UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

PEDRO ALVES EUZEBIO
Orientadora: Profa. Dra. Andrea Gomes Campos Bianchi
Coorientador: Prof. Dr. Rafael Alves Bonfim de Queiroz

INTERPRETABILIDADE DE MODELOS DE APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO AUTOMÁTICA DE CÉLULAS CERVICAIS

Ouro Preto, MG
2022

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

PEDRO ALVES EUZEBIO

INTERPRETABILIDADE DE MODELOS DE APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO AUTOMÁTICA DE CÉLULAS CERVICAIS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Andrea Gomes Campos Bianchi
Coorientador: Prof. Dr. Rafael Alves Bonfim de Queiroz

Ouro Preto, MG
2022

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Pedro Alves Euzebio

Interpretabilidade de modelos de aprendizagem de máquina para classificação automática de células cervicais

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 27 de Outubro de 2022.

Membros da banca

Andrea Gomes Campos Bianchi (Orientadora) - Doutora - Universidade Federal de Ouro Preto
Rafael Alves Bonfim de Queiroz (Coorientador) - Doutor - Universidade Federal de Ouro Preto
Eduardo José da Silva Luz (Examinador) - Doutor - Universidade Federal de Ouro Preto
Daniela Costa Terra (Examinadora) - Mestre - Instituto Federal de Minas Gerais

Andrea Gomes Campos Bianchi, Orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/10/2022.

Documento assinado eletronicamente por **Andrea Gomes Campos Bianchi**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/11/2022, às 15:25, 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 http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0416024** e o código CRC **C6304E4C**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

À Universidade Federal de Ouro Preto (UFOP), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Resumo

A utilização de modelos de aprendizado de máquina e profundo para automatização de processos de tomada de decisões obtém sucesso em diferentes contextos. Devido à difusão desses algoritmos, torna-se importante que eles sejam interpretáveis para que se mostrem confiáveis ao serem adotados por humanos. Embora o aprendizado alcance resultados de ponta em aplicações do mundo real, seu número excessivo de parâmetros não é bem compreendido pelos humanos. Nesta monografia apresentamos uma investigação da interpretabilidade de métodos de aprendizado de máquina utilizando o método *Local interpretable model agnostic explanations (LIME)* para prover explicações. Os modelos sobre os quais as estratégias de interpretabilidade serão aplicados são de florestas aleatórias, usados no contexto de classificação de imagens de células cervicais. Os resultados para modelos de árvore de decisão foram explicações individuais das instâncias mais representativas da base de dados, seguido de uma análise de ocorrências das características em meio as explicações geradas, além do cálculo da *permutation importances* para fins de comparação.

Palavras-chave: Interpretabilidade. Aprendizagem de Máquina. Rede Neural Convolucional. Floresta Aleatória. Célula cervical. classificação

Abstract

Using machine and deep learning models for automated decision-making processes is successful in different contexts. Due to the dissemination of these algorithms, it becomes important that they are interpretable to be reliable when adopted by humans. Although learning achieves cutting-edge results in real-world applications, its excessive number of parameters is not well understood by humans. We present an interpretability investigation of machine learning methods using the *Local interpretable model agnostic explanations (LIME)*. The interpretability strategies will be applied at Random Forests, used in the context of cervical cell image classification.The results for decision tree models were individual explanations of the most representative instances of the database, followed by an analysis of occurrences of the characteristics among the generated explanations, in addition to the calculation of permutation importances for comparison purposes.

Keywords: Interpretability, Machine Learning, Convolutional Neural Network, Random Forest, Cervical Cell, Classification.

Lista de Ilustrações

Figura 3.1 – Recortes 90x90 para cada classe presente no CRIC Cervix: A) NILM; (B) ASC-US; (C) LSIL; (D) ASC-H; (E) HSIL; (F) SCC. Diniz et al. (2021a). . 14
Figura 3.2 – Arquitetura Proposta. 15
Figura 3.3 – Estrutura das etapas apresentadas na Figura 3.2. 15
Figura 3.4 – Explicação gerada pelo LIME Tabular (GARREAU; LUXBURG, 2020b) . 16
Figura 4.1 – Etapa 1: Explicação para uma instância da classe de células alteradas 19
Figura 4.2 – Etapa 1: Explicação para uma instância da classe NILM 19
Figura 4.3 – Etapa 1: Histograma de características presentes nas explicações do SP-LIME 20
Figura 4.4 – Etapa 1: Histograma de extratores de características presentes nas explicações do SP-LIME 20
Figura 4.5 – Etapa 1: 20 maiores valores de *Feature Importance* para o modelo 21
Figura 4.6 – Etapa 2: Explicação para uma instância da classe de células maduras. 21
Figura 4.7 – Etapa 2: Explicação para uma instância da classe de células jovens. 22
Figura 4.8 – Etapa 2: Histograma de características presentes nas explicações do SP-LIME. 22
Figura 4.9 – Etapa 2: Histograma de extratores de características presentes nas explicações do SP-LIME. 23
Figura 4.10–Etapa 2: 20 maiores valores de *Feature Importance* para o modelo. 23
Figura 4.11–Etapa 3: Explicação de uma instância da classe ASC-H. 24
Figura 4.12–Etapa 3: Explicação de uma instância da classe HSIL. 24
Figura 4.13–Etapa 3: Explicação de uma instância da classe SC. 25
Figura 4.14–Etapa 3: Histograma de características presentes nas explicações do SP-LIME. 25
Figura 4.15–Etapa 3: Histograma de extratores de características presentes nas explicações do SP-LIME. 26
Figura 4.16–Etapa 3: 20 maiores valores de *Feature Importance* para o modelo. 26
Figura 4.17–Etapa 4: Explicação para uma instância da classe ASC-US. 27
Figura 4.18–Etapa 4: Explicação para uma instância da classe LSIL. 27
Figura 4.19–Etapa 4: Histograma de características presentes nas explicações do SP-LIME. 28
Figura 4.20–Etapa 4: Histograma de extratores de características presentes nas explicações do SP-LIME. 28
Figura 4.21–Etapa 4: 20 maiores valores de *Feature Importance* para o modelo. 29

Lista de Tabelas

Tabela 3.1 – Número de células classificadas na base de dados CRIC *Cervix*. Diniz et al. (2021a) 14
Tabela 4.1 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células normais/alteradas (etapa 1). 30
Tabela 4.2 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células alteradas (etapa 2). 30
Tabela 4.3 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células jovens (etapa 3). 31
Tabela 4.4 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células maduras (etapa 4). 31

Lista de Abreviaturas e Siglas

LIME	*Local Interpretable Model-Agnostic Explanation*
SP-LIME	*Submodular Pick* - LIME
CNN	*Convolutional Neural Networks*
NN	*Neural Networks*
ReLU	Rectified linear unit
CRIC	*Center for recognition and inspection of cells*
XAI	*Explainable Artificial Intelligence*
LSIL	Lesão intraepitelial escamosa de baixo grau
NILM	Negativo para lesão intraepitelial ou malignidade
ASC-H	Células escamosas atípicas não podendo excluir HSIL
ASC-US	Células escamosas atípicas de significado indeterminado
HSIL	Lesão intraepitelial escamosa de alto grau
SCC	Carcinoma de células escamosas

Sumário

1 Introdução 1
1.1 Justificativa 2
1.2 Objetivos 3
1.3 Organização do Trabalho 3
2 Revisão Bibliográfica 4
2.1 Fundamentação Teórica 4
2.1.1 Câncer de Colo de Útero 4
2.1.2 Classificação de Imagens 4
2.1.2.1 Classificação Hierárquica 4
2.1.3 Floresta Aleatória 5
2.1.3.1 *Ensemble learning* 5
2.1.3.2 Importância das Características 5
2.1.4 Extração de Características 6
2.1.4.1 *Region Props* 6
2.1.4.2 Haralick 6
2.1.4.3 Zernike 7
2.1.4.4 Padrão Binário Local 8
2.1.4.5 *Threshold Adjancency Statistics* 8
2.1.5 Interpretabilidade 8
2.1.5.1 *Local Interpretable Model-Agnostic Explanations* (LIME) . . 9
2.2 Trabalhos Relacionados ao LIME 10
2.2.1 Discussão 12
3 Desenvolvimento 13
3.1 Base de Dados: CRIC *Cervix* 13
3.2 Métodos 14
3.2.1 Floresta Aleatória 14
3.2.1.1 Passo 1: Lime Tabular 16
3.2.1.2 Passo 2: SP-LIME 16
3.2.1.3 Passo 3: Contagem de ocorrências das características contidas nas explicações do SP-LIME 17
3.2.1.4 Passo 4: Agrupamento das ocorrências por algoritmo extrator 17
4 Resultados e Discussões 18
4.1 Floresta Aleatória 18
4.1.1 Etapa 1: Normais/Alteradas 19
4.1.2 Etapa 2: Alteradas 21
4.1.3 Etapa 3: Jovens 24

4.1.4 Etapa 4: Maduras 26
4.2 Discussões 29
5 Conclusões e Trabalhos Futuros 32

Referências 33

1 Introdução

A cada ano, mais de 500.000 mil mulheres são diagnosticadas com câncer de colo do útero e a doença gera mais de 300.000 mortes pelo mundo (COHEN et al., 2019), sendo a quarta maior causa de mortes relacionadas ao câncer em mulheres. Desse modo, o câncer cervical se mostra um problema de saúde pública extremamente relevante, em que a prevenção por meio de diagnóstico precoce revela-se eficaz.

O advento de exames de rastreamento, tais como o Papanicolaou, permitiram o diagnóstico precoce e a diminuição de mortes. Todavia, o processo de análise clínica se mostra árduo devido ao grande número de células presentes em cada amostra (WRIGHT, 2007). Portanto, uma alternativa de apoio ao profissional de saúde é o uso citologia digital, com o objetivo de tornar o processo de análise mais eficiente e menos custoso.

Em 1988, foi concebido o sistema Bethesda para classificação de lesões cervicais, que tinha como princípio o entendimento de uma constante atualização nas terminologias, para que possam refletir os entendimentos mais atualizados relacionados à patologia em questão. Desse modo, após algumas atualizações, a mais recente, em 2014, apresentam-se as seguintes terminologias para classificação de células escamosas: Células escamosas atípicas de significado indeterminado, possivelmente não neoplásicas (ASC-US); Células escamosas atípicas, não se pode excluir lesão de alto grau (ASC-H); Lesão intraepitelial escamosa de baixo grau (LSIL); Lesão intraepitelial escamosa de alto grau (HSIL); Lesão intraepitelial escamosa de alto grau não podendo excluir micro-invasão; Carcinoma de células escamosas (SCC) e Negativo para lesão intraepitelial ou malignidade (NILM), para células sem lesão (NAYAR; WILBUR, 2017).

No contexto da classificação, o algoritmo de floresta aleatória possui diversas aplicações bem sucedidas no âmbito da medicina em imagens digitais. Em classificação de células cervicais,Sun Shaobo Li e Lang (2017) fez uso do floresta aleatória aplicado a 20 atributos extraídos da base de dados Harlev (MARINAKIS; DOUNIAS; JANTZEN, 2009), em que as classes são dissonantes do sistema Bethesda e as imagens representam células únicas. As Convolutional Neural Networks (CNNs) também são amplamente utilizadas nesse cenário, Chen et al. (2022) fizeram uso das arquiteturas *Xception*, *MobileNet* e *MobileNetV2* para classificação de células cervicais usando a mesma base de dados.

Em consonância com o sistema Bethesda, a base de dados CRIC *cervix* é um acervo de imagens reais de esfregaços do colo do útero retratando uma variedade de lesões em células cervicais, em que os núcleos das células possuem seis classificações diferentes. Mantida pela plataforma CRIC, a base de dados conta com 400 imagens e 11534 células classificadas. Fazendo uso da base de dados supracitada, Diniz et al. (2021b) propuseram uma metodologia de classificação hierárquica fazendo uso do floresta aleatória em atributos extraídos de imagens dos

núcleos e Diniz et al. (2021a) apresentaram um método de agrupamento (*ensemble learning*) aplicado a modelos de aprendizado profundo responsáveis por classificar células cervicais.

Com a consolidação do uso de métodos de aprendizado de máquina caixa preta para tomada de decisões, surge a necessidade de que eles sejam interpretáveis, visto que a confiabilidade de um modelo é diretamente relacionado à possibilidade de interpretá-lo (ZHANG et al., 2021). No contexto clínico, especificamente na classificação de células cervicais, a utilização de técnicas de interpretabilidade visa prover ao profissional de saúde, além das células cervicais classificadas, explicações relativas aos critérios de decisão adotados pelos modelos propostos.

Para gerar explicações em modelos classificadores, métodos de interpretação foram concebidos. Neste trabalho, será usado o método *Local Interpretable Model-Agnostic Explanations* (LIME) (RIBEIRO; SINGH; GUESTRIN, 2016) como explicador e os objetos de estudo serão os trabalhos de Diniz et al. (2021b) e Diniz et al. (2021a).

1.1 Justificativa

Desde a sua introdução por Papanicolaou em 1927 como exame de rastreamento para câncer do colo do útero, o exame preventivo de Papanicolaou provou ser o teste de triagem mais viável em questão de custo-benefício na história da medicina. De acordo com Instituto Nacional do Câncer, a taxa de mortalidade por câncer do colo do útero caiu 70% nos últimos 70 anos. Porém, atualmente existe uma realidade de incidências de falsos negativos, que varia entre 5% a 10% até mesmo nos melhores laboratórios (ROCK et al., 2000).

A partir desse contexto, com o objetivo de tornar o processo de diagnóstico mais eficiente, diminuir o tempo de espera e a sobrecarga dos profissionais de saúde, foram propostos métodos de automatização desse processo que é realizado de forma manual. Dentre esses métodos, o uso de algoritmos de aprendizado de máquina para classificar células cervicais possuem papel importante no processo de apoio automatizado de diagnósticos. Autores como Isidoro. et al. (2020) usaram métodos tradicionais de aprendizado de máquina para realizar as classificações, enquanto Hussain et al. (2020) usaram CNNs para fazer a mesma tarefa.

No entanto, a confiabilidade de um método de aprendizado de máquina é um ponto crucial para a adesão ao modelo (RIBEIRO; SINGH; GUESTRIN, 2016). Em modelos caixa preta, as explicações de instâncias do problema provem uma visão de quais características o algoritmo leva em consideração para tomar uma decisão. Ademais, no contexto médico, é de grande valia a integração dos especialistas no processo de desenho das estratégias de interpretação, caso contrário, o modelo dificilmente será integrado na rotina clínica (VELLIDO, 2020). Desse modo, os algoritmos de classificação de células cervicais podem ser submetidos a estratégias de interpretabilidade, em especial o uso de explicadores como o método LIME proposto por Ribeiro, Singh e Guestrin (2016). Portanto, justifica-se a aplicação de estratégias de explicabilidade em modelos de classificação voltados para citologia, em especial ao contexto de células cervicais.

1.2 Objetivos

Este trabalho tem como objetivo aplicar estratégias de explicabilidade em modelos de classificação de citologia cervical. São objetivos específicos desse trabalho:

- Analisar as predições dos modelos estudados para a definição do método de interpretabilidade;
- Prover explicações para classificações celulares explorando diferentes visualizações usando o método LIME em modelos de aprendizagem de máquina, tal como: floresta aleatória;
- Gerar a permutação de importâncias para os modelos;

1.3 Organização do Trabalho

O texto da monografia está organizado como segue. O Capítulo 2 elucida conceitos importantes e traz trabalhos com temas correlatos. No Capítulo 3 é descrita a metodologia definida. No Capítulo 4 é feita a apresentação dos resultados, bem como as discussões relacionadas. Por último, o Capítulo 5 expõe as considerações finais sobre o trabalho e apresenta propostas para trabalhos futuros.

2 Revisão Bibliográfica

Nos últimos anos com a ascensão da temática de interpretabilidade, devido à necessidade latente de desvendar modelos caixa preta e conceber modelos interpretáveis. Dito isso, este capítulo apresentará trabalhos relacionados ao tema, bem como conceitos importantes para melhor entendimento do leitor.

2.1 Fundamentação Teórica

Nesta seção, serão apresentados conceitos relacionados ao tema em pauta.

2.1.1 Câncer de Colo de Útero

O câncer de colo de útero consiste em uma lesão intrauterina, podendo ser causada por algumas variações do Papalomavírus Humano, o HPV (COHEN et al., 2019). A partir desse cenário, é utilizado o exame de Papanicolaou, que consiste em um exame microscópico, descrito por Georgios N. Papanicolaou, que faz uso de células raspadas do colo do útero e tem por objetivo detectar tumores cancerosos ou pré-cancerosos do colo do útero (WAGGONER, 2003).

2.1.2 Classificação de Imagens

Nos dias atuais a classificação se faz presente em vários âmbitos com o objetivo de auxiliar na tomada de decisões. A necessidade de classificar um objeto surge sempre que este é colocado em um grupo específico conforme seus atributos. Tratando-se de imagens, existe uma enorme produção de conteúdos gráficos, que cria a necessidade de classificá-los para que a acessibilidade se torne mais rápida e fácil (DEY et al., 2014).

Uma das aplicações da classificação de imagens é no âmbito das imagens médicas, onde o desafio se encontra em obter informações analisando e provendo apoio ao diagnóstico de uma doença. Este desafio está intimamente relacionado ao uso de métodos na exploração de resultados de processamento de imagens, reconhecimento de padrões e técnicas de classificação, que posteriormente provém resultados que são validados em conhecimento médico especializado (MIRANDA; ARYUNI; IRWANSYAH, 2016).

2.1.2.1 Classificação Hierárquica

Para problemas de classificação que envolvem uma relação hierárquica entre as classes, utiliza-se uma abordagem que engloba a natureza de classes e promove uma classificação hierárquica. Desse modo, a classificação é feita por meio da divisão do problema em problemas

menores de classificação. Um bom exemplo de caso onde esse tipo de abordagem é recomendável é a taxonomia biológica, onde indivíduos são classificados em termos de espécie, gênero, família e ordem, e essas classes apresentam uma relação hierárquica entre elas (GORDON, 1987).

2.1.3 Floresta Aleatória

Floresta Aleatória ou *Random Forest* (BREIMAN, 2001) é um método de aprendizado de máquina que consiste na combinação de árvores de decisão podendo ser usada para classificação e regressão. No que tange à classificação, o resultado do algoritmo é a classe que possui predominância de escolha na maioria das árvores.

Segundo Cutler, Cutler e Stevens (2012), é possível elencar motivos pelos quais a floresta aleatória é uma solução atraente em diversos casos. Do ponto de vista computacional, consegue lidar com regressão e classificação, é relativamente rápido para fase de treinamento e a fase de predição, pode ser usado para problemas de alta dimensão e é facilmente implementado usando paralelismo.

2.1.3.1 *Ensemble learning*

Ensemble learning consiste na combinação de vários modelos individuais com o objetivo de obter um melhor desempenho em uma generalização. Em um conjunto de modelos de *deep learning* combinam as vantagens dos modelos em questão de modo a obter um modelo final que possui um desempenho de generalização melhor (GANAIE et al., 2021).

2.1.3.2 Importância das Características

Se caracteriza pela geração de *insights* sobre como as características preditivas afetam a variável de interesse, ou seja, como as características impactam a predição do modelo.

Juntamente com a concepção do algoritmo de floresta aleatória também são propostas duas medidas para classificação de características, a variável importância (VI) e Importância Gini (IG). No entanto, foi constatado que em classificadores categóricos, ambas as medidas se mostram tendenciosas (ALTMANN et al., 2010).

Em decorrência desse fato, foi desenvolvido o algoritmo de *Permutation Importances* (PIMP). O algoritmo PIMP permuta o vetor de resposta s vezes. Para cada permutação do vetor resposta, a relevância para todas as variáveis preditoras é avaliada. Isso leva a um vetor de s medidas de importância para cada variável, que é chamado de as importâncias nulas. O algoritmo PIMP ajusta uma distribuição de probabilidade para a população de importâncias nulas, que o usuário pode escolher dentre as seguintes: Gaussiano, lognormal ou gamma (ALTMANN et al., 2010).

2.1.4 Extração de Características

Extração de características é um método no qual se tenta desenvolver uma transformação do espaço de entrada para o subespaço de baixa dimensão que preserva a maioria das informações relevantes, transformando características individuais em outras mais significantes (KHALID; KHALIL; NASREEN, 2014).

Desse modo, existem diversos algoritmos que extraem diferentes tipos de características de uma base de dados. Para este trabalho serão extraídas informações relacionadas a morfologia da região de interesse e também as textura da região..

2.1.4.1 *Region Props*

Region Props concentra a extração de características ligadas a região, muito usada como *toolbox* no MATLAB e também presente na biblioteca *skimage* e conta com um total de 24 características (SHOUMY et al., 2016), sendo que para o trabalho de Diniz et al. (2021a) foram usadas:

- Circularidade;
- intensidade mínima;
- intensidade máxima;
- intensidade média;
- área;
- perímetro;
- número de Euler
- extensão;
- eixo maior;
- eixo menor;
- excentricidade;
- solidez.

2.1.4.2 Haralick

As características de textura Haralick consistem na matriz de co-ocorrência projetada a partir da imagem (CHOWDHARY; ACHARJYA, 2020). As matrizes de co-ocorrência são descartadas com o objetivo de expor lateralmente a dependência espacial do nível de cinza como

relações angulares, conselhos verticais e horizontais nas imagens. Ao combinar a matriz de co-ocorrência, várias texturas alteradas apresentadas podem ser formadas. Ao todo são extraídas 13 características, a saber:

- diferença de variância;
- diferença de entropia;
- soma dos quadrados;
- correlação;
- contraste;
- soma média;
- medidas de informação de correlação;
- variância da soma;
- segundo momento angular;
- entropia;
- soma da entropia;
- coeficiente máximo de correlação;
- momento da diferença inversa.

2.1.4.3 Zernike

Momentos de Zernike são usados em aplicações de reconhecimento de padrões como descritores invariantes da forma da imagem, sendo que seu princípio baseia-se em um conjunto de polinômios sobre o espaço de coordenadas polares dentro de um único círculo. A partir disso, eles provaram serem superiores em relação a outras funções de momento, como momentos geométricos, em termos de suas capacidades de representação de características e robustez na presença de erro e ruído de quantização de imagem. Sua propriedade de ortogonalidade ajuda a alcançar um valor próximo de zero quanto à redundância medida em um conjunto de funções de momento. Assim, momentos de ordens diferentes correspondem a características independentes da imagem (CHONG; RAVEENDRAN; MUKUNDAN, 2003).

2.1.4.4 Padrão Binário Local

O padrão binário local (*Local Binary Pattern*-LBP) é um extrator de características de texturas que, apesar de ser simples, mostra-se eficiente. O funcionamento do LBP consiste em rotular os pixels de uma imagem limitando a vizinhança de cada pixel, gerando um resultado binário. Dentre as propriedades do método, a que mais se destaca é a robustez a mudanças monotônicas em escalas de cinza decorrentes de variações de iluminação, por exemplo. Desse modo, o padrão binário local pode ser visto como uma abordagem unificadora para os modelos estatísticos e estruturais tradicionalmente divergentes de análise de textura (PIETIKÄINEN, 2005).

2.1.4.5 *Threshold Adjancency Statistics*

O *Threshold Adjancency Statistics*(TAS) surgiu com o objetivo de ser um meio computacionalmente rápido e simples de diferenciar localização sub-celulares de uma imagem. Para isso, o algoritmo consiste na aplicação da operação de *Thresholding* para criar uma imagem binária a partir da imagem original. A partir disso, a intensidade média da imagem μ é calculada usando os pixels com intensidade de pelo menos 30. Sendo assim, a imagem é limitada ao intervalo de μ-30 a μ +30. Tal intervalo foi escolhido para evidenciar a diferença visual entre as imagens onde a operação foi aplicada e imagens visualmente semelhantes mas com localizações diferentes. Portanto, foram geradas um total de 27 estatísticas seguindo esse princípio (HAMILTON et al., 2007).

2.1.5 Interpretabilidade

A partir da popularização de algoritmos caixa preta e a adoção deles para tomar decisões antes confiadas a humanos, surgiram desconfianças sobre a veracidade de seus resultados. Sendo assim, tornou-se necessário que esses mecanismos se expliquem. Um exemplo disso, foi o trabalho de Angwin et al. (2016), que analisou o perfil de gerenciamento de infratores correcionais de sanções alternativas (COMPAS), um sistema criminal amplamente usado para avaliação de riscos e descobriu que suas previsões eram não confiáveis e racialmente tendenciosas (GILPIN et al., 2018).

Não existe uma definição matemática de interpretabilidade. Uma definição (não matemática) dada por Miller (2019) gira em torno da noção de que a interpretabilidade é relacionada ao grau em que um ser humano pode entender a motivação de uma decisão. No contexto de sistemas de aprendizado de máquina, Kim, Khanna e Koyejo (2016) descrevem interpretabilidade baseada no grau em que um ser humano pode prever consistentemente o resultado do modelo. Isso significa que a interpretabilidade de um modelo é maior se for mais fácil para uma pessoa raciocinar e rastrear por que uma previsão foi feita pelo modelo. Comparativamente, um modelo é mais interpretável do que outro modelo se as decisões do anterior são mais fáceis de entender do que as decisões do último (CARVALHO; PEREIRA; CARDOSO, 2019). Segundo Zhang et al.

(2021), interpretabilidade pode ser definida pela capacidade de fornecer explicações em termos compreensíveis a humanos. Explicações, idealmente, devem ser ou podem ser transformadas em regras de decisão lógica. Já os termos compreensíveis devem ser do domínio de conhecimento da aplicação. Tendo essa definição, é possível desbravar novas perspectivas em interpretabilidade.

A partir do conceito de interpretabilidade foi cunhado o termo sistema de IA explicável (XAI), que tem por objetivo tornar comportamento desses algoritmos mais inteligível para os humanos, dando explicações. O sistema XAI deve ser capaz de explicar suas capacidades e entendimentos; explicar o que fez, o que está fazendo agora e o que acontecerá a seguir (GUNNING et al., 2019).

Em uma definição simples, pode-se dizer que a interpretabilidade pode tomar dois caminhos: a criação de modelos interpretáveis em sua concepção e a criação de métodos de explicação voltados para modelos caixa preta (CARVALHO; PEREIRA; CARDOSO, 2019).

Para o contexto de explicar modelos caixa preta, é válido pontuar quais características constituem a base para um algoritmo explicador que faz esse papel. Como característica fundamental o explicador deve gerar explicações interpretáveis, ou seja, o conteúdo deve ser compreensível por humanos. Além disso, as explicações devem ter fidelidade local. Com relação a variedade de modelos suportados pelo explicador, ele deve ser capaz de explicar qualquer modelo, e portanto, ser agnóstico em relação ao modelo. Por fim, é essencial que o explicador forneça uma perspectiva de visão global do modelo analisado (RIBEIRO; SINGH; GUESTRIN, 2016).

2.1.5.1 *Local Interpretable Model-Agnostic Explanations* (LIME)

Tendo em vista a concepção de um explicador agnóstico, algumas técnicas visam desempenhar a tarefa de fornecer informações quantitativas e qualitativas de como modelos caixa preta tomam decisões. A partir disso, é notória a percepção da complexidade global de algoritmos caixa preta. Sendo assim, muitas vezes a compreensão nesses modelos é obtida através de um ponto de vista local, gerado pela interpretação de uma instância específica presente na base de dados (GARREAU; LUXBURG, 2020a).

Desse modo, Ribeiro, Singh e Guestrin (2016) desenvolveram o LIME, um método que tem a capacidade de explicar predições de qualquer classificador ou regressor de forma confiável, através de prover um aprendizado de um modelo localmente em torno da predição. Além disso, foi desenvolvida uma extensão do LIME, denominada Submodular Pick LIME (SP-LIME), que consiste em um algoritmo que seleciona um grupo de instâncias representativas com suas respectivas explicações para lidar com o problema de confiabilidade no modelo.

O SP-LIME por sua vez possui como objetivo obter uma compreensão mais ampla do modelo ao explicar um conjunto selecionado de instâncias individuais, visto que o usuário pode não ter disponibilidade de examinar explicações para todas as instâncias da base de dados. A

escolha das explicações é feita de modo a cobrir o maior número de atributos importantes, evitando redundâncias que podem ocorrer ao escolher instâncias com explicações similares.

2.2 Trabalhos Relacionados ao LIME

Nesta seção, serão relatados alguns trabalhos relacionados à interpretabilidade de modelos. Sendo assim, para maior entendimento do leitor, a seção contará com trabalhos que empregam interpretabilidade em diferentes contextos, abarcando a aplicabilidade da ferramenta LIME usando seus diferentes módulos. Desse modo, alguns artigos de referência serão relacionados ao contexto das redes neurais convolucionais, enquanto outros tratarão de métodos tradicionais.

Para entender melhor como as técnicas de interpretabilidade podem ser agrupadas, Zhang et al. (2021) conceberam uma taxonomia organizada em três dimensões: tipo de engajamento (abordagens de interpretação passiva versus ativa), o tipo de explicação e o foco (do local ao global). Esta taxonomia fornece uma visão 3D significativa de distribuição de artigos da literatura relevantes, conseguindo classificar os explicadores a partir destes prismas. Por fim, foi feito um resumo do estado da arte no que tange à interpretabilidade e métodos existentes de avaliação e sugeridos possíveis direções de pesquisa inspirado pela taxonomia criada.

No contexto clínico e de assistência médica, o LIME pode desempenhar um papel de apoio aos profissionais de saúde na tomada de decisões médicas. A partir disso, Kumarakulasinghe et al. (2020) avaliaram como o LIME realiza explicações a partir de um modelo em comparação com explicações feitas por profissionais da saúde de forma independente. Além disso, também foi investigada a relevância clínica e a confiabilidade das explicações obtidas por meio do LIME. Como resultados, houve o indicativo da relevância das explicações, bem como a consonância em relação as explicações feitas por médicos.

Magesh, Myloth e Tom (2020) propuseram uma abordagem para classificação de imagens digitalizadas das regiões do putâmen e núcleo caudado do cérebro, com o objetivo de auxiliar no diagnóstico para doença de Parkinson. Para a fase de treinamento foi usada uma CNN, mais especificamente a VGG16, implementada usando Keras (CHOLLET et al., 2015) e com o modelo constituído pelas camadas convolucionais, max-pooling, de ativação e camadas totalmente conectadas. Com essa configuração, o modelo obteve 95% de precisão e a partir da consolidação do modelo foi possível criar explicações de instâncias. Para isso, foi usado o LIME. Como resultado disso, foi percebido que o modelo nas instâncias analisadas enfatizavam regiões anormais do putâmen e caudado de um paciente sem a doença de Parkinson, denotando que essas regiões influenciaram a classificação dos dados.

Cervantes e Chan (2021) analisaram a confiabilidade de modelos baseados em CNNs com os objetivo de classificar imagens de raio-X do tórax saudáveis, com COVID e com pneumonia viral. Para realizar esse procedimento foram utilizadas quatro CNNs ImageNet: VGG16, DenseNet201, ResNet50, e EfficientNetB3. A partir disso, a DenseNet201 e a VGG16 adquiriram

as acurácias mais altas e o LIME foi aplicado em todos os casos. Dentre as CNNs com maior desempenho, a partir da explicação obtida da VGG16, percebeu-se que a rede não aprendeu características importantes para o diagnóstico da doença, apesar da acurácia elevada.

Um ponto de partida para a interpretabilidade em modelos de aprendizagem é definir o que seria um modelo interpretável (IML). Segundo Agarwal e Das (2020), um modelo interpretável pode ser definido como a medida em que um humano pode compreender as decisões tomadas por modelos de aprendizado de máquina em seu processo de tomada de decisão. Além disso, a importância de ser ter um modelo interpretável existe devido às seguintes características: (1) Viés - as previsões da garantia IML de modelos de ML não são tendenciosos e não discriminam grupos, (2) Privacidade - a confidencialidade dos dados é garantida por meio da interpretação do modelo, (3) Robustez - garante que as previsões do modelo de ML permaneceram consistentes e não têm mudanças quando poucas mudanças são feitas nos dados, (4) Causalidade - garante que apenas relações causais sejam escolhidas acima, (5) Crença - os indivíduos podem confiar no interpretável sistema facilmente comparado ao sistema de caixa preta.

Para avaliar especificamente métodos tradicionais de aprendizado de máquina em termos de desempenho e interpretabilidade no contexto de toxicologia, Robinson et al. (2017) propuseram uma análise comparativa entre os métodos de máquina de vetores de suporte, quadrado mínimo parcial e floresta aleatória. Para isso, foram aplicados os algoritmos em bases de dados de domínio público relacionadas à classificação binária e regressão no contexto da toxicologia e a avaliação de interpretabilidade partiu da utilização de esquemas de pontuação para avaliar imagens de mapas de calor de contribuições subestruturais dos algoritmos. Como resultados foi concluído que a floresta aleatória normalmente produz resultados comparáveis ou possivelmente de melhor desempenho preditivo do que as abordagens de modelagem linear e suas previsões podem ser interpretadas de uma forma química e biologicamente significativas.

Cruz, Schneider e Schapranow (2019) examinaram a ocorrência de insuficiência renal aguda em pacientes que fizeram cirurgias cardíacas e propuseram um modelo de predição para pacientes pós cirurgia. Para realizar esse procedimentos, foram aplicadas árvores de decisão convencionais e as árvores de decisão com *gradient boosting* (BAHAD; SAXENA, 2020) em um recorte da base de dados clínica MIMIC-III. Após a criação dos modelos, foi utilizado o LIME para prover explicações no modelo com melhor desempenho, o baseado em árvore de decisão com *gradient boosting*. A partir dos insumos gerados pelo LIME, foram reveladas informações sobre quais recursos são mais relevantes para a tomada de decisão do modelo, bem como foram elucidadas algumas inconsistências entre características tratadas como importantes pelo algoritmo e o consenso médico.

Davagdorj, Li e Ryu (2021) propuseram um framework para modelos de predição de hipertensão arterial, em que o problema é abordado com o uso de interpretabilidade de modelos caixa preta, através do método LIME. Nos experimentos, foram usados classificadores com o objetivo de determinar o melhor modelo para o contexto de hipertensão em relação a uma base

de dados da população coreana. Como resultado, o XGBoost (CHEN; GUESTRIN, 2016) obteve melhor acurácia. Com base nisso, o LIME foi utilizado para prover explicações personalizadas, servindo como auxílio para o profissional de saúde em diagnosticar o paciente.

2.2.1 Discussão

Os trabalhos de Davagdorj, Li e Ryu (2021) e Cruz, Schneider e Schapranow (2019) fizeram uso do módulo tabular presente do LIME. Tal módulo provê uma visualização gráfica para as explicações do modelo e em ambos os trabalhos, a análise dos insumos que o explicador gerou se ativeram a essa visualização. Logo, não foram exploradas outras formas de visualização que podem originar novos *insights* ao relacionar, por exemplo, várias explicações individuais e suas regras.

No que tange ao uso do LIME em imagens usando CNNs, os trabalhos de Cervantes e Chan (2021), Magesh, Myloth e Tom (2020),Kumarakulasinghe et al. (2020) fizeram uso do módulo Lime Image, que por usa vez, não gera explicações das instâncias mais representativas da base de dados. Desse modo, as análises se ativeram a explicação de instâncias individuais, por uma limitação da ferramenta. Apesar desse fato, nos trabalhos supracitados, a integração de profissionais de saúde como maneira de avaliar as explicações geradas pelo algoritmo se mostraram consistentes, visto que as explicações foram interpretadas com sucesso por humanos.

3 Desenvolvimento

Neste capítulo são descritas as etapas do desenvolvimento deste projeto. Na Seção 3.1 é descrita a base de dados e como ela se apresenta no artigo de referência, na Seção 3.2 são apresentados os métodos, abarcando detalhadamente a especificação de cada passo em termos teóricos e de implementação.

3.1 Base de Dados: CRIC *Cervix*

A plataforma de banco de imagens CRIC (disponível em: https://database.cric.com.br), que tem como uma de suas bases de dados o CRIC *Cervix*, é um consórcio colaborativo entre pesquisadores que visa fornecer coleções de células para a comunidade científica. As imagens do CRIC Cervix possuem características em comum as obtidas nos exames de Papanicolaou, contando com muitas células por imagem e uma resolução de 150 dpi. A base de dados disponibilizada conta com 400 imagens e 11.534 células classificadas (REZENDE et al., 2021).

No que tange à classificação das células presentes na base de dados, o sistema de nomenclatura Bethesda (NAYAR; WILBUR, 2017) é utilizado. Sendo assim, as células presentes no CRIC *Cervix* são classificadas em seis classes, exemplificadas na Figura 3.1 e na Tabela 3.1, sendo elas: negativo para lesão intraepitelial ou malignidade (NILM); células escamosas atípicas de significado indeterminado, possivelmente não neoplásicas (ASC-US); lesão intraepitelial escamosa de baixo grau (LSIL); células escamosas atípicas, não pode excluir lesão de alto grau (ASC-H); lesão intraepitelial escamosa de alto grau (HSIL); e carcinoma de células escamosas (SCC). A distribuição de amostras das classes apresentadas na base de dados pode ser observada pela Tabela 3.1.

Para o trabalho de Diniz et al. (2021b) a base de dados CRIC *Cervix* foi submetida a diferentes extratores de atributos. O algoritmo *Region Props* foi responsável por extrair atributos relacionados a morfologia dos núcleos das células. Já o LBP e o *Gray Level Co-occurrence Matrices* foram aplicados para gerar características de textura das células. O TAS foi usado para diferenciar imagens de localização subcelular distinta rapidamente e com alta precisão. Por último, o *Zernike Moments* foi utilizado por medir como a massa do núcleo é distribuída em uma certa região.

Tabela 3.1 – Número de células classificadas na base de dados CRIC *Cervix*. Diniz et al. (2021a)

Classificação da Célula	Quantidade de células classificadas - CRIC Cervix
NILM	6.770
ASC-US	606
ASC-H	925
LSIL	1.360
HSIL	1.703
SCC	161
Total	11.534

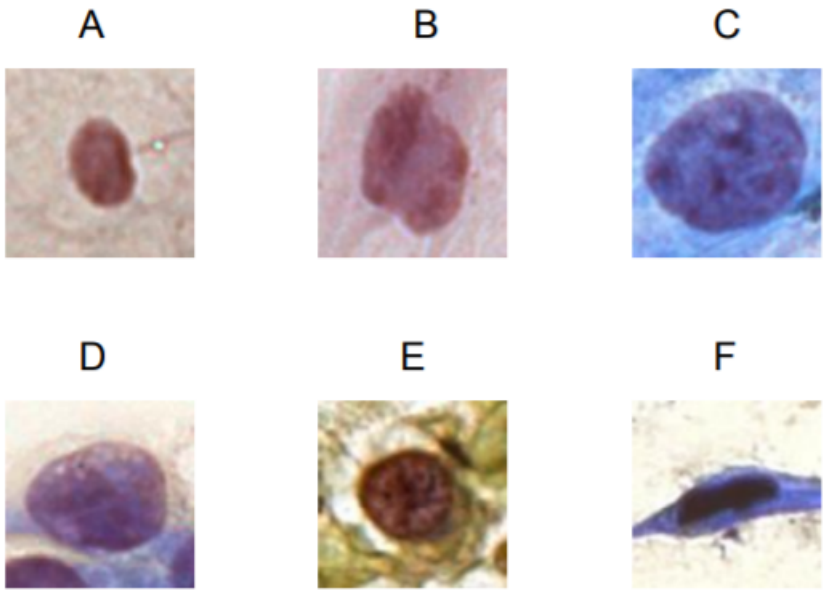

Figura 3.1 – Recortes 90x90 para cada classe presente no CRIC Cervix: A) NILM; (B) ASC-US; (C) LSIL; (D) ASC-H; (E) HSIL; (F) SCC. Diniz et al. (2021a).

3.2 Métodos

3.2.1 Floresta Aleatória

A base de dados usada será a base *CRIC Cervix* com características extraídas a partir dos métodos *Regionpros, TAS, Haralick, Zernike* e LBP. O uso do LIME a cada etapa será feito através da instanciação de um explicador do módulo *Lime Tabular*, as explicações geradas passarão pelo SP-LIME (https://github.com/marcotcr/lime).

Conforme dito anteriormente, um dos trabalhos onde as estratégias de interpretabilidade serão utilizadas tem como metodologia para executar a classificação hierárquica obtida com florestas aleatórias. Desse modo, é possível extrair vários modelos, sendo que cada um deles faz a classificação de um nível da hierarquia, como pode ser notado na Figura 3.2 no total são 3 hierarquias de explicações. Na primeira são separadas as células normais das alteradas. Na

segunda as alteradas são avaliadas conforme o grau das lesões (alto e baixo grau). Já na terceira, são discriminadas as classes das lesões. Desse modo, em cada classificador presente nos níveis da hierarquia haverá uma análise de interpretabilidade, denotada pelo nome de etapa, totalizando 4 etapas.

Além disso, cada etapa conta com uma sequência de passos que constroem os resultados. Como é especificado na Figura 3.3, cada etapa se divide em 4 passos, que se repetem para todas as etapas e serão descritos a seguir.

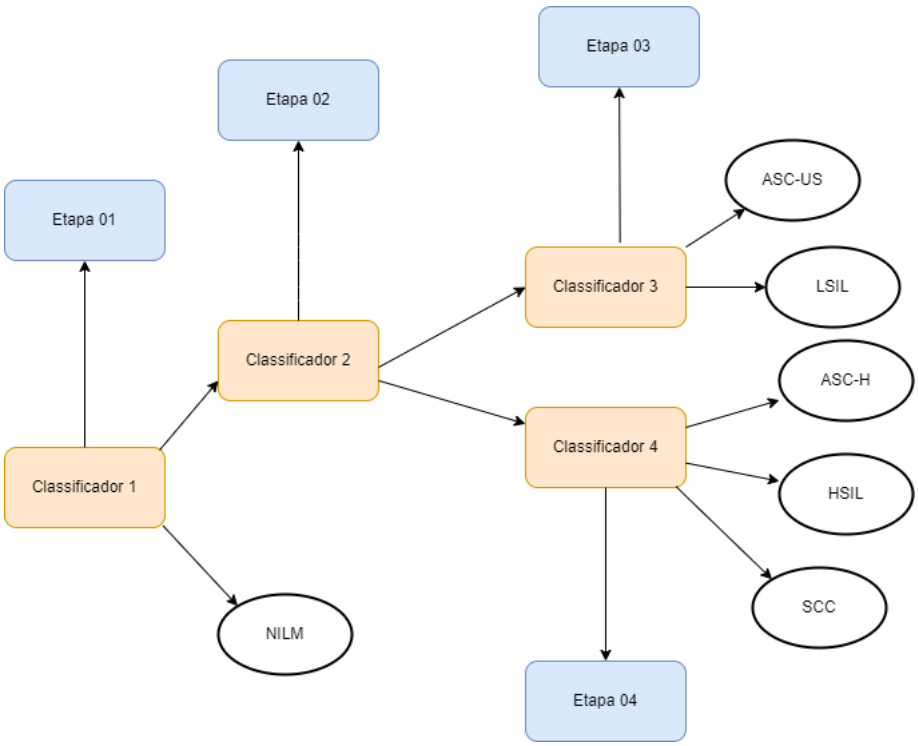

Figura 3.2 – Arquitetura Proposta.

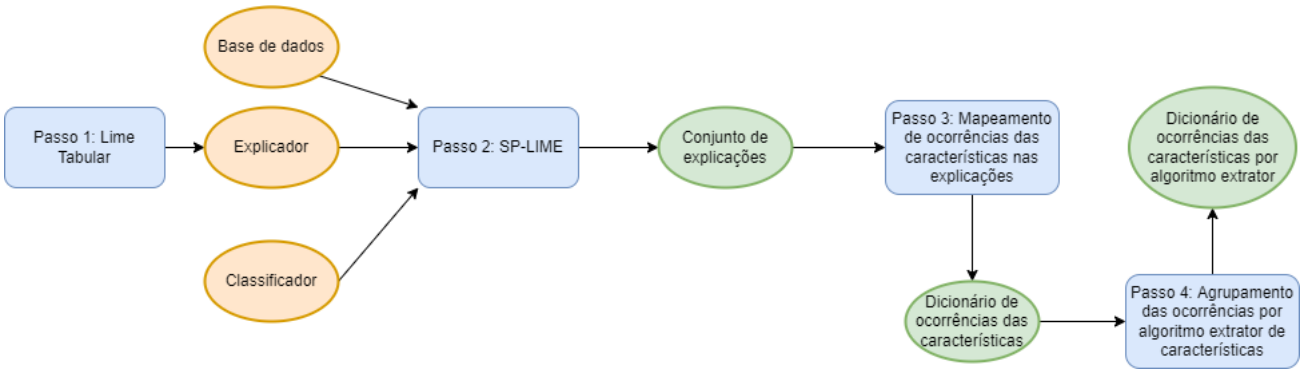

Figura 3.3 – Estrutura das etapas apresentadas na Figura 3.2.

3.2.1.1 Passo 1: Lime Tabular

Conforme dito na Seção 2.1.5.1, o LIME Tabular é um submódulo do LIME que lida com dados tabulares. Sendo assim, para lidar com o trabalho de Diniz et al. (2021b), será criado um explicador para cada etapa apresentada na Seção 3.2. Desse modo, como primeiro passo é necessário fazer uso da função principal que o LIME oferece, chamada de *explainer*, que nos permite chamar uma observação específica e obter uma interpretação como resultado. Ao instanciar o *explainer* é necessário definir todos os parâmetros manualmente. Para começar, para essa aplicação, será inserida a base de dados, além da definição do *mode* do explicador para *classification*, a lista de atributos e a lista de classes. A partir disso, é obtido o explicador que poderá ser usado para prover explicações, como é indicado na Figura 3.3.

3.2.1.2 Passo 2: SP-LIME

A partir da obtenção do explicador, é possível usar o *SP-LIME* através da chamada da função principal do módulo *submodular_pick*. Como parâmetros, serão usados o explicador, a base de dados, o método *predict_proba*, que retorna a probabilidade da predição para cada classe do modelo, o *num_features*, que representa o número de atributos que serão incluídos para a definição das instâncias mais representativas, além do parâmetro *method* ser definido como *full*, denotando que todas as instâncias da base de dados serão explicadas e o *num_exps_desired*, que representa a quantidade de explicações que serão geradas.

Como pode ser visto na Figura 3.4, ressaltando como é disposto o gráfico gerado pelo LIME, onde o eixo x representa os graus de importância das regras geradas através das características, que se encontram no eixo y. Além disso, é válido ressaltar que as barras de cor verde representam regras que contribuíram à favor da classe a qual o classificador fez a predição, enquanto as barras vermelhas denotam regras que contribuíram contra.

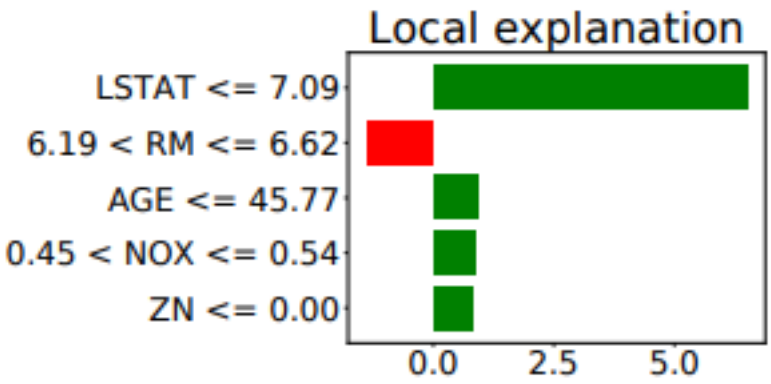

Figura 3.4 – Explicação gerada pelo LIME Tabular (GARREAU; LUXBURG, 2020b)

3.2.1.3 Passo 3: Contagem de ocorrências das características contidas nas explicações do SP-LIME

Gerados os conjuntos de explicações, o método *as_map* do LIME foi usado para obter uma lista de tuplas para cada uma das explicações contendo o *id* da característica como chave e seu peso como valor. Como foram geradas explicações do SP-LIME retornando apenas a quantidade de características mais importantes de acordo com o valor do atributo *num_features*, foi possível para todas as explicações calcular a ocorrência das características. Desse modo, o total de ocorrências para todas as explicações é dado por : $num\_features \times num\_exps\_desired$.

Portanto, como resultado desse passo temos um dicionário onde as chaves são dadas pelo *id* das características que apareceram nas explicações, enquanto os valores são representados pela quantidade de ocorrências da característica.

3.2.1.4 Passo 4: Agrupamento das ocorrências por algoritmo extrator

Para o último passo, tendo em mãos o algoritmo que gerou a base de dados usada no trabalho de Diniz et al. (2021b), foi possível identificar qual grupo de *id* das características fazia parte de cada algoritmo extrator, visto que as características de um mesmo extrator possuem *id* adjacentes na base de dados e o algoritmo que gerou a base de dados indica explicitamente o intervalo de *id* que delimita o conjunto de cada tipo de característica. Desse modo, o dicionário gerado no passo anterior foi comparado com esse insumo, gerando um novo dicionário, onde as chaves são os nomes do algoritmos extratores e os valores correspondem a contagens de ocorrências para cada um deles.

4 Resultados e Discussões

Neste capítulo são apresentados e analisados os resultados da metodologia de explicações de classificadores aplicados na base de dados *CRIC Cervix*. Para a realização dos experimentos, foi alocada uma máquina virtual da ferramenta *Google Colaboratory*, que conta com uma CPU Intel(R) Xeon(R) CPU com frequência de 2.30GHz, uma GPU: Tesla K80 com 12GB de memória GDDR5 VRAM e 12.6 GB de memória RAM.

4.1 Floresta Aleatória

Para o trabalho envolvendo a floresta aleatória, como dito na seção anterior, foi instanciado um explicador pelo *LIME Tabular*. Como parâmetros foram usados a base de dados *CRIC Cervix* balanceada, todas as *features* contidas na base, as classes representadas pelos números de 1 a 6, além do *mode* definido para *classification*. Com o explicador criado, foram geradas as explicações por meio do *SP-LIME*. Para isso, cada classificador proveu suas respectivas explicações usando os recortes das bases de dados utilizados na classificação, além disso foram definidos o *sample_size* como 20, *num_features* igual a 40 e o *num_exps_desired* estabelecido foi 10.

A partir dessas explicações geradas pelo SP-LIME, o passo 3 descrito na Seção 3.2.1.3 gerou um dicionário com a ocorrência de características, que por sua vez foi convertido em histogramas contendo as 20 características mais presentes no conjunto de explicações, como pode ser observado pela figura 4.3. Já a partir do passo 4, elucidado pela Seção 3.2.1.4 , o dicionário gerado contendo o agrupamento dessas características por algoritmo usado para extrai-las (*Region Props, Haralick, TAS, LBP, Zernike*) , também foi convertido em histograma, como pode ser visto na figura 4.4 .Além disso, o método *permutation_importances* da biblioteca *sklearn* foi usado para calcular a importância de cada uma das características analisadas pelo modelo, sendo representada por um gráfico com 20 características com maior pontuação exemplificado na figura 4.5.

Portanto, para maior entendimento, esta seção será dividida em etapas, onde cada uma das etapas contará com os histogramas e gráficos de *permutation_importances* supracitados. A etapa 1, tratará do explicador instanciado para a classificação de células entre as classes alteradas e normais. A etapa 2 contém explicações relacionadas a classificação de classes alteradas, contendo as classes jovens e maduras. Já a etapa 3 trata-se de explicações para as classes de células jovens (ASC-H, HSIL, SC). Por fim, a etapa 4 engloba as explicações de células maduras (ASC-US e LSIL).

4.1.1 Etapa 1: Normais/Alteradas

Como primeiro classificador da hierarquia, temos a classificação entre células normais e células alteradas. A partir disso, a seguir segue uma explicação das geradas pelo SP-LIME para cada uma das classes. As classes presentes são a NILM, que representa as células normais, e a classe de células alteradas.

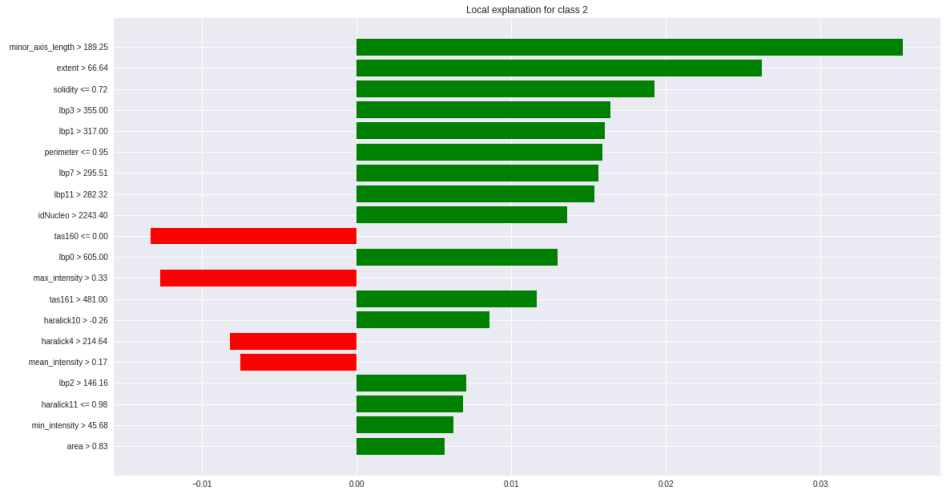

Figura 4.1 – Etapa 1: Explicação para uma instância da classe de células alteradas

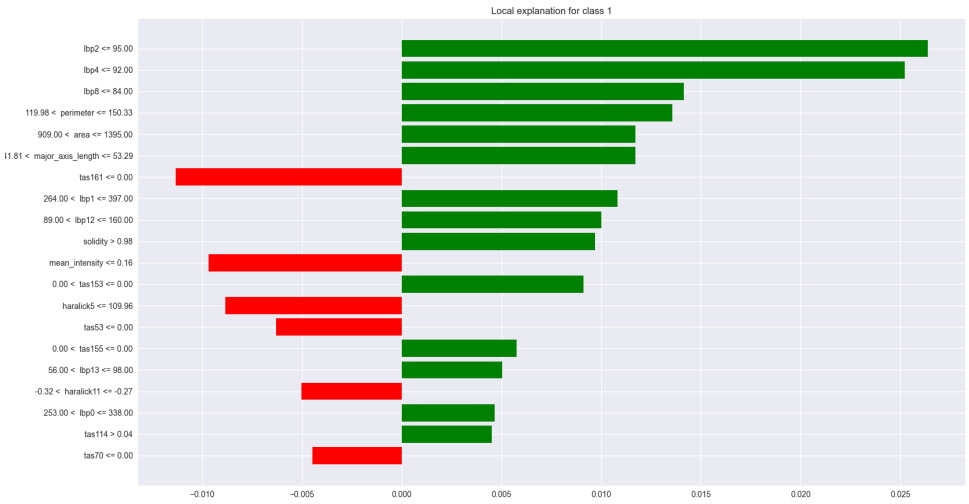

Figura 4.2 – Etapa 1: Explicação para uma instância da classe NILM

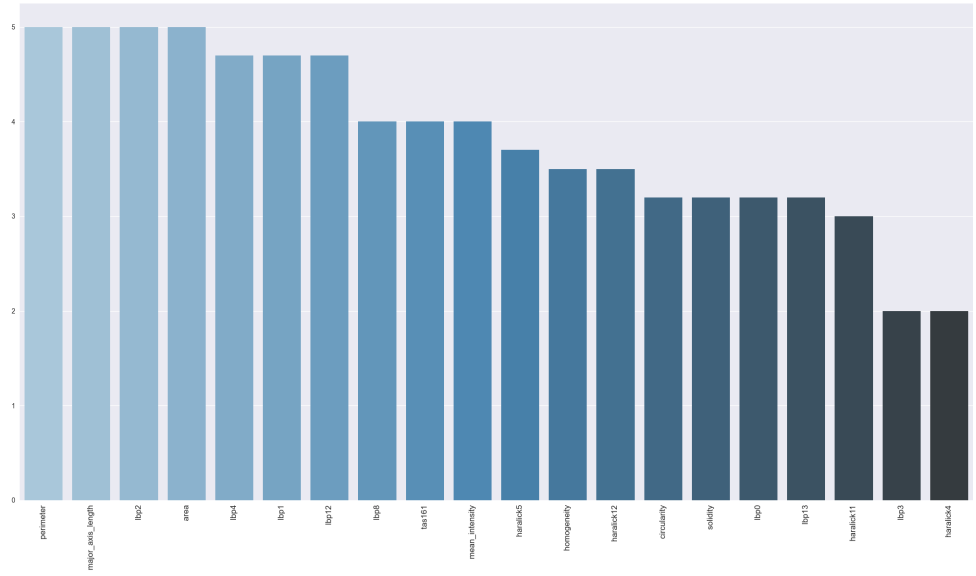

Figura 4.3 – Etapa 1: Histograma de características presentes nas explicações do SP-LIME

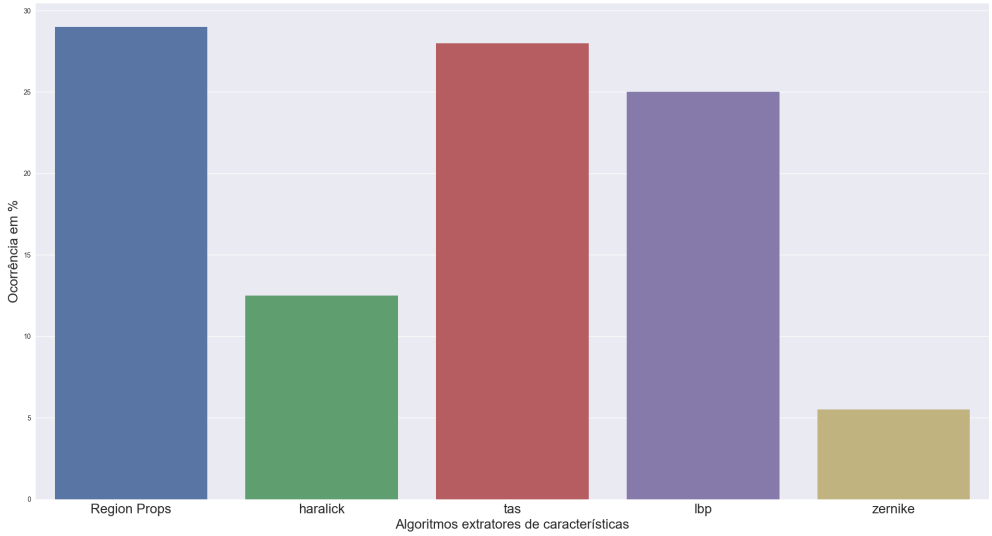

Figura 4.4 – Etapa 1: Histograma de extratores de características presentes nas explicações do SP-LIME

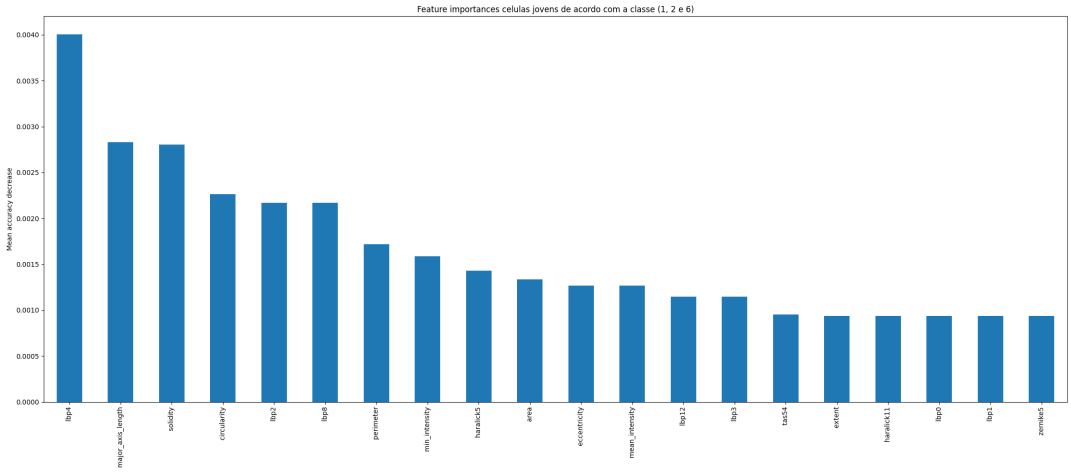

Figura 4.5 – Etapa 1: 20 maiores valores de *Feature Importance* para o modelo

4.1.2 Etapa 2: Alteradas

No caso das células classificadas como alteradas, é usado um outro classificador para predizer qual das classes que representam células alteradas é relacionado a instância. As classes são jovens e maduras.

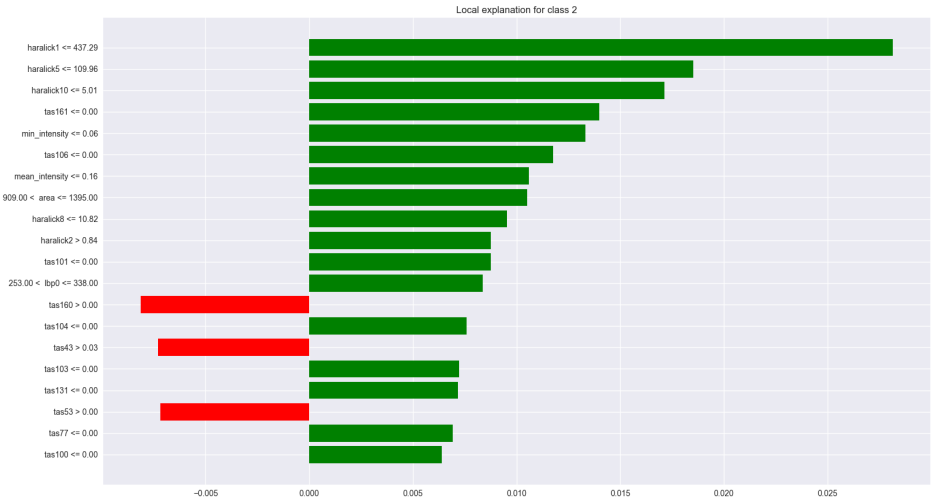

Figura 4.6 – Etapa 2: Explicação para uma instância da classe de células maduras.

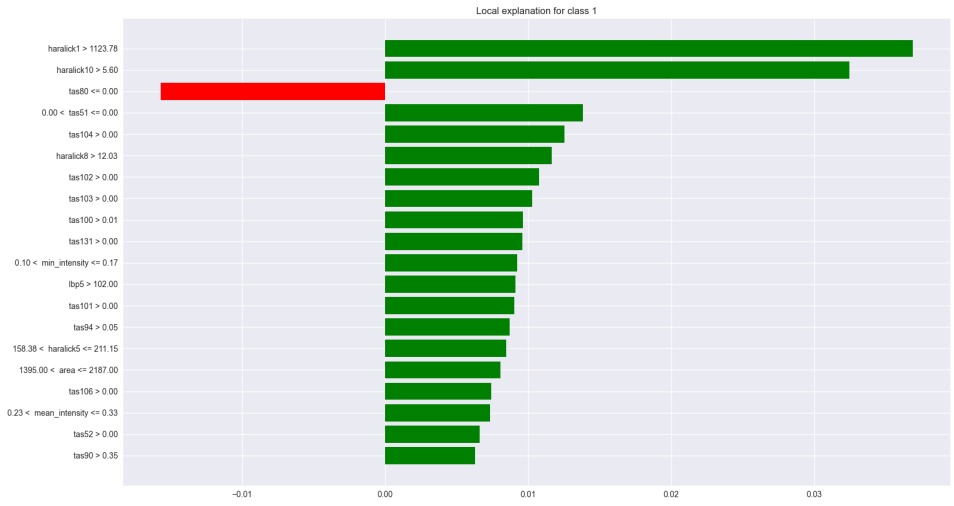

Figura 4.7 – Etapa 2: Explicação para uma instância da classe de células jovens.

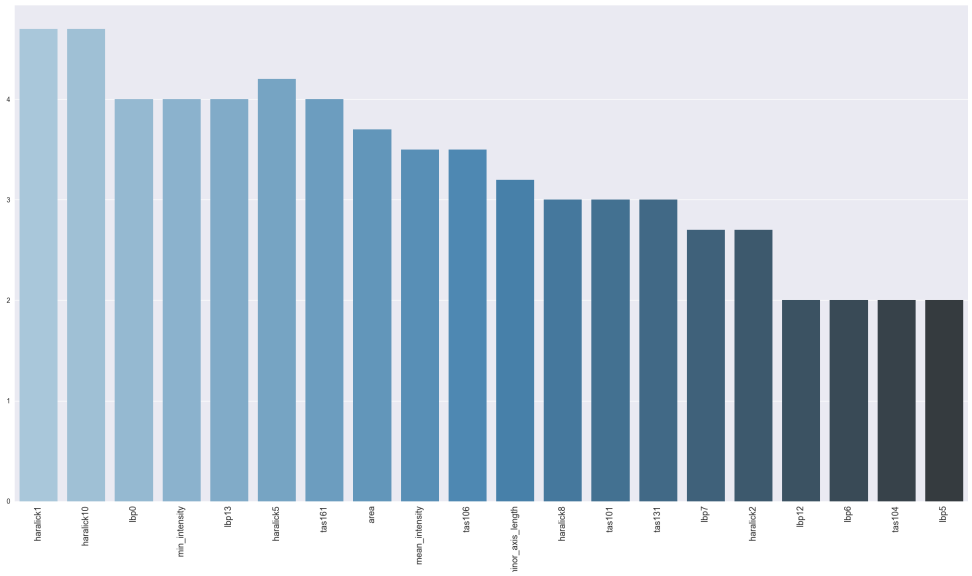

Figura 4.8 – Etapa 2: Histograma de características presentes nas explicações do SP-LIME.

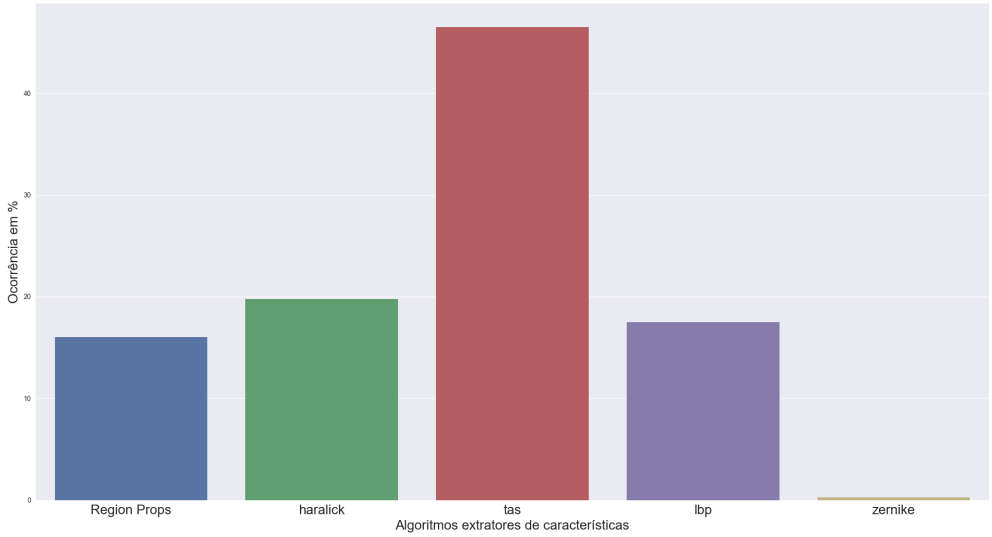

Figura 4.9 – Etapa 2: Histograma de extratores de características presentes nas explicações do SP-LIME.

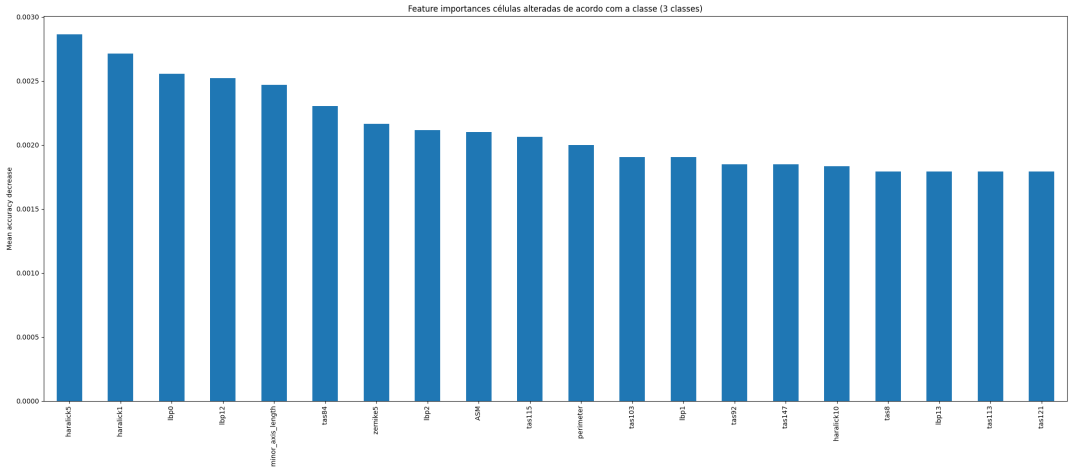

Figura 4.10 – Etapa 2: 20 maiores valores de *Feature Importance* para o modelo.

4.1.3 Etapa 3: Jovens

Para as células jovens, o classificação pode resultar em três classes. São elas: ASC-H, HSIL, SCC

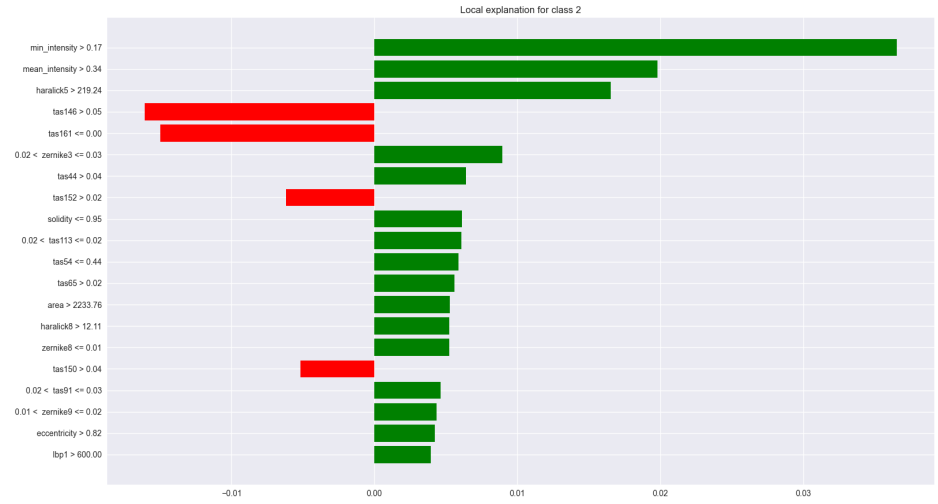

Figura 4.11 – Etapa 3: Explicação de uma instância da classe ASC-H.

]

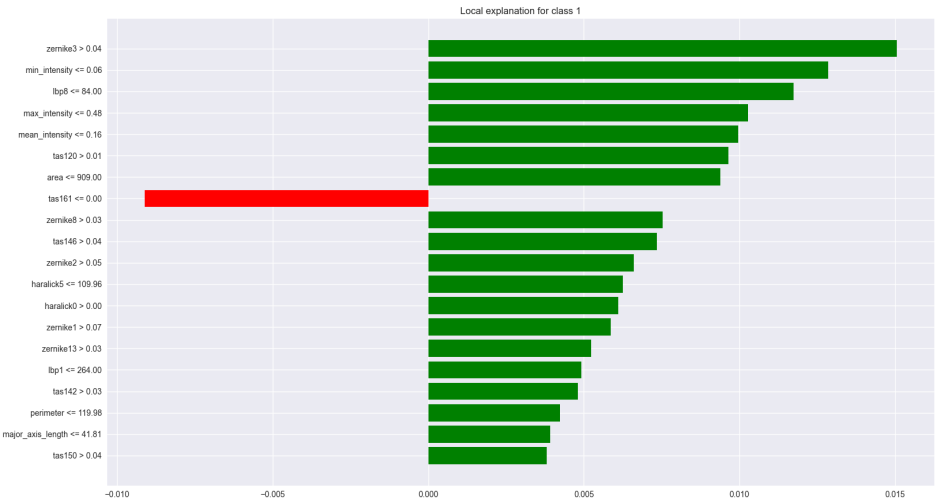

Figura 4.12 – Etapa 3: Explicação de uma instância da classe HSIL.

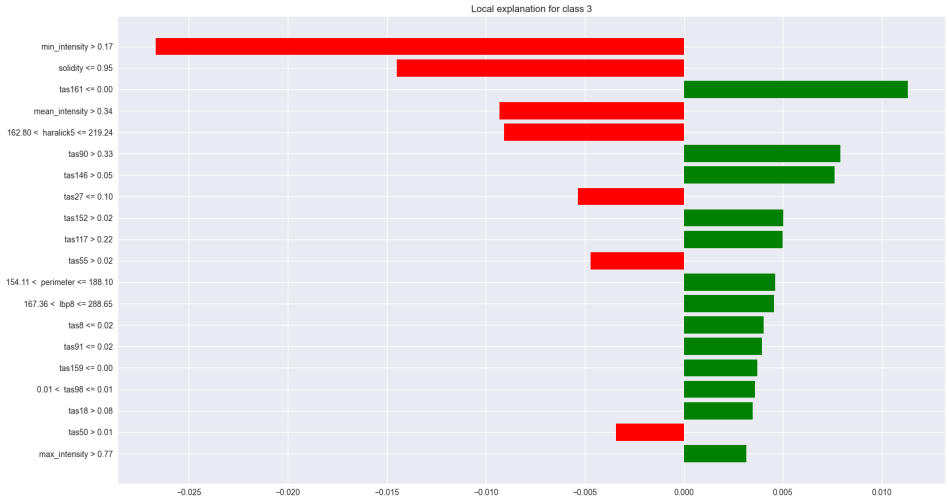

Figura 4.13 – Etapa 3: Explicação de uma instância da classe SC.

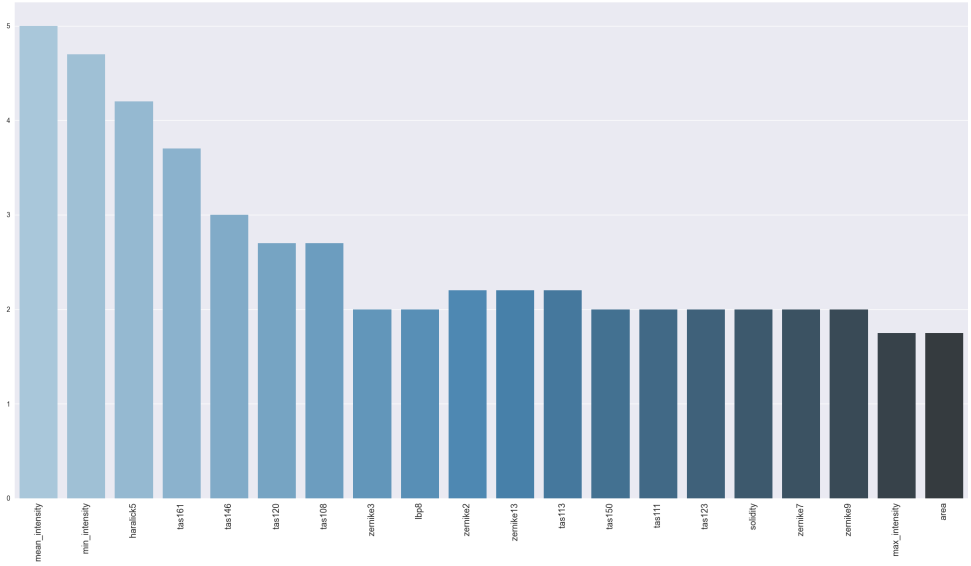

Figura 4.14 – Etapa 3: Histograma de características presentes nas explicações do SP-LIME.

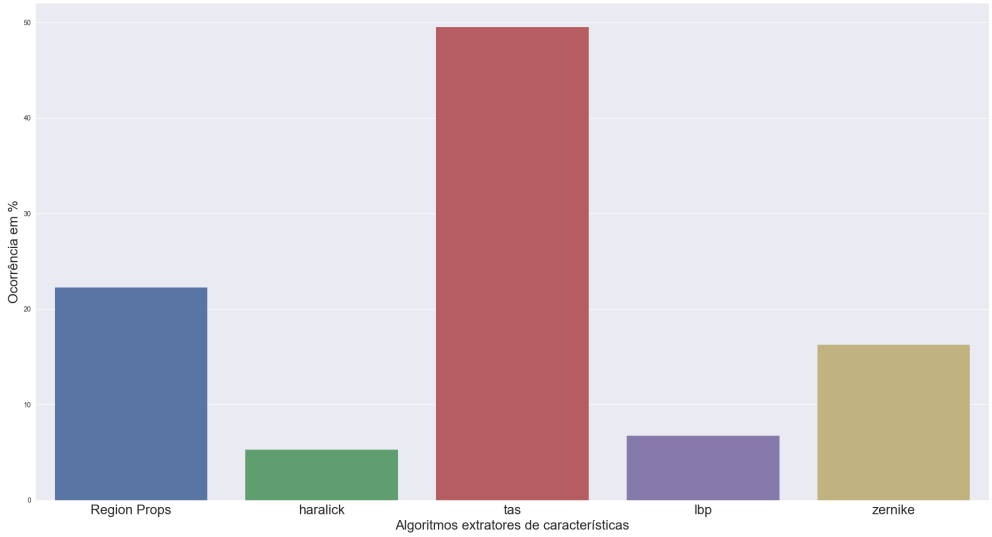

Figura 4.15 – Etapa 3: Histograma de extratores de características presentes nas explicações do SP-LIME.

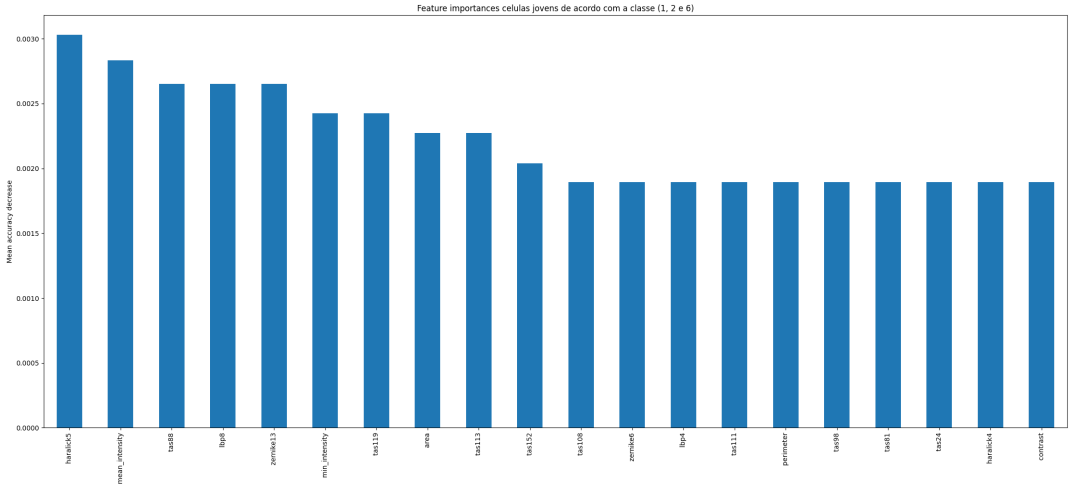

Figura 4.16 – Etapa 3: 20 maiores valores de *Feature Importance* para o modelo.

4.1.4 Etapa 4: Maduras

As células maduras são células alteradas, assim como as células jovens. As classes que representam células maduras são ASC-US e LSIL

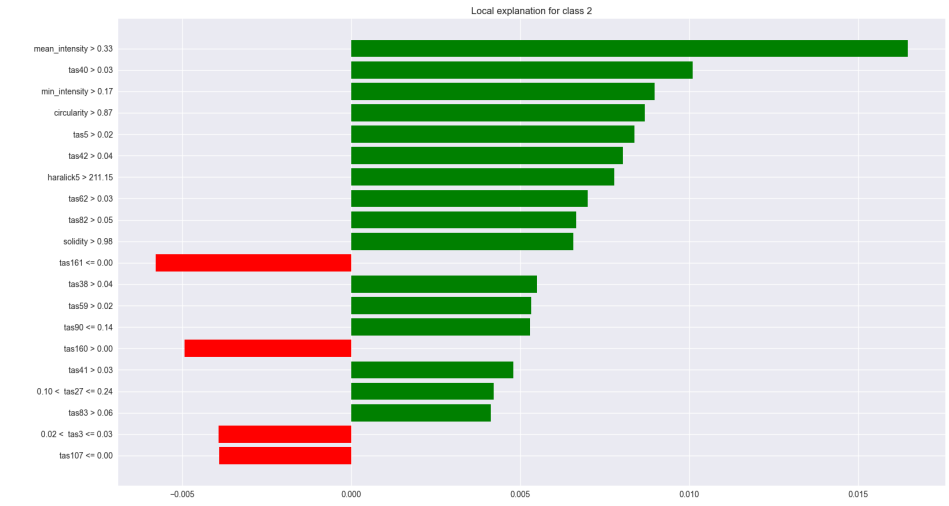

Figura 4.17 – Etapa 4: Explicação para uma instância da classe ASC-US.

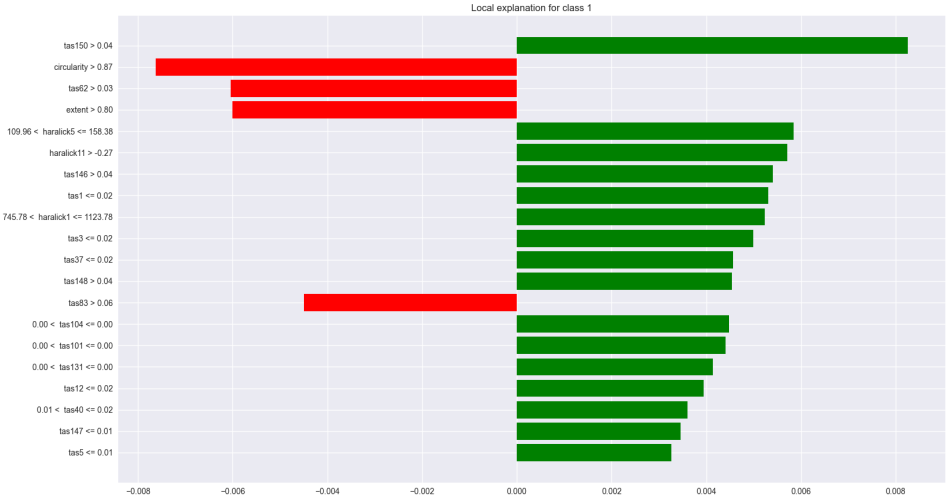

Figura 4.18 – Etapa 4: Explicação para uma instância da classe LSIL.

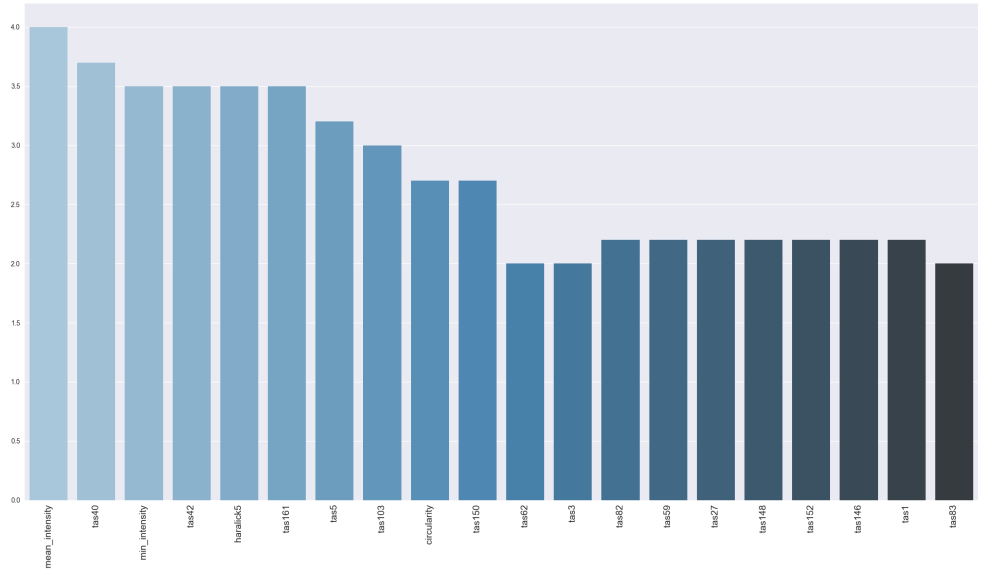

Figura 4.19 – Etapa 4: Histograma de características presentes nas explicações do SP-LIME.

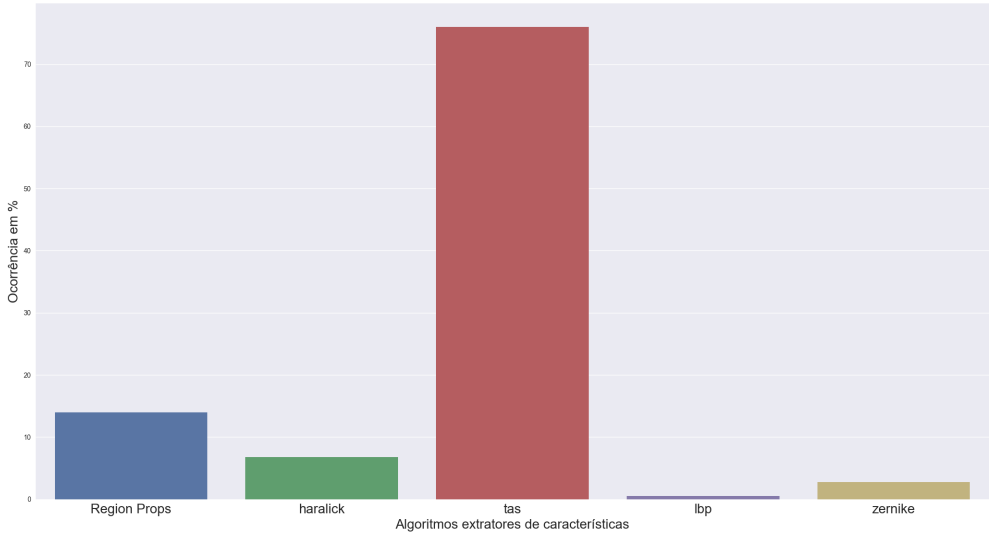

Figura 4.20 – Etapa 4: Histograma de extratores de características presentes nas explicações do SP-LIME.

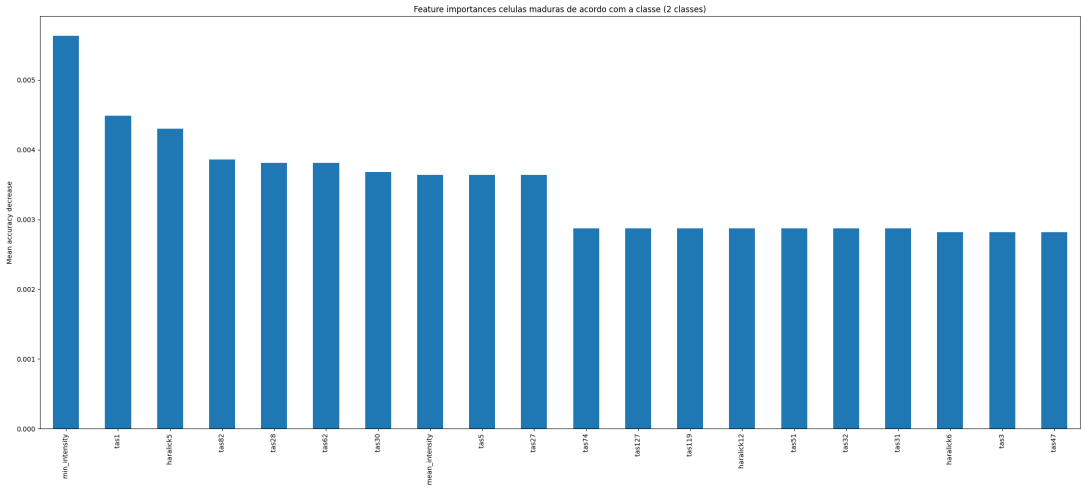

Figura 4.21 – Etapa 4: 20 maiores valores de *Feature Importance* para o modelo.

4.2 Discussões

O modelo de floresta aleatória, devido ao fato de trabalhar com características, possui em suas explicações regras baseadas nessas características. Porém, esses dados não denotam explicitamente a verdade biológica, os grupos de características estão codificados de acordo com o extrator que os gerou. Desse modo, para um citopatologista, a interpretabilidade do modelo não é totalmente transparente, visto que seria necessário decodificar os códigos mencionados.

Levando isso em consideração, os histogramas gerados a partir das explicações do SP-LIME para cada classificador visaram entender quais características e quais algoritmos extratores foram mais relevantes no processo de classificação. Em adição, o cálculo da *permutation_importance* foi usado como comparativo. Desse modo, é válido perceber as semelhanças e diferenças entre os métodos de forma a realizar uma análise dos resultados individual para cada etapa do desenvolvimento.

No modelo de floresta aleatória que classificou células normais e alteradas houve uma predominância de características morfológicas ressaltadas pela alta recorrência das características obtidas pelo algoritmo *Region Props*, como descrito pela Tabela 4.1, enquanto o resultado de *permutation_importance* além da presença de características morfológicas, também apresentou características voltadas para textura como representativas. Portanto, existe uma certa consonância nos resultados em ambos os métodos.

Tabela 4.1 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células normais/alteradas (etapa 1).

Extrator	Quantidade de aparições nas explicações do SP-LIME
Region Props	116
Haralick	50
TAS	112
LBP	100
Zernike	22
Total	400

Já no modelo que classifica células alteradas, as características relacionadas a localizações sub-celulares se mostraram recorrentes e dominantes nas explicações do SP-LIME representadas pelo extrator TAS, LBP e Haralick também se mostram relevantes, como mostrado na Tabela 4.2. Já no gráfico de importâncias é perceptível a presença de LBP e Haralick, em contrapartida as características que fazem parte do TAS não possuem números expressivos. Desse modo, os métodos possuem semelhanças nos resultados, porém não são plenamente iguais.

Tabela 4.2 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células alteradas (etapa 2).

Extrator	Quantidade de aparições nas explicações do SP-LIME
Region Props	64
Haralick	79
TAS	186
LBP	70
Zernike	1
Total	400

Em relação ao classificador de células maduras, o SP-LIME apresentou uma dominância de características relacionadas a localizações sub-celulares, mais especificamente geradas pelo algoritmo TAS, como exemplificado na Tabela 4.4, sendo que os resultados de *permutation_importance* também mostram a grande presença dessa classe de características. Em contrapartida, o classificador de células jovens não obteve a mesma harmonia entre os resultados, enquanto o SP-LIME ressaltou a importância de atributos gerados pelo TAS, com uma enorme diferença de ocorrência em relação aos outros extratores, vide a tabela 4.3 os resultados de importância da permutação foram mais distribuídos entre os algoritmos extratores, gerando uma diversidade no gráfico da figura 4.16.

Tabela 4.3 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células jovens (etapa 3).

Extrator	Quantidade de aparições nas explicações do SP-LIME
Region Props	89
Haralick	21
TAS	198
LBP	27
Zernike	65
$\boldsymbol{Total}$	400

Tabela 4.4 – Número de características por extrator presentes nas explicações do SP-LIME para o classificador de células maduras (etapa 4).

Extrator	Quantidade de aparições nas explicações do SP-LIME
Region Props	56
Haralick	27
TAS	304
LBP	2
Zernike	11
$\boldsymbol{Total}$	400

5 Conclusões e Trabalhos Futuros

A classificação de células cervicais através de algoritmos de aprendizado de máquina se mostra uma alternativa de auxílio a análise manual realizada por profissionais da saúde. Trabalhos como o realizado por Diniz et al. (2021b) apresentam alta taxa de efetividade em testes realizados a partir de bases de dados como a CRIC *Cervix*. Porém, a interpretabilidade desses modelos necessita ser explorada no contexto de citologia cervical, visto que para a adoção dos modelos supracitados na rotina média passa também pela possibilidade de interpretá-los.

Neste trabalho, foi realizada uma análise de interpretabilidade do modelo proposto por Diniz et al. (2021b), evidenciando através das explicações geradas pelo LIME, quais características foram mais importantes para os classificadores tomarem as decisões. Além disso, o agrupamento das características em algoritmos extratores forneceu uma visão mais ampla de quais tipos de características foram mais relevantes para o modelo. Tudo isso, comparado ao uso do *feature importances* gera insumos mais palpáveis para o profissional da saúde, visto que revelar os algoritmos extratores mais e menos importantes no processo de classificação denota também qual tipo de característica biológica é mais relevante, agregando assim um maior nível de informação ao modelo de classificação que já se mostrou eficaz nos testes feitos pela literatura.

Portanto, a metodologia proposta se mostrou capaz de revelar detalhes sobre o modelo até então ocultos e gerar uma análise de interpretabilidade estruturada em etapas de modo a exceder somente o uso as ferramentas fornecidas pelo explicador agnóstico LIME.

A partir dos resultados e conforme dito nas discusões apresentadas, é notável que as explicações geradas pelo LIME possuem limitações. Sendo assim, trabalhos futuros podem se concentrar em prover ao modelo de classificação que faz uso de CNN explicações de uma perspectiva global, visto que o Lime Image produz apenas explicações individuais. Já no modelo de Árvores Aleatórias, é interessante validar as explicações geradas e as ocorrências das características com profissionais da saúde, a fim de avaliar biologicamente cada etapa do desenvolvimento. Além disso, traçar comparativos entre o LIME e outros explicadores, tal qual o SHAP (LUNDBERG; LEE, 2017), é uma boa maneira de validar as explicações e evidenciar as semelhanças e diferenças entre os resultados dos explicadores no contexto de citologia cervical.

Referências

AGARWAL, N.; DAS, S. Interpretable machine learning tools: A survey. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. [S.l.: s.n.], 2020. p. 1528–1534.

ALTMANN, A.; TOLOşI, L.; SANDER, O.; LENGAUER, T. Permutation importance: a corrected feature importance measure. *Bioinformatics*, v. 26, n. 10, p. 1340–1347, 04 2010. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btq134>.

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. Machine bias. In: *Ethics of Data and Analytics*. [S.l.]: Auerbach Publications, 2016. p. 254–264.

BAHAD, P.; SAXENA, P. Study of adaboost and gradient boosting algorithms for predictive analytics. In: TOMAR, G. S.; CHAUDHARI, N. S.; BARBOSA, J. L. V.; AGHWARIYA, M. K. (Ed.). *International Conference on Intelligent Computing and Smart Communication 2019*. Singapore: Springer Singapore, 2020. p. 235–244. ISBN 978-981-15-0633-8.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 10 2001.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, v. 8, n. 8, 2019. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/8/8/832>.

CERVANTES, E. G.; CHAN, W.-Y. Lime-enabled investigation of convolutional neural network performances in covid-19 chest x-ray detection. In: *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. [S.l.: s.n.], 2021. p. 1–6.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

CHEN, W.; GAO, L.; LI, X.; SHEN, W. Lightweight convolutional neural network with knowledge distillation for cervical cells classification. *Biomedical Signal Processing and Control*, v. 71, p. 103177, 2022. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809421007746>.

CHOLLET, F. et al. *Keras*. GitHub, 2015. Disponível em: <https://github.com/fchollet/keras>.

CHONG, C.-W.; RAVEENDRAN, P.; MUKUNDAN, R. A comparative analysis of algorithms for fast computation of zernike moments. *Pattern Recognition*, v. 36, n. 3, p. 731–742, 2003. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320302000912>.

CHOWDHARY, C. L.; ACHARJYA, D. Segmentation and feature extraction in medical imaging: A systematic review. *Procedia Computer Science*, v. 167, p. 26–36, 2020. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S187705092030644X>.

COHEN, P. A.; JHINGRAN, A.; OAKNIN, A.; DENNY, L. Cervical cancer. *The Lancet*, v. 393, n. 10167, p. 169–182, 2019. ISSN 0140-6736. Disponível em: <https://www.sciencedirect.com/science/article/pii/S014067361832470X>.

CRUZ, H. F. D.; SCHNEIDER, F.; SCHAPRANOW, M.-P. Prediction of acute kidney injury in cardiac surgery patients: Interpretation using local interpretable model-agnostic explanations. In: *HEALTHINF*. [S.l.: s.n.], 2019.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Ensemble machine learning. *Ensemble Mach. Learn*, 2012.

DAVAGDORJ, K.; LI, M.; RYU, K. H. Local interpretable model-agnostic explanations of predictive models for hypertension. In: PAN, J.-S.; LI, J.; RYU, K. H.; MENG, Z.; KLASNJA-MILICEVIC, A. (Ed.). *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Singapore: Springer Singapore, 2021. p. 426–433. ISBN 978-981-33-6757-9.

DEY, N.; MISHRA, G.; KAR, J.; CHAKRABORTY, S.; NATH, S. A survey of image classification methods and techniques. In: . [S.l.: s.n.], 2014.

DINIZ, D. N.; REZENDE, M. T.; BIANCHI, A. G. C.; CARNEIRO, C. M.; LUZ, E. J. S.; MOREIRA, G. J. P.; USHIZIMA, D. M.; MEDEIROS, F. N. S. de; SOUZA, M. J. F. A deep learning ensemble method to assist cytopathologists in pap test image classification. *Journal of Imaging*, v. 7, n. 7, 2021. ISSN 2313-433X. Disponível em: <https://www.mdpi.com/2313-433X/7/7/111>.

DINIZ, D. N.; REZENDE, M. T.; BIANCHI, A. G. C.; CARNEIRO, C. M.; USHIZIMA, D. M.; MEDEIROS, F. N. S. de; SOUZA, M. J. F. A hierarchical feature-based methodology to perform cervical cancer classification. *Applied Sciences*, v. 11, n. 9, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/9/4091>.

GANAIE, M. A.; HU, M.; MALIK, A. K.; TANVEER, M.; SUGANTHAN, P. N. *Ensemble deep learning: A review*. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2104.02395>.

GARREAU, D.; LUXBURG, U. von. Explaining the explainer: A first theoretical analysis of lime. In: CHIAPPA, S.; CALANDRA, R. (Ed.). *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 2020. (Proceedings of Machine Learning Research, v. 108), p. 1287–1296. Disponível em: <https://proceedings.mlr.press/v108/garreau20a.html>.

GARREAU, D.; LUXBURG, U. von. *Looking Deeper into Tabular LIME*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2008.11092>.

GILPIN, L. H.; BAU, D.; YUAN, B. Z.; BAJWA, A.; SPECTER, M.; KAGAL, L. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1806.00069>.

GORDON, A. D. A review of hierarchical classification. *Journal of the Royal Statistical Society: Series A (General)*, v. 150, n. 2, p. 119–137, 1987. Disponível em: <https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2981629>.

GUNNING, D.; STEFIK, M.; CHOI, J.; MILLER, T.; STUMPF, S.; YANG, G.-Z. Xai—explainable artificial intelligence. *Science Robotics*, American Association for the Advancement of Science, v. 4, n. 37, p. eaay7120, 2019.

HAMILTON, N.; PANTELIC, R.; HANSON, K.; TEASDALE, R. Fast automated cell phenotype classification. *BMC bioinformatics*, v. 8, p. 110, 03 2007.

HUSSAIN, E.; MAHANTA, L. B.; DAS, C. R.; TALUKDAR, R. K. A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network. *Tissue and Cell*, v. 65, p. 101347, 2020. ISSN 0040-8166. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0040816619304872>.

ISIDORO., D.; CARNEIRO., C.; RESENDE., M.; MEDEIROS., F.; USHIZIMA., D.; BIANCHI., A. Automatic classification of cervical cell patches based on non-geometric characteristics. In: INSTICC. *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*. [S.l.]: SciTePress, 2020. p. 845–852. ISBN 978-989-758-402-2. ISSN 2184-4321.

KHALID, S.; KHALIL, T.; NASREEN, S. A survey of feature selection and feature extraction techniques in machine learning. In: *2014 Science and Information Conference*. [S.l.: s.n.], 2014. p. 372–378.

KIM, B.; KHANNA, R.; KOYEJO, O. O. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, v. 29, 2016.

KUMARAKULASINGHE, N. B.; BLOMBERG, T.; LIU, J.; LEAO, A. S.; PAPAPETROU, P. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.: s.n.], 2020. p. 7–12.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MAGESH, P. R.; MYLOTH, R. D.; TOM, R. J. An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, v. 126, p. 104041, 2020. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482520303723>.

MARINAKIS, Y.; DOUNIAS, G.; JANTZEN, J. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Computers in Biology and Medicine*, v. 39, n. 1, p. 69–78, 2009. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482508001674>.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, Elsevier, v. 267, p. 1–38, 2019.

MIRANDA, E.; ARYUNI, M.; IRWANSYAH, E. A survey of medical image classification techniques. In: *2016 International Conference on Information Management and Technology (ICIMTech)*. [S.l.: s.n.], 2016. p. 56–61.

NAYAR, R.; WILBUR, D. C. The bethesda system for reporting cervical cytology: a historical perspective. *Acta cytologica*, Karger Publishers, v. 61, n. 4-5, p. 359–372, 2017.

PIETIKÄINEN, M. Image analysis with local binary patterns. In: SPRINGER. *Scandinavian Conference on Image Analysis*. [S.l.], 2005. p. 115–118.

REZENDE, M. T.; SILVA, R.; BERNARDO, F. d. O.; TOBIAS, A. H. G.; OLIVEIRA, P. H. C.; MACHADO, T. M.; COSTA, C. S.; MEDEIROS, F. N. S.; USHIZIMA, D. M.; CARNEIRO, C. M.; BIANCHI, A. G. C. Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific Data*, v. 8, n. 1, p. 151, Jun 2021. ISSN 2052-4463. Disponível em: <https://doi.org/10.1038/s41597-021-00933-8>.

RIBEIRO, M.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, 2016. p. 97–101. Disponível em: <https://aclanthology.org/N16-3020>.

ROBINSON, R. L. M.; PALCZEWSKA, A.; PALCZEWSKI, J.; KIDLEY, N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *Journal of Chemical Information and Modeling*, v. 57, n. 8, p. 1773–1792, 2017. PMID: 28715209. Disponível em: <https://doi.org/10.1021/acs.jcim.6b00753>.

ROCK, C. L.; MICHAEL, C. W.; REYNOLDS, R.; RUFFIN, M. T. Prevention of cervix cancer. *Critical Reviews in Oncology/Hematology*, v. 33, n. 3, p. 169–185, 2000. ISSN 1040-8428. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1040842899000736>.

SHOUMY, N. J.; EHKAN, P.; YAAKOB, S. N.; ALI, M. S.; KHATUN, S. Feature extraction for neural network pattern recognition for bloodstain analysis. *International Journal of Applied Engineering Research*, v. 11, n. 15, p. 8583–8589, 2016.

SUN SHAOBO LI, Y. C. G.; LANG, F. Cervical cancer diagnosis based on random forest. *International Journal of Performability Engineering*, Int J Performability Eng, v. 13, n. 4, p. 446, 2017. Disponível em: <http://www.ijpe-online.com/EN/abstract/article_3888.shtml>.

VELLIDO, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, v. 32, p. 1–15, 12 2020.

WAGGONER, S. E. Cervical cancer. *The Lancet*, v. 361, n. 9376, p. 2217–2225, 2003. ISSN 0140-6736. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0140673603137786>.

WRIGHT, T. C. Cervical cancer screening in the 21st century: is it time to retire the pap smear? *Clinical obstetrics and gynecology*, v. 50, n. 2, p. 313—323, June 2007. ISSN 0009-9201. Disponível em: <https://doi.org/10.1097/GRF.0b013e31804a8285>.

ZHANG, Y.; TINO, P.; LEONARDIS, A.; TANG, K. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 5, n. 5, p. 726–742, oct 2021. Disponível em: <https://doi.org/10.1109%2Ftetci.2021.3100641>.