiabilidade das predições na base de displasia.

Categoria	% de Amostras	Acurácia	Probabilidade MLP
Confiável	96,9%	93,5%	0,959
Incerta	3,0%	56,8%	0,617
Inconclusiva	0	-	-

Um fator importante a ser destacado em relação a essa base é que as características dos núcleos capturadas nesse tipo de lesão são diferentes. A taxa de acerto do modelo multiclasse nesse experimento foi de 92,4%. Como esse problema tem 4 classes, foram necessários 6 modelos binários auxiliares, um para cada par de classes: Saudável e Leve; Saudável e Moderado; Saudável e Severo; Leve e Moderado; Leve e Severo; e Moderado e Severo. Suas taxas de acerto, considerando apenas amostras de teste, foram de, respectivamente, 99,2%; 99,3%; 99,3%; 92,4%; 92,2%; e 96,4%. Nas colunas Acurácia e Probabilidade MLP da Tabela 10, percebe-se que predições da categoria Incerta têm menor acurácia média e menor estimativa da probabilidade de classificação calculada pelo modelo baseado em MLP, mesmo comportamento observado no linfoma. Por isso, pode-se dizer que o método de estimação de confiabilidade proposto também têm bons resultados com este segundo caso de uso.

4.3.2 Exploração dos Dados

Utilizando o mesmo limite de correlação de 0,97 utilizado na base de linfomas, o método proposto formou 17 grupos de atributos, o quais são apresentados na Tabela 11. No total, 37 atributos, dos 99 presentes no experimento apareceram nestes grupos. Neste caso de uso também há ocorrências de atributos dos canais de intensidade de brilho verde e cinza no mesmo grupo, porém nesse conjunto de dados, também houve a formação de grupos com misturas de outras cores, como vermelho e cinza; e azul, verde e cinza.

Outra informação importante que pode ser obtida com a correlação de Pearson é a correlação entre cada atributo e a variável alvo do problema. Na Figura 30 é apresentado a correlação entre os atributos para as classes de displasia. O atributo com maior correlação foi a métrica média obtida por meio do calculo da média do brilho do canal de cores verde

Tabela 11 – Grupos de descritores do caso de uso de displasias com alta correlação de Pearson entre si (> 0,97).

Avr_Areas Avr_EquivDiameters	Avr_Min_r Median_Min_r	Avr_Avr_r Median_Avr_r
Avr_StdDev_r Avr_StdDev_gray	Avr_Max_g Avr_Max_gray Median_Max_g	Avr_Min_b Median_Min_b
Avr_Avr_b Median_Avr_b	Avr_StdDev_b Median_StdDev_b	Avr_Min_gray Median_Min_gray
Median_StdDev_r Median_StdDev_gray	Median_Avr_g Median_Median_g	Median_Avr_gray Median_Median_gray
StdDev_Avr_r StdDev_Median_r	StdDev_Min_g StdDev_Min_gray	StdDev_Max_g StdDev_Max_gray
StdDev_Avr_g StdDev_Median_g StdDev_Avr_gray	Mode_Avr_g Mode_Avr_b Mode_Avr_gray	

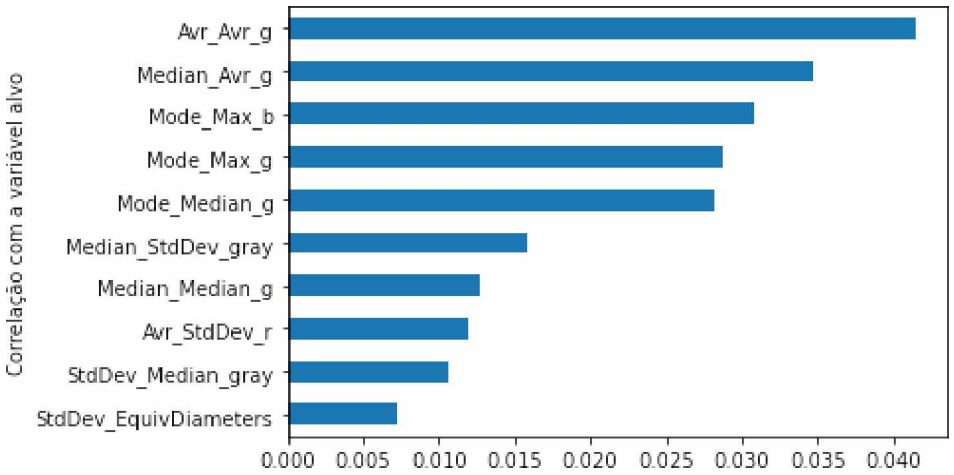


Figura 30 – Atributos com maior correlação com a classe do problema no caso de uso de displasias.

(*Avr_Avr_g*), com valor próximo a 0,04. Isso representa metade do valor calculado para o atributo de maior correlação do caso de uso de linfomas. Este resultado indica que, para este caso de uso, é ainda mais difícil reduzir o número de atributos utilizados na classificação sem prejudicar os resultados, já que os atributos, isoladamente, têm pouca capacidade preditiva. Apenas um dos 10 atributos mais correlacionados com a variável alvo é morfológico, indicando um menor valor preditivo deste grupo de descritores. Neste caso de uso, também houve um aumento na correlação dos atributos provenientes do canal de cor verde entre os mais importantes (cinco atributos entre os 10 melhores, enquanto que na análise dos dados de linfomas havia apenas um).

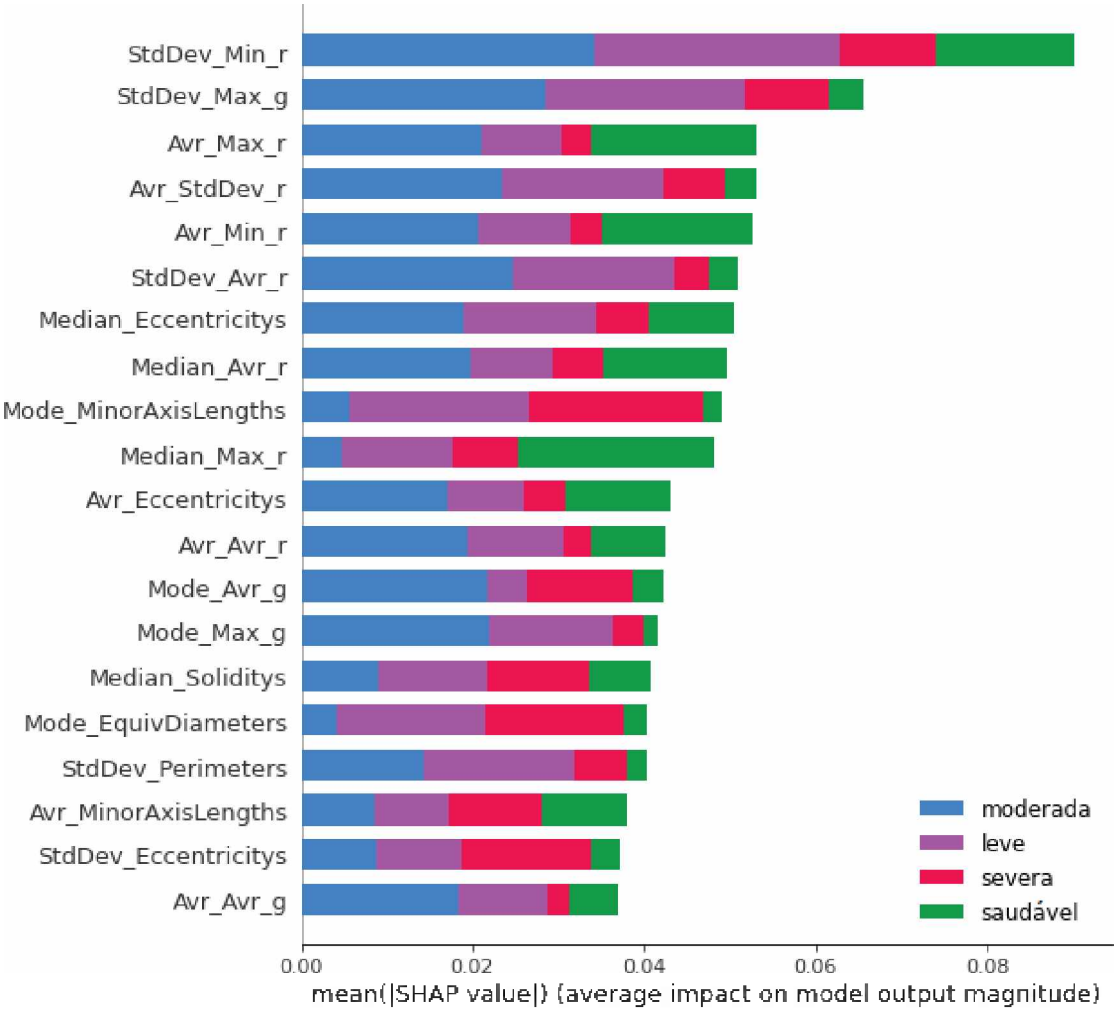


Figura 31 – Impacto médio dos 20 atributos mais impactantes no modelo de predição, especificado por cada classe predita, no caso de uso de classificação de imagens histológicas de displasias.

4.3.3 Análise do Modelo Baseada em Explicações Globais

Para o modelo de classificação de imagens histológicas de displasias, o atributo mais importante, segundo a explicação do SHAP, foi obtido pelo desvio padrão dos valores mínimos de intensidade de brilho nos núcleos celulares no canal de cor vermelha (StdDev_Min_r). Conforme pode ser observado na Figura 31, de acordo com a análise do SHAP, dos 20 atributos mais importantes para as predições do modelo, oito são descritores morfológicos e doze são não-morfológicos. Dentre os descritores não-morfológicos, oito foram obtidos do canal vermelho. Isso pode indicar que essa cor é importante para uma tomada de decisão no caso da displasia. Já o canal de cor verde apareceu em 4 atributos, e a intensidade de brilho azul e cinza não apareceram dentre as 20 características mais importantes.

O experimento destinado a análise da capacidade de seleção de atributos do método SHAP também foi conduzido para o modelo de displasias, e os resultados são apresentados

Tabela 12 – Acurácia média de modelos baseado em MLP no caso de uso de displasias utilizando diferentes seletores de atributos.

Atributos	PCA	ANOVA	SHAP
80% (91)	0.920 ±0.04	0.930 ±0.04	0.924 ±0.04
60% (68)	0.916 ±0.06	0.910 ±0.05	0.916 ±0.05
40% (45)	0.893 ±0.05	0.927 ±0.04	0.882 ±0.06

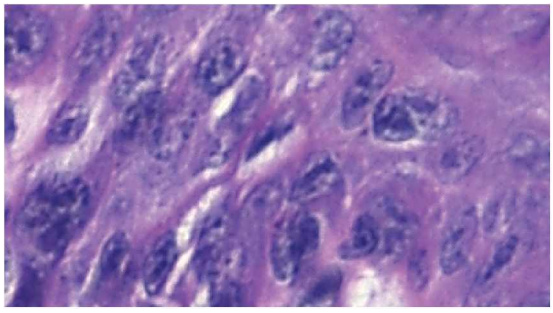


Figura 32 – Imagem histológica de displasia severa, classificada corretamente pelo modelo, mas categorizada como incerta.

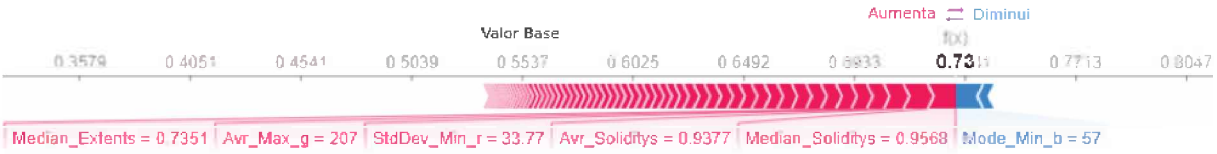
na Tabela 12. O método obteve resultados mais relevantes nos grupos com 80% e 60% do número de descritores. Isso mostra o mesmo comportamento do caso investigado com linfomas. No entanto, quando houve uma redução para 40% dos descritores, o desempenho do modelo não manteve padrões semelhantes ao linfoma. Nesse caso, pode-se observar que essa redução provocou uma queda maior em relação do desempenho.

4.3.4 Análise de Predições Baseado em Explicação Local

Para analisar a explicação local da metodologia proposta no caso de uso de displasias, uma amostra da base de testes foi selecionada. Esta amostra pertence a classe de displasias severas, e foi corretamente classificada. Esta predição foi categorizada como incerta, e apenas a classe de tecido saudável foi descartada, já que nenhum modelo de classificação binária classificou a amostra como saudável. A Figura 32 mostra a imagem histológica da amostra.

A Figura 33 mostra as explicações locais da predição baseadas em SHAP e *Anchors*. Pode-se perceber pela explicação local do SHAP na Figura 33(a) que os descritores baseados no valor mediano da solidez ($Median\_Solidity s$) e na métrica média da solidez ($Avr\_Solidity$) foram os que mais impactaram para que esta amostra fosse classificada como Severa, enquanto a mediana dos valores de diâmetro equivalente ($Median\_EquivDiameters$) e a moda dos valores de brilho mínimo no canal de cor azul ($Mode\_Min\_b$) tiveram impacto negativo nesta decisão. O último atributo não pode ser visualizado na figura por uma limitação de espaço, porém, esta explicação pode ser exibida em um ambiente interativo, possibilitando ao usuário selecionar qualquer segmento das barras vermelha e azul e verificar os dados do atributo correspondente. Apenas dois

a) Explicação do SHAP



b) Explicações do Anchors

Predição: Severa (Incerta)

Porque:
```
StdDev_Avr_r <= 27,51 E
Median_Solidity s <= 0,97 E
StdDev_StdDev_b > 8,38
```

Quando estas condições são verdadeiras (0,9% dos casos), o modelo prediz Severa em 100% dos casos

---

Porque não Saudável:
```
Median_Max_r > 193,25 E
Avr_Areas <= 939,03
```

Quando estas condições são verdadeiras (15,3% dos casos), o modelo prediz diferente de Saudável em 100% dos casos

---

Porque não Leve:
```
Median_Solidity s <= 0,97 E
StdDev_Min_r > 29,85
```

Quando estas condições são verdadeiras (14,4% dos casos), o modelo prediz diferente de leve em 100% dos casos

---

Porque não Moderada:
```
Mode_Extents > 0,56 E
Avr_StdDev_b > 24,33
```

Quando estas condições são verdadeiras (14,7% dos casos), o modelo prediz diferente de moderada em 99,1% dos casos

Figura 33 – Explicações locais dos métodos a) SHAP; e b) *Anchors*, para a amostra do caso de uso de displasias.

atributos da amostra contribuíram contra a predição da classe Severa. Isso pode indicar que a categorização da amostra como incerta pode ter sido causada por erro de generalização em algum dos classificadores binários auxiliares. O especialista pode utilizar essas informações para sua tomada de decisão.

Com a primeira âncora da Figura 33(b), o usuário é capaz de perceber que, na amostra analisada, o desvio padrão do valor de brilho médio no canal de cor vermelha (*StdDev_Avr_r*) é baixo, a mediana do valor de solidez (*Median_Solidity s*) é baixo, e o desvio padrão da métrica desvio padrão de todos os núcleos pertencentes a imagem no canal de cor azul (*StdDev_StdDev_b*) é alto. Essas condições são verdadeiras para a amostra analisada, e outras amostras que também tiverem esses atributos dentro dos limites estabelecidos, também são classificadas como severas pelo modelo preditivo.

A visualização da regra do *Anchors*, juntamente com as 25 instâncias mais próximas

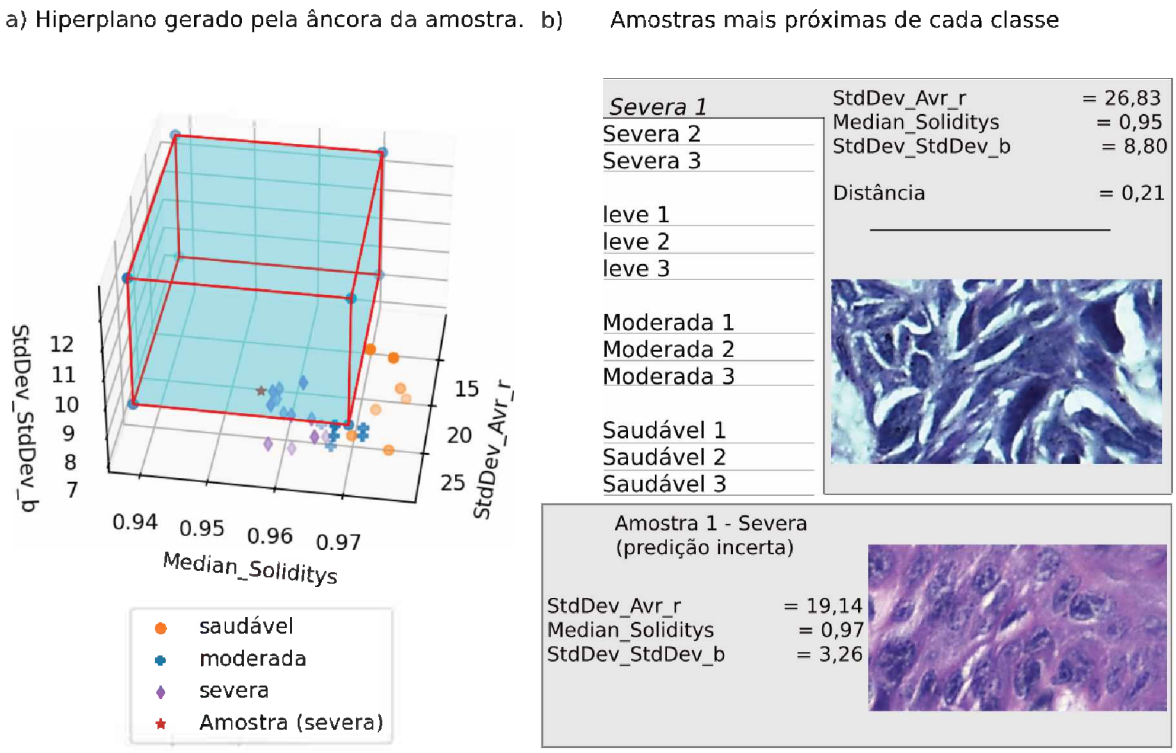
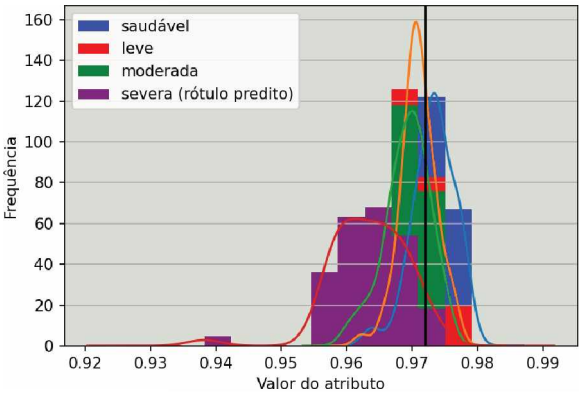


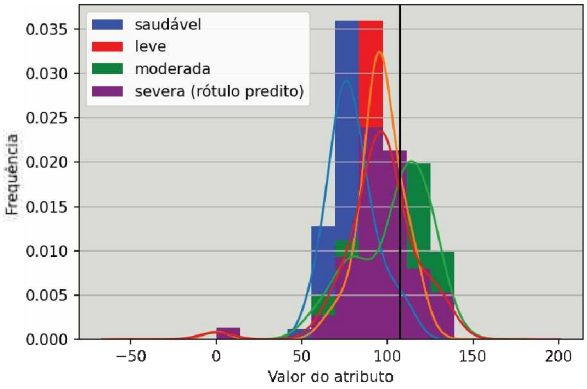
Figura 34 – Visualizações geradas a partir da explicação local do método âncora para displasias: (a) visão espacial da explicação do âncora; e (b) dados das três instâncias mais próximas de cada classe e da amostra investigada.

da amostra estão apresentadas na Figura 34(a), enquanto os valores dos atributos e as imagens histológicas das três amostras mais próximas de cada classe estão apresentadas na Figura 34(b). Neste exemplo, não houve nenhuma amostra da classe de displasias leves presentes nas 25 instâncias da vizinhança. A instância mais próxima da amostra tem valor próximo de mediana da solidez (*Median_Solidity*), mas valores maiores de desvio padrão do valor médio de brilho no canal de cor vermelho (*StdDev_Avr_r*); e da métrica desvio padrão de informações de desvio padrão dos núcleos presentes na imagem do valor de brilho da cor azul (*StdDev_StdDev_b*).

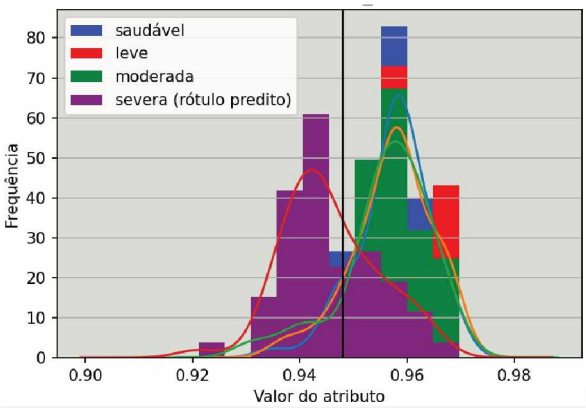
Na Figura 35 são apresentados os histogramas com a distribuição da quantidade de instâncias de cada classe pelos valores dos atributos mais relevantes, de acordo com a explicação local do SHAP. A Figura 35(a) mostra que a métrica mediana dos valores de solidez da amostra (*Median_Soliditys*) tem um valor dentro de um intervalo comum às classes saudável, leve e moderada, porém entre as instâncias da classe predita (severa), esse intervalo de valor, representado pela barra de cor roxa é inferior, o que pode indicar que houve algum problema no treinamento do modelo ou que o impacto deste atributo na predição esta relacionado a valores de outros atributos. Na Figura 35(b), o histograma indica que o atributo moda dos valores mínimos de brilho no canal de cor azul da amostra (*Mode_Min_b*) está em uma faixa de valores que é mais frequente na classe severa do que em qualquer outra. Na Figura 35(c), pode-se observar que a distribuição dos



(a) Histograma - Median_Soliditys



(b) Histograma - Mode_Min_b



(c) Histograma - Avr_Soliditys

Figura 35 – Histogramas do caso de uso de displasias com a distribuição da quantidade de instâncias de cada classe pelos valores dos atributos mais relevantes, de acordo com a explicação local do SHAP. O valor da amostra sendo explicada é marcado pela linha vertical preta.

valores de média de solidez (*Avr_Soliditys*) na classe severa está invertida em relação as distribuições das outras classes, ou seja, enquanto a distribuição da métrica para a classe severa apresenta uma assimetria para a esquerda (valores em torno de 0,94), a distribuição da métrica para as demais classes possui uma assimetria para a direita (valores em torno de 0,96). Pelo histograma também é possível notar que o valor da amostra investigada está próximo do ponto de encontro entre as curvas de distribuição.

Utilizando os mesmos atributos (os três mais relevantes para a decisão da amostra) apresentados nos histogramas da Figura 35, é construído uma visão tridimensional da vizinhança da amostra investigada. Pode-se observar pela Figura 36(a) que as instâncias estão espalhadas pelo espaço amostral. Visualmente, é difícil dizer quais estão mais próximas, assim como determinar a qual classe a amostra pertence. Porém, utilizando as informações detalhadas fornecidas pelo método (Figura 36(b)), percebe-se que a distância para a instância mais próxima da classe moderada é de 0,44, enquanto que a menor distância para uma instância da classe severa é de 0,54. Essa situação pode indicar que a amostra se encontra próxima a borda de divisão entre essas classes, mas ainda do lado da

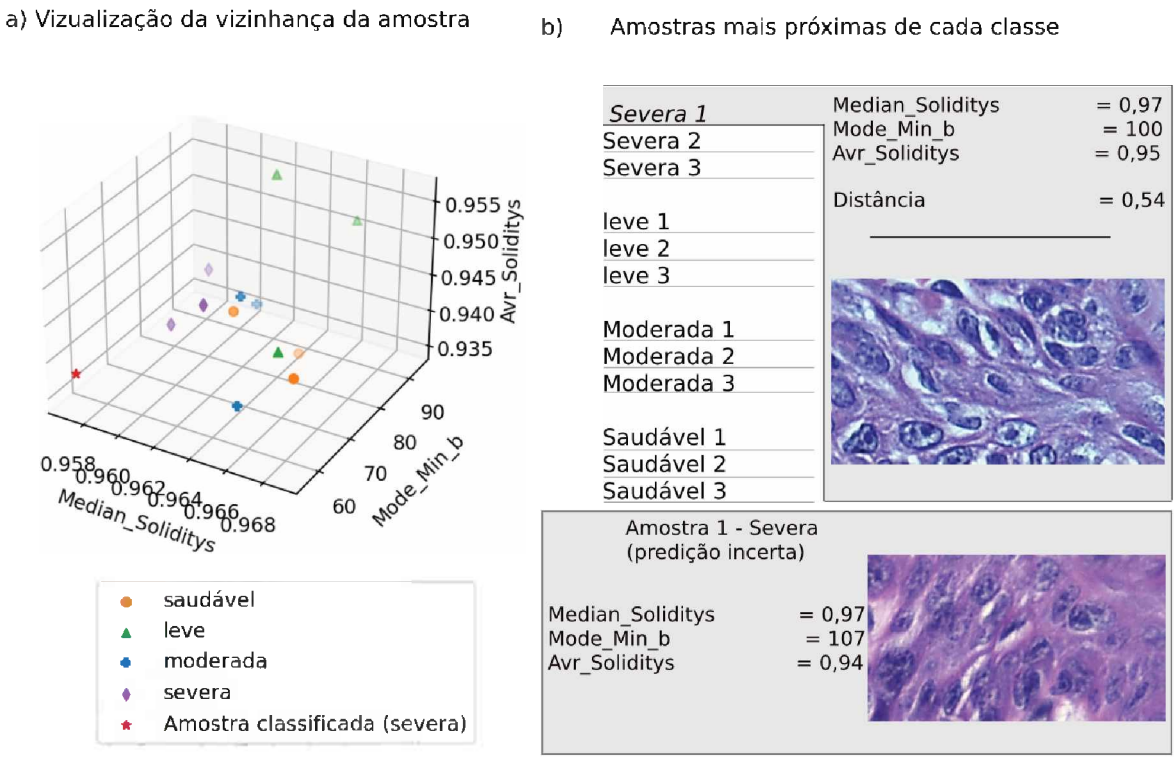


Figura 36 – Visualizações geradas a partir da explicação local do método SHAP para a amostra (incerta) de displasia: (a) Visão espacial das instâncias mais próximas de cada classe; e (b) dados das três instâncias mais próximas de cada classe e da amostra explicada.

classe predita. Uma análise comparativa entre a amostra investigada e os vizinhos mais próximos dessas duas classes pode indicar o motivo dessa região do espaço amostral ser de difícil classificação, o que auxiliaria na melhora do sistema.

Capítulo 5

Conclusões

Neste trabalho, foi apresentado um método de classificação de imagens histológicas que emprega descritores interpretáveis e fáceis de extrair e uma abordagem multiclasse. A combinação de diferentes técnicas de XAI permite que o usuário interprete os dados, o modelo e as previsões individuais, e possibilita um diagnóstico mais informado, intuitivo e confiável pelo especialista humano. Nossa abordagem mostrou-se relevante, retornando resultados importantes na tarefas de classificação multiclasse de imagens histológicas de pacientes com linfomas e de imagens histológicas com ou sem displasia oral epitelial. Os resultados experimentais indicam que é possível obter um desempenho importante em relação aos trabalhos presentes na literatura utilizando um conjunto de descritores morfológicos e não morfológicos. Utilizando um classificador baseado em MLP, os resultados de acurácia médio foram de 92,8% e 92,1% para as imagens de linfoma e displasia, respectivamente. Também foi proposta uma combinação de diferentes técnicas de explicação, o que pode resultar em uma melhor interpretação do sistema.

Uma exploração dos dados foi realizada utilizando a correlação de *Pearson*, indicando grupos de atributos correlacionados. Essas informações se mostraram úteis na identificação de padrões de correlação nos dados, o que pode ser utilizado como base para futuras investigações sobre as características da doença e seu diagnóstico. Uma metodologia de categorização de predições foi proposta, e se mostrou eficiente em indicar quais predições têm maior chance de estarem corretas. Isso possibilitou a utilização de uma metodologia de recomendação de análises e explicações diferentes para predições com menor probabilidade de acerto, assim oferecendo mais informação para auxiliar o especialista na decisão do diagnóstico. Uma abordagem de explicação do modelo de classificação utilizando o SHAP foi utilizada para apresentar o impacto médio de cada atributo nas suas decisões. Nos experimentos realizados, essa informação foi importante para analisar quais grupos de descritores têm maior importância, o que poderia ser utilizado para investigar se o classificador utiliza algum tipo de informação não condizente com o problema tratado, que poderia ter sido incluído dos dados de treinamento por algum erro de coleta. Explicações locais utilizando âncoras e um gráfico de forças construído pelo SHAP foram

utilizados para mostrar como os atributos se relacionam e influenciam a decisão. Durante a análise dos experimentos, foi demonstrado como essas explicações podem ser utilizadas para entender melhor a decisão do classificador, e com isso, aumentar a confiabilidade no sistema. Para complementar a interpretação proporcionada pela âncora, foi utilizada uma visualização espacial de sua regra, permitindo que, de forma interativa, o usuário investigue amostras vizinhas que seriam classificadas da mesma forma, ou contra exemplos, que seriam rotulados de forma diferente a amostra investigada.

Durante a execução dos experimentos, também foi possível observar limitações da metodologia proposta. O conjunto de descritores utilizados, apesar de se basear em características de forma e cor dos núcleos, não são perfeitamente interpretáveis. As características morfológicas baseadas em tamanhos foram calculadas em quantidades de pixel. Nestes casos, é possível fazer análises comparativas entre células de imagens registradas com a mesma ampliação, ou que foram propriamente ajustadas em pré-processamento, porém pode ser difícil utilizar conhecimentos prévios, obtidos em outra escala ou medida. Dentre os descritores não morfológicos, também há alguns de difícil interpretação, ocasionados pelo uso de duas operações estatísticas no cálculo de cada característica. Diversas abordagens para redução do número de atributos foram investigadas, porém não foi possível encontrar um método que diminuísse significativamente o tamanho do vetor de características sem afetar o desempenho da classificação das amostras. Por isso, o sistema proposto utiliza uma quantidade relativamente grande de atributos, o que dificulta as análises das explicações. A abordagem proposta para categorização das predições necessita da construção de modelos binários, o que acarreta em um aumento no custo de processamento e armazenamento do sistema.

Com base nas análises de interpretação realizadas neste trabalho, é possível melhorar a compreensão dos resultados para o especialista médico sobre o comportamento de um modelo multiclasse. O uso dos métodos também pode auxiliar no aprendizado da relação desses descritores com os tipos das doenças analisadas, assim como contribuir para investigação de classificadores mais robustos e complexos em sistemas CAD. Os experimentos foram avaliados em um conjunto de dados público de 375 imagens de imagens de NHL e em um conjunto de 296 imagens de DOE. Com essas diferentes bases, os métodos propostos foram capazes de obter uma classificação relevante para um sistema CAD. Acredita-se que essa abordagem pode ser útil para aplicação em ambientes de rotina diagnóstica contribuindo como uma solução de leitura complementar a decisão do especialista.

5.1 Principais Contribuições

Este trabalho contribui com a apresentação de uma metodologia nova de criação de um sistema CAD explicável, combinando diferentes técnicas já existentes com ferramentas para adaptá-las ao contexto de imagens histológicas multiclasse. Foi apresentada nesta

dissertação soluções para diversos problemas encontrados para realizar estas adaptações, além de técnicas de visualização que ajudaram as explicações a se tornarem ainda mais interpretáveis.

Por ter se mostrado um método robusto o suficiente para ser utilizado em um segundo caso de uso sem a necessidade de adaptações ou mudanças de parâmetros, este trabalho fornece uma metodologia que pode ser utilizada posteriormente em diversos tipos de problemas.

Como a área de classificação de imagens histológicas de linfomas e displasias apresenta uma escassez de publicações focadas em XAI, este trabalho se apresenta com um estudo inicial nessa área. Espera-se que essa abordagem possa trazer avanços científicos para área e novas contribuições possa surgir a partir dessa ferramenta.

5.2 Trabalhos Futuros

Apesar da maioria dos descritores morfológicos e não morfológicos utilizados serem interpretáveis, há ainda muitos que dependem da habilidade de abstração do usuário. As medidas de distância são calculadas em pixels, e as comparações entre cores são feitas utilizando números em uma escala de 0 a 255. Por isso, utilizar descritores ainda mais fáceis de serem interpretados é uma importante tarefa para trabalhos futuros.

Com o foco na criação de uma metodologia interpretável, foram utilizadas neste trabalho técnicas de classificação e engenharia de atributos simples. Assim, trabalhos futuros também podem avaliar novas técnicas de classificação e parametrização dos métodos, o que deve melhorar ainda mais o desempenho do classificador, sem comprometer sua interpretabilidade.

Dentre as técnicas de explicação abordados, a explicação dos dados pode ser a que possui mais trabalhos já publicados, encontrados na área de análise de dados exploratória, já que ela existe antes do termo XAI. Por isso, em trabalhos futuros pretende-se realizar novas investigações para melhorar a metodologia de explicação dos dados.

Os métodos de explicação local *pós-hoc* geralmente são projetados para problemas de classificação binária, ou de regressão. Neste trabalho, foram adotadas técnicas e realizadas adaptações para esses métodos de explicação binária serem empregados em um contexto de abordagem multiclasse. Investigar uma nova abordagem para tratar as explicações multiclasse pode ser importante e trazer contribuições para a área.

5.3 Contribuições em Produção Bibliográfica

Os seguintes artigos foram publicados em conferências como resultados obtidos da pesquisa desenvolvida durante o mestrado:

- FARIA, T. P. de; NASCIMENTO, M. Z. do; MARTINS, L. G. A.. A Method For Multiclass Lymphoma Classification Based on Morphological and Non-Morphological Descriptors. Anais do WORKSHOP DE VISÃO COMPUTACIONAL (WVC), Sociedade Brasileira de Computação (SBC), páginas 184-189, 2021.
- FARIA, T. P. de; NASCIMENTO, M. Z. do; MARTINS, L. G. A.. Understanding the multiclass classification of lymphomas from simple descriptors. 2021 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, páginas 1202-1208, 2021.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics**, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. Disponível em: <https://doi.org/10.1002/wics.101>.

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). **IEEE access**, IEEE, v. 6, p. 52138–52160, 2018. Disponível em: <https://doi.org/10.1109/ACCESS.2018.2870052>.

ADEL, D.; MOUNIR, J.; EL-SHAFEY, M.; ELDIN, Y. A.; MASRY, N. E.; ABDELRAOUF, A.; ELHAMID, I. S. A. Oral epithelial dysplasia computer aided diagnostic approach. In: IEEE. **2018 13th International Conference on Computer Engineering and Systems (ICCES)**. 2018. p. 313–318. Disponível em: <https://doi.org/10.1109/ICCES.2018.8639452>.

AGGARWAL, C. C. **Data mining: the textbook**. Springer, 2015. Disponível em: <https://doi.org/10.1007/978-3-319-14142-8>.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LÓPEZ, S.; MOLINA, D.; BENJAMINS, R. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. **Information Fusion**, Elsevier, v. 58, p. 82–115, 2020. Disponível em: <https://doi.org/10.1016/j.inffus.2019.12.012>.

BAI, J.; JIANG, H.; LI, S.; MA, X. Nhl pathological image classification based on hierarchical local information and googlenet-based representations. **BioMed research International**, 2019. Disponível em: <https://doi.org/10.1155/2019/1065652>.

BAIK, J.; YE, Q.; ZHANG, L.; POH, C.; ROSIN, M.; MACAULAY, C.; GUILLAUD, M. Automated classification of oral premalignant lesions using image cytometry and random forests-based algorithms. **Cellular Oncology**, Springer, v. 37, n. 3, p. 193–202, 2014. Disponível em: <https://doi.org/10.1007/s13402-014-0172-x>.

BANIECKI, H.; BIECEK, P. The grammar of interactive explanatory model analysis. **arXiv preprint**, 2020. Disponível em: <https://arxiv.org/abs/2005.00497>.

BECK, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. **Science Translational Medicine**, American Association for the Advancement of Science, v. 3, n. 108, p. 108–113, 2011. Disponível em: <https://doi.org/10.1126/scitranslmed.3002564>.

BENESTY, J.; CHEN, J.; HUANG, Y. On the importance of the pearson correlation coefficient in noise reduction. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 16, n. 4, p. 757–765, 2008. Disponível em: <https://doi.org/10.1109/TASL.2008.919072>.

BENTAIEB, A.; HAMARNEH, G. Adversarial stain transfer for histopathology image analysis. **IEEE Transactions on Medical Imaging**, v. 37, n. 3, p. 792–802, 2018. Disponível em: <https://doi.org/10.1109/TMI.2017.2781228>.

BI, Y.; XIANG, D.; GE, Z.; LI, F.; JIA, C.; SONG, J. An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. **Molecular Therapy-Nucleic Acids**, Elsevier, v. 22, p. 362–372, 2020. Disponível em: <https://doi.org/10.1016/j.omtn.2020.08.022>.

BIECEK, P.; BURZYKOWSKI, T. **Explanatory model analysis: Explore, explain and examine predictive models**. Chapman and Hall/CRC, 2021. Disponível em: <https://doi.org/10.1201/9780429027192>.

BIESIADA, J.; DUCH, W. Feature selection for high-dimensional data—a pearson redundancy based filter. In: **Computer recognition systems 2**. Springer, 2007. p. 242–249. Disponível em: <https://doi.org/10.1007/978-3-540-75175-5_30>.

BISHOP, C. M. **Pattern recognition and machine learning**. springer, 2006. Disponível em: <https://dl.acm.org/doi/10.5555/1162264>.

BRAZIL, M. da Saúde Instituto Nacional do C. **Estimativa 2020: Incidência de Câncer no Brasil**. 2020. Disponível em: <https://www.inca.gov.br/publicacoes/livros/estimativa-2020-incidencia-de-cancer-no-brasil>.

CODELLA, N.; MORADI, M.; MATASAR, M.; SVEDA-MAHMOOD, T.; SMITH, J. R. Lymphoma diagnosis in histopathology using a multi-stage visual learning approach. In: **Medical Imaging 2016: Digital Pathology**. [s.n.], 2016. v. 9791, p. 97910H. Disponível em: <https://doi.org/10.1117/12.2217158>.

CONFALONIERI, R.; COBA, L.; WAGNER, B.; BESOLD, T. R. A historical perspective of explainable artificial intelligence. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 11, n. 1, p. e1391, 2021. Disponível em: <https://doi.org/10.1002/widm.1391>.

DEMIR, C.; YENER, B. Automated cancer diagnosis based on histopathological images: a systematic survey. **Rensselaer Polytechnic Institute, Tech. Rep**, Citeseer, 2005. Disponível em: <https://www.researchgate.net/publication/228640139_Automated_cancer_diagnosis_based_on_histopathological_images_A_systematic_survey>.

DIF, N.; ELBERRICHI, Z. Efficient regularization framework for histopathological image classification using convolutional neural networks. **International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)**, IGI Global, v. 14, n. 4, p. 62–81, 2020. Disponível em: <https://doi.org/10.4018/IJCINI.2020100104>.

DING, H.; FENG, P.-M.; CHEN, W.; LIN, H. Identification of bacteriophage virion proteins by the anova feature selection and analysis. **Molecular BioSystems**, Royal Society of Chemistry, v. 10, n. 8, p. 2229–2235, 2014. Disponível em: <https://doi.org/10.1039/C4MB00316K>.

DOI, K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. **Computerized medical imaging and graphics**, Elsevier, v. 31, n. 4-5, p. 198–211, 2007. Disponível em: <https://doi.org/10.1016/j.compmedimag.2007.02.002>.

DOST, F.; CAO, K. L.; FORD, P.; ADES, C.; FARAH, C. Malignant transformation of oral epithelial dysplasia: a real-world evaluation of histopathologic grading. **Oral surgery, oral medicine, oral pathology and oral radiology**, Elsevier, v. 117, n. 3, p. 343–352, 2014. Disponível em: <https://doi.org/10.1016/j.oooo.2013.09.017>.

ELSHAWI, R.; AL-MALLAH, M. H.; SAKR, S. On the interpretability of machine learning-based model for predicting hypertension. **BMC medical informatics and decision making**, BioMed Central, v. 19, n. 1, p. 1–32, 2019. Disponível em: <https://doi.org/10.1186/s12911-019-0874-0>.

ELSHAWI, R.; SHERIF, Y.; AL-MALLAH, M.; SAKR, S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. **Computational Intelligence**, Wiley Online Library, v. 37, n. 4, p. 1633–1650, 2021. Disponível em: <https://doi.org/10.1111/coin.12410>.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. C. P. L. F. **Inteligência artificial: uma abordagem de aprendizado de máquina**. LTC, 2021. 304 p. Disponível em: <https://books.google.com.br/books/about/Intelig\%C3\%AAncia_artificial.html?id=4DwelAEACAAJ&redir_esc=y>.

GANGULY, A.; DAS, R.; SETUA, S. Histopathological image and lymphoma image classification using customized deep learning models and different optimization algorithms. In: IEEE. **2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)**. 2020. p. 1–7. Disponível em: <https://doi.org/10.1109/ICCCNT49239.2020.9225616>.

HAYKIN, S. **Neural networks and learning machines, 3/E**. Pearson Education India, 2010. Disponível em: <https://books.google.com.br/books/about/Neural_Networks_and_Learning_Machines.html?hl=pt-BR&id=K7P36lKzI_QC&redir_esc=y>.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. Mask r-cnn. In: **IEEE international conference on computer vision**. [s.n.], 2017. p. 2961–2969. Disponível em: <https://doi.org/10.1109/ICCV.2017.322>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **IEEE conference on computer vision and pattern recognition**. [s.n.], 2016. p. 770–778. Disponível em: <https://doi.org/10.1109/CVPR.2016.90>.

IRFAN, M.; BASUKI, S.; AZHAR, Y. Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning. **Bulletin of Electrical Engineering and Informatics**, v. 10, n. 3, p. 1621–1633, 2021. Disponível em: <https://doi.org/10.11591/eei.v10i3.2344>.

JANOCHA, K.; CZARNECKI, W. M. On loss functions for deep neural networks in classification. **arXiv preprint arXiv:1702.05659**, 2017. Disponível em: <https://doi.org/10.4467/20838476SI.16.004.6185>.

KAUFMANN, E.; KALYANAKRISHNAN, S. Information complexity in bandit subset selection. In: PMLR. **Conference on Learning Theory**. 2013. p. 228–251. Disponível em: <https://www.researchgate.net/publication/287279572_Information_complexity_in_bandit_subset_selection>.

KAUSHAL, C.; BHAT, S.; KOUNDAL, D.; SINGLA, A. Recent trends in computer assisted diagnosis (cad) system for breast cancer diagnosis using histopathological images. **Irbm**, Elsevier, v. 40, n. 4, p. 211–227, 2019. Disponível em: <https://doi.org/10.1016/j.irbm.2019.06.001>.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In: **Advances in neural information processing systems**. [s.n.], 2017. p. 3146–3154. Disponível em: <https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint**. Disponível em: <https://arxiv.org/abs/1412.6980>.

KRISHNAN, M. M. R.; CHAKRABORTY, C.; PAUL, R. R.; RAY, A. K. Hybrid segmentation, characterization and classification of basal cell nuclei from histopathological images of normal oral mucosa and oral submucous fibrosis. **Expert Systems with Applications**, Elsevier, v. 39, n. 1, p. 1062–1077, 2012. Disponível em: <https://doi.org/10.1016/j.eswa.2011.07.107>.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: **IEEE conference on computer vision and pattern recognition**. [s.n.], 2015. p. 3431–3440. Disponível em: <https://doi.org/10.1109/CVPR.2015.7298965>.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, J.; BANSAL, N.; LEE, S.-I. From local explanations to global understanding with explainable ai for trees. **Nature machine intelligence**, Nature Publishing Group, v. 2, n. 1, p. 56–67, 2020. Disponível em: <https://doi.org/10.1038/s42256-019-0138-9>.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: **Proceedings of the 31st international conference on neural information processing systems**. [s.n.], 2017. p. 4768–4777. Disponível em: <https://dl.acm.org/doi/10.5555/3295222.3295230>.

LUNDBERG, S. M.; NAIR, B.; VAVILALA, M. S.; HORIBE, M.; EISSES, M. J.; ADAMS, T.; LISTON, D. E.; LOW, D. K.-W.; NEWMAN, S.-F.; KIM, J. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. **Nature biomedical engineering**, Nature Publishing Group, v. 2, n. 10, p. 749–760, 2018. Disponível em: <https://doi.org/10.1016/j.future.2020.04.038>.

MA, L.; XIAO, Z.; LI, K.; LI, S.; LI, J.; YI, X. Game theoretic interpretability for learning based preoperative gliomas grading. **Future Generation Computer Systems**, Elsevier, v. 112, p. 1–10, 2020. Disponível em: <https://doi.org/10.1016/j.future.2020.04.038>.

MARTINS, A. S.; NEVES, L. A.; FARIA, P. R. de; TOSTA, T. A.; LONGO, L. C.; SILVA, A. B.; ROBERTO, G. F.; NASCIMENTO, M. Z. do. A hermite polynomial algorithm

for detection of lesions in lymphoma images. **Pattern Analysis and Applications**, Springer, p. 1–13, 2020. Disponível em: <https://doi.org/10.1007/s10044-020-00927-z>.

MENG, T.; LIN, L.; SHYU, M.-L.; CHEN, S.-C. Histology image classification using supervised classification and multimodal fusion. In: **2010 IEEE International symposium on multimedia**. [s.n.], 2010. p. 145–152. Disponível em: <https://doi.org/10.1109/ISM.2010.29>.

MILLER JUNIOR, R. G. **Beyond ANOVA: basics of applied statistics**. CRC press, 1997. Disponível em: <https://doi.org/10.1201/b15236>.

MOLNAR, C. **Interpretable machine learning**. 2022. Disponível em: <https://christophm.github.io/interpretable-ml-book/pdp.html>.

NANNIA, L.; GHIDONI, S.; BRAHNAM, S. Ensemble of convolutional neural networks for bioimage classification. **Applied Computing and Informatics**, Emerald Publishing Limited, 2020. Disponível em: <https://doi.org/10.1016/j.aci.2018.06.002>.

NASCIMENTO, M.; NEVES, L.; DUARTE, S.; DUARTE, Y.; BATISTA, V. R. Classification of histological images based on the stationary wavelet transform. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. 2015. v. 574, n. 1, p. 012133. Disponível em: <https://doi.org/10.1088/1742-6596/574/1/012133>.

NASCIMENTO, M. Z. do; MARTINS, A. S.; TOSTA, T. A. A.; NEVES, L. A. Lymphoma images analysis using morphological and non-morphological descriptors for classification. **Computer methods and programs in biomedicine**, Elsevier, v. 163, p. 65–77, 2018. Disponível em: <https://doi.org/10.1016/j.cmpb.2018.05.035>.

NOWAK, A. S.; RADZIK, T. The shapley value for n-person games in generalized characteristic function form. **Games and Economic Behavior**, Elsevier, v. 6, n. 1, p. 150–161, 1994. Disponível em: <https://doi.org/10.1006/game.1994.1008>.

ORLOV, N. V.; CHEN, W. W.; ECKLEY, D. M.; MACURA, T. J.; SHAMIR, L.; JAFFE, E. S.; GOLDBERG, I. G. Automatic classification of lymphoma images with transform-based global features. **IEEE Transactions on Information Technology in Biomedicine**, v. 14, n. 4, p. 1003–1013, 2010. Disponível em: <https://doi.org/10.1109/TITB.2010.2050695>.

OZA, N. C.; TUMER, K. Classifier ensembles: Select real-world applications. **Information fusion**, Elsevier, v. 9, n. 1, p. 4–20, 2008. Disponível em: <https://doi.org/10.1016/j.inffus.2007.07.002>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <https://arxiv.org/abs/1201.0490>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: **22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [s.n.], 2016. p. 1135–1144. Disponível em: <https://doi.org/10.1145/2939672.2939778>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Anchors: High-precision model-agnostic explanations. In: **AAAI conference on artificial intelligence**. [s.n.], 2018. v. 32, n. 1. Disponível em: <https://doi.org/10.1609/aaai.v32i1.11491>.

RICE, J. A. **Mathematical statistics and data analysis**. Cengage Learning, 2006. Disponível em: <https://doi.org/10.2307/3619963>.

ROBERTO, G. F.; LUMINI, A.; NEVES, L. A.; NASCIMENTO, M. Z. do. Fractal neural network: A new ensemble of fractal geometry and convolutional neural networks for the classification of histology images. **Expert Systems with Applications**, Elsevier, v. 166, p. 114103, 2021. Disponível em: <https://doi.org/10.1016/j.eswa.2020.114103>.

ROBERTO, G. F.; NEVES, L. A.; NASCIMENTO, M. Z.; TOSTA, T. A.; LONGO, L. C.; MARTINS, A. S.; FARIA, P. R. Features based on the percolation theory for quantification of non-hodgkin lymphomas. **Computers in biology and medicine**, Elsevier, v. 91, p. 135–147, 2017. Disponível em: <https://doi.org/10.1016/j.compbiomed.2017.10.012>.

RODGERS, J. L.; NICEWANDER, W. A. Thirteen ways to look at the correlation coefficient. **The American Statistician**, Taylor & Francis, v. 42, n. 1, p. 59–66, 1988. Disponível em: <https://doi.org/10.1080/00031305.1988.10475524>.

RODRIGUEZ, J. D.; PEREZ, A.; LOZANO, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 3, p. 569–575, 2009. Disponível em: <https://doi.org/10.1109/TPAMI.2009.187>.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. P.Hall, 2002. Disponível em: <https://dl.acm.org/doi/book/10.5555/1671238>.

SAIN, S. R. **The nature of statistical learning theory**. Taylor & Francis, 1996. Disponível em: <https://doi.org/10.1007/978-1-4757-3264-1>.

SANTOS, F. P. S.; FERNANDES, G. S. **Linfomas não-Hodgkin**. [S.l.], 2008. Disponível em: <https://assinantes.medicinanet.com.br/conteudos/revisoes/99/linfomas_nao_hodgkin.htm>.

SHAMIR, L.; ORLOV, N.; ECKLEY, D. M.; MACURA, T. J.; GOLDBERG, I. G. Iicbu 2008: a proposed benchmark suite for biological image analysis. **Medical & biological engineering & computing**, Springer, v. 46, n. 9, p. 943–947, 2008. Disponível em: <https://doi.org/10.1007/s11517-008-0380-5>.

SHANKLAND, K. R.; ARMITAGE, J. O.; HANCOCK, B. W. Non-hodgkin lymphoma. **The Lancet**, Elsevier, v. 380, n. 9844, p. 848–857, 2012. Disponível em: <https://doi.org/10.1016/S0140-6736(12)60605-9>.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965. Disponível em: <https://doi.org/10.1093/biomet/52.3-4.591>.

SILVA, A. B.; MARTINS, A. S.; TOSTA, T. A. A.; NEVES, L. A.; SERVATO, J. P. S.; ARAÚJO, M. S. de; FARIA, P. R. de; NASCIMENTO, M. Z. do. Computational analysis of histological images from hematoxylin and eosin-stained oral epithelial dysplasia tissue

sections. **Expert Systems with Applications**, Elsevier, p. 116456, 2022. Disponível em: <https://doi.org/10.1016/j.eswa.2021.116456>.

SILVA, A. B. et al. Métodos computacionais para análise e classificação de displasias em imagens da cavidade bucal. Universidade Federal de Uberlândia, 2019. Disponível em: <http://dx.doi.org/10.14393/ufu.di.2019.2390>.

SILVA, A. B.; SANTOS, D. F. D.; TOSTA, T. A.; MARTINS, A. S.; NEVES, L. A.; TRAVENÇLO, B. A.; FARIA, P. R. de; NASCIMENTO, M. Z. do. Segmentation of oral epithelial dysplasias employing mask r-cnn and color normalization. In: IEEE. **2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. 2020. p. 2818–2824. Disponível em: <http://dx.doi.org/10.1109/BIBM49941.2020.9313101>.

SMITH, J.; RATTAY, T.; MCCONKEY, C.; HELLIWELL, T.; MEHANNA, H. Biomarkers in dysplasia of the oral cavity: a systematic review. **Oral oncology**, Elsevier, v. 45, n. 8, p. 647–653, 2009. Disponível em: <https://doi.org/10.1016/j.oraloncology.2009.02.006>.

SOCIETY, A. **Cancer Statistics Center**. 2020. Disponível em: <https://cancerstatisticscenter.cancer.org>.

SOMARATNE, U. V.; WONG, K. W.; PARRY, J.; SOHEL, F.; WANG, X.; LAGA, H. Improving follicular lymphoma identification using the class of interest for transfer learning. In: IEEE. **2019 Digital Image Computing: Techniques and Applications (DICTA)**. 2019. p. 1–7. Disponível em: <https://doi.org/10.1109/DICTA47822.2019.8946075>.

SONG, F.; GUO, Z.; MEI, D. Feature selection using principal component analysis. In: **International Conference on System Science, Engineering Design and Manufacturing Informatization**. [s.n.], 2010. v. 1, p. 27–30. Disponível em: <https://doi.org/10.1109/ICSEM.2010.14>.

SONG, Y.; CAI, W.; HUANG, H.; FENG, D.; WANG, Y.; CHEN, M. Bioimage classification with subcategory discriminant transform of high dimensional visual descriptors. **BMC bioinformatics**, Springer, v. 17, n. 1, p. 465, 2016. Disponível em: <https://doi.org/10.1186/s12859-016-1318-9>.

SONG, Y.; LI, Q.; HUANG, H.; FENG, D.; CHEN, M.; CAI, W. Low dimensional representation of fisher vectors for microscopy image classification. **IEEE transactions on medical imaging**, v. 36, n. 8, p. 1636–1649, 2017. Disponível em: <https://doi.org/10.1109/TMI.2017.2687466>.

TOSTA, T. A. A. et al. Método computacional para segmentação não supervisionada de imagens histológicas de linfoma. Universidade Federal de Uberlândia, 2016. Disponível em: <http://doi.org/10.14393/ufu.di.2016.17>.

TUKEY, J. W. et al. **Exploratory data analysis**. Reading, MA, 1977. v. 2. Disponível em: <https://doi.org/10.1007/978-0-387-32833-1_136>.

WONG, T.-T.; YEH, P.-Y. Reliable accuracy estimates from k-fold cross validation. **IEEE Transactions on Knowledge and Data Engineering**, v. 32, n. 8, p. 1586–1594, 2019. Disponível em: <https://doi.org/10.1109/TKDE.2019.2912815>.

WU, H.; RUAN, W.; WANG, J.; ZHENG, D.; LIU, B.; GENG, Y.; CHAI, X.; CHEN, J.; LI, K.; LI, S. et al. Interpretable machine learning for covid-19: an empirical study on severity prediction task. **IEEE Transactions on Artificial Intelligence**, IEEE, 2021. Disponível em: <https://doi.org/10.1109/TAI.2021.3092698>.

YOO, T. K.; RYU, I. H.; CHOI, H.; KIM, J. K.; LEE, I. S.; KIM, J. S.; LEE, G.; RIM, T. H. Explainable machine learning approach as a tool to understand factors used to select the refractive surgery technique on the expert level. **Translational vision science & technology**, The Association for Research in Vision and Ophthalmology, v. 9, n. 2, p. 8–8, 2020. Disponível em: <https://doi.org/10.1167/tvst.9.2.8>.