Classificações Explicáveis para Imagens de Células Infectadas por Malária

Iam Palatnik de Sousa[1], **Marley M.B.R. Vellasco**[1], **Eduardo Costa da Silva**[1]

[1]Departamento de Engenharia Elétrica – Pontifícia Universidade Católica do Rio de Janeiro

iam.palat@gmail.com

Abstract. *This work presents the development of an explainable image classifier, trained for the task of determining whether a cell is infected by malaria. The classifier consists of a residual neural network, with a classification accuracy of 96%, trained on the National Institute of Health Malaria Dataset. Explainable Artificial Intelligence (XAI) techniques are employed to make classifications more interpretable. These explanations are generated by using two different methodologies: Local Interpretable Model Agnostic Explanations (LIME) and SquareGrid. The explanations provide novel, key insights into the decision making patterns of such a high performing model for a medical task.*

Resumo. *Este trabalho apresenta o desenvolvimento de um classificador explicável de imagens, treinado para a tarefa de determinar se uma célula foi infectada por malária. O classificador consiste em uma rede neural residual, com acurácia de classificação de 96%, treinada com o dataset de Malária do National Health Institute. Técnicas de Inteligência Artificial Explicável foram aplicadas para tornar as classificações mais interpretáveis. Estas explicações são geradas usando duas metodologias: Local Interpretable Model Agnostic Explanations (LIME) e SquareGrid. As explicações fornecem perspectivas novas e importantes para os padrões de decisão de modelos como este, de alto desempenho para tarefas médicas.*

1. Introdução

Segundo a Organização Mundial da Saúde (World Health Organization – WHO), a malária ocasiona centenas de milhões de infecções por ano, resultando em centenas de milhares de mortes [WHO 2019]. Ela é uma doença infecciosa causada por parasitas do gênero Plasmodium, um tipo de protozoário que infecta hemácias ao entrar na corrente sanguínea. A morfologia deste parasita permite que ele seja visualmente identificado em esfregaços de sangue, conforme descrito nas diretrizes do Centro para Controle de Doenças dos Estados Unidos (*Center for Disease Control* – CDC). Ainda segundo o CDC (2020), a microscopia continua sendo o "padrão-ouro" para confirmação laboratorial de malária.

Consequentemente, a principal forma de diagnóstico para esta doença envolve a identificação de células infectadas por um microscopista treinado. Adicionalmente, ressalta-se que o diagnóstico rápido desta enfermidade, em seus estágios iniciais, é tido como um dos principais fatores para um tratamento efetivo [Sayyed et al. 2019]. Esses aspectos, contudo, aumentam a demanda por centros de diagnóstico com patologistas treinados para esta tarefa.

Em 2017 [OPAS 2019], estima-se que houve 219 milhões de casos de malária em 90 países, com um total de 435 mil mortes. Por sua vez, países do continente Africano concentram uma carga desproporcionalmente alta dos casos de malária, tendo sido os responsáveis por 92% do total de casos notificados e por 93% das mortes registradas. Dados os altos índices de infecção por malária, a demanda por centros com patologistas treinados pode se tornar insustentável em alguns casos, especialmente em países de baixa renda [Sayyed et al. 2019].

Gradualmente, a utilização de sistemas de Inteligência Artificial (IA) para análises de imagens em aplicações médicas tem sido relatada na literatura [Kermany et al. 2018]. Estes sistemas de auxílio à decisão em muitos casos apresentam desempenhos similares ao de patologistas humanos, com a vantagem adicional da escalabilidade. Isso torna a presença de sistemas de *Machine Learning* (ML) e *Deep Learning* (DL) cada vez mais comuns na área médica [Miotto et al. 2017]. Dado este panorama, a utilização de redes neurais e outros sistemas relacionados para a identificação de células infectadas por malária tem se tornado uma área cada vez mais estudada.

Recentemente, a maior base de dados já publicada para esta tarefa, com mais de 27000 imagens, foi disponibilizada pelo Instituto Nacional de Saúde americano (*National Institute of Health* – NIH) [Rajaraman et al. 2018]. Com esta base de dados, Rajaraman et al. (2018) treinaram diversos classificadores, demonstrando que a rede ResNet50 tem o melhor desempenho quando comparada com outras arquiteturas tradicionalmente usadas para tarefas de classificação de imagem (em particular: VGG-16, AlexNet, Xception e DenseNet-121). A acurácia obtida foi cerca de 95%, utilizando validação cruzada 5-fold. Similarmente, Sayyed et al. (2019) também obtiveram acurácias de aproximadamente 95% na mesma base de dados, utilizando modelos customizados de redes neurais convolucionais (*Convolutional Neural Network* – CNN) e Capsule Networks.

Vale salientar, contudo, que estes modelos treinados não estão disponíveis publicamente. Por esta razão, um dos objetivos deste trabalho é treinar uma rede neural utilizando a mesma base de dados do NIH, que gradualmente vem se tornando a principal para esta tarefa. Baseado nos resultados de Rajaraman et al. (2018), a arquitetura escolhida para o presente estudo é a ResNet50 [He et al. 2016].

Entretanto, ressalta-se que a utilização de redes neurais e sistemas similares para tarefas de classificação de imagem possui um problema considerável, que se torna ainda mais crítico quando se trata de aplicações médicas. Estes modelos não lineares, complexos, em geral apresentam um comportamento descrito na literatura como "caixa-preta" ou black box. Isso significa que não é possível determinar exatamente o que ocasionou uma saída gerada pelo modelo em função de uma dada entrada, de modo que não se consegue avaliar o processo de tomada de decisão do sistema de classsificação.Em outras palavras, o modelo não é transparente para um usuário humano.

Essa opacidade cria obstáculos para a adoção mais abrangente de sistemas de IA na clínica, pois a área médica, bem como outras áreas, requer transparência em relação aos aspectos considerados para decisões críticas [Adadi and Berrada 2018]. Recentemente essa questão tem sido mais discutida na literatura científica, e uma área conhecida como Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) passou a ser abordada com mais frequência.

Desde então, diversas técnicas para gerar explicações a partir de classificações foram desenvolvidas. Para o caso de modelos de classificação de imagens, estas técnicas focam em criar visualizações compreensíveis para um usuário humano, sendo capazes de mostrar quais partes de uma imagem são mais relevantes para uma dada classificação. Isso permite auditar um sistema de classificação baseado em IA, aumentando a transparência do sistema de auxílio à decisão.

Dentre as principais técnicas de XAI publicadas nos últimos anos, a *Local Interpretable Model Agnostic Explanations* (LIME), de Ribeiro et al. (2016), ganhou destaque por sua universalidade. Devido à sua natureza agnóstica ao modelo, é possível aplicar esta técnica em qualquer classificador. Isso torna as análises facilmente comparáveis para quaisquer novos classificadores treinados para uma tarefa ou base de dados em particular.

Embora a necessidade de modelos de IA explicáveis para medicina esteja claramente delineada na literatura, ainda há poucos estudos que apliquem estas técnicas recentemente desenvolvidas em tarefas relacionadas à base de dados médicas. Um dos primeiros estudos neste sentido [Palatnik de Sousa et al. 2019] mostrou que é possível usar a técnica LIME para gerar explicações para as classificações de uma CNN que identifica metástases em patches de histopatologia da base de dados Patch-Camelyon [Veeling et al. 2018]. Adicionalmente, ficou demonstrado que estas explicações estavam majoritariamente de acordo com anotações feitas por médicos patologistas [Palatnik de Sousa et al. 2019]. Esse estudo também evidenciou que, ao menos no caso das metástases, o algoritmo de segmentação utilizado para gerar as explicações influencia bastante os resultados. Dessa forma, foi proposta uma alternativa para geração de explicações simplificadas, sem parâmetros e com custo computacional significativamente menor, chamada de *SquareGrid*.

Considerando a relevância da potencial aplicação de classificadores automáticos para identificação de malária, e dada a discussão acima, denota-se a preocupação com a transparência de tais sistemas. Dessa forma, o objetivo principal deste trabalho é treinar um classificador explicável com a base de dados de malária do NIH. Como discutido anteriormente, a arquitetura escolhida é a ResNet50 e a estratégia para geração de explicações inclui as técnicas LIME e *SquareGrid*. Consequentemente, ressalta-se que o foco principal do presente trabalho não é necessariamente treinar uma rede com uma acurácia mais alta que todas as anteriores, mas sim treinar uma rede com acurácia similarmente alta em relação àquelas da literatura [Rajaraman et al. 2018] [Sayyed et al. 2019], e torná-la mais transparente ao estudar seu comportamento com técnicas de XAI.

2. Materiais e Métodos

2.1. Base de Dados

A base de dados de imagens de malária do NIH [Rajaraman et al. 2018] possui 27558 imagens de microscopia, com tamanhos variados, incluindo uma hemácia cada. Por sua vez, o fundo de cada uma destas imagens é excluído através de pré-processamento e convertido em pixels pretos.

As imagens foram coletadas de 150 pacientes infectados por Plasmodium falciparum e 50 não-infectados, usando uma câmera de celular acoplada a um microscópio no Chittagong Medical College de Bangladesh. A base de dados é balanceada entre duas

classes com exatamente o mesmo número de instâncias. A classe das imagens foi manualmente anotada por um especialista em leitura de lâminas de extensão sanguínea da Mahidol-Oxford Tropical Research Unit em Bangkok, sendo que a classe 0 corresponde à hemácias saudáveis e a classe 1 à hemácias infectadas. A Figura 1 mostra alguns exemplos de imagens de ambas as classes da base de dados.

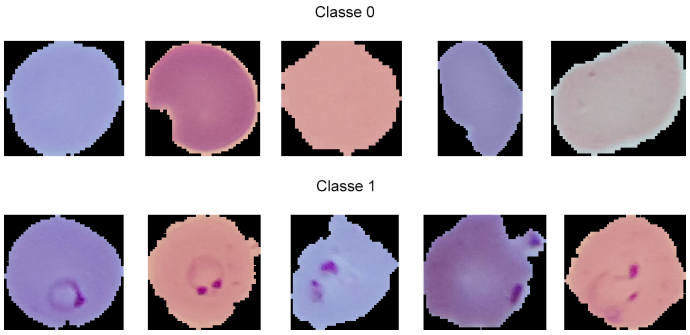

Figura 1. Exemplos de imagens da base de dados do NIH. A classe 0 corresponde às células não-infectadas. A classe 1 corresponde às células infectadas.

Como as imagens da base de dados têm dimensões variadas, é necessário redimensioná-las para um tamanho uniforme, antes de efetuar o treinamento da CNN. Dessa forma, no presente trabalho, as imagens foram redimensionadas para 64 por 64 pixels. A escolha destas dimensões permite reduzir o custo computacional associado ao treinamento da rede, enquanto possibilita a obtenção de acurácias comparáveis com às explicitadas na literatura, conforme será discutido mais adiante.

2.2. Classificador

O classificador escolhido para este estudo é uma rede neural residual, mais especificamente o modelo ResNet50 de He et al. (2016). Em particular, foi utilizada a versão da ResNet50 integrada ao *framework* Keras, para realização dos experimentos descritos ao longo deste trabalho. A ResNet50 é uma arquitetura consagrada para classificação de imagens, que consiste em vários blocos de convolução, dispostos de forma similar a uma CNN. No entanto, as redes residuais incluem conexões que pulam algumas camadas, conhecidas como *skip-connections*, que permitem que a rede identifique quais blocos de convolução estão contribuindo mais significativamente para o aprendizado. Esta característica possibilitou o treino de redes consideravelmente mais profundas do que as CNN previamente utilizadas [He et al. 2016].

Ressalta-se que a única adaptação necessária da arquitetura base da ResNet50 para o problema aqui tratado é a substituição do bloco de classificação por uma camada densamente conexa com 2 neurônios (correspondendo às duas classes) e ativação softmax.

2.3. Treinamento

A base de dados descrita na subseção 2.1 foi dividida em três subconjuntos: treino/validação/teste. Essa divisão foi feita em porcentagens de 80/10/10 do total de imagens, respectivamente, correspondendo aa 22046/2756/2756 imagens em cada um desses subconjuntos.

Após testes preliminares, foi notado que a acurácia de validação da rede converge rapidamente (em geral já na terceira época de treino) para valores ao redor de 95%.

Ainda assim, o treinamento foi estendido até 20 épocas, usando um checkpoint baseado na acurácia do conjunto de validação. Na sequência, o melhor modelo obtido ao final das 20 épocas de treinamento foi usado para avaliar as imagens do conjunto de teste, que não foram apresentadas para a rede durante o treino/validação.

Destaca-se que o tamanho do batch utilizado foi 128, a função de Loss aplicada foi a '*categorical crossentropy*' e o otimizador utilizado foi o Gradiente Descendente Estocástico (*Stocastic Gradient Descent* – SGD). A decisão de usar SGD ao invés de outros otimizadores adaptativos comumente usados, tais como Adam, foi feita por observações e experimentos da literatura que mostram que frequentemente o SGD, apesar de convergir mais lentamente, gera modelos que generalizam melhor para além do subconjunto de treinamento [Keskar and Socher 2017].

2.4. LIME

A técnica LIME consiste em gerar explicações para um modelo do tipo caixa-preta, apresentando instâncias que desejem ser explicadas individualmente [Ribeiro et al. 2016]. Para o caso discutido neste trabalho, isso significa apresentar imagens individuais da base de dados, para que sejam geradas explicações associadas a cada imagem específica. A seguir, são descritas as principais etapas do mecanismo utilizado para obtenção destas explicações.

Inicialmente, a imagem cuja classificação se deseja explicar é dividida em regiões representativas. Esse processo é denominado na literatura como segmentação, sendo responsável por gerar sub-regiões frequentemente chamadas de segmentos ou superpixels. Por sua vez, estes superpixels em geral são áreas contíguas da imagem com características similares, como por exemplo texturas ou cores. O algoritmo de segmentação escolhido foi o de Felzenszwalb-Huttenlocher (FHA) [Felzenszwalb and Huttenlocher 2004], por motivos delineados na próxima subseção.

Em seguida, são geradas diversas versões perturbadas da imagem segmentada, em que os superpixels são aleatoriamente cobertos pela cor preta, criando uma distribuição de imagens perturbadas. Frequentemente, centenas ou milhares de imagens são geradas nesta etapa. Em particular, ressalta-se que no presente estudo foram utilizadas 1000 imagens perturbadas por explicação. Na sequência, a distribuição de imagens perturbadas é apresentada ao modelo estudado (ResNet50), o qual realiza as respectivas classificações.

Com a saída softmax do modelo para as imagens perturbadas e a distribuição de superpixels perturbados, um modelo linear é treinado a fim de determinar o peso de cada superpixel no processo de classificação do modelo caixa-preta. O resultado deste modelo linear permite atribuir a cada superpixel um peso explicativo (explanation weight – *xw*). Caso um superpixel contribua a favor da classificação para uma dada classe seu peso explicativo será positivo, caso contrário será negativo. Ademais, o valor absoluto dos pesos explicativos será tão maior quanto mais influente for o superpixel para aquela classificação.

No último passo do processo de geração das explicações, estes pesos explicativos são plotados para cada superpixel, gerando um *heatmap* que mostra as regiões mais relevantes para a classificação, compondo a explicação final para a classificação de uma dada imagem.

2.5. Algoritmo de Segmentação

Devido à natureza do método aqui descrito, a divisão inicial da imagem em superpixels representativos é um passo extremamente importante para geração de explicações significativas. Essas regiões devem fazer sentido para o usuário humano, que busca uma explicação sobre a classificação da IA.

A segmentação FHA já foi aplicada com sucesso, na literatura, para uma tarefa similar de classificação de imagens histopatológicas [Palatnik de Sousa et al. 2019]. Esse algoritmo consiste numa *oversegmentation* da imagem feita através de *clustering* das regiões com base no método de minimum spanning trees [Felzenszwalb and Huttenlocher 2004].

O tamanho e número de superpixels gerados só pode ser controlado indiretamente. O parâmetro regulador para tal, scale, cria uma escala de observação para a função de *thresholding* do algoritmo FHA, durante o cálculo da diferença mínima entre os componentes da imagem [Felzenszwalb and Huttenlocher 2004]. Em geral valores maiores de scale criam superpixels maiores, e em menor número.

Testes preliminares realizados com a base de dados de malária analisada no presente trabalho mostraram que o algoritmo FHA também consegue gerar boas segmentações para este novo problema, separando adequadamente as manchas roxas características da classe 1 do citoplasma (ver Figura 2). Verificou-se ainda que, para imagens da base de dados de malária, o desempenho do FHA é aprimorado ao se ajustar o parâmetro scale, que indiretamente influencia o tamanho dos superpixels, para 50 (o valor padrão é 1). Esta heurística parece funcionar bem para obter superpixels que separam as manchas roxas e algumas nuances de coloração do citoplasma. Por esse motivo, ela foi utilizada em todas as explicações geradas neste trabalho.

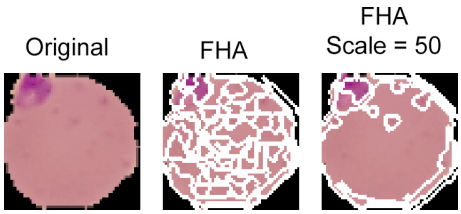

Figura 2. Segmentação FHA de uma imagem do banco de dados. A coluna do meio mostra os superpixels encontrados com os parâmetros padrão. A coluna da direita mostra o resultado quando o parâmetro scale é ajustado para um valor de 50, separando melhor as manchas roxas do citoplasma.

Uma varredura mais detalhada de parâmetros e a análise de outros métodos de segmentação certamente poderiam ser proveitosos. No entanto, isso fica fora do escopo deste trabalho, dado o tempo e custo computacional requerido para todos os testes, e dado que a heurística encontrada, com este algoritmo já usado para aplicação similar [Palatnik de Sousa et al. 2019], parece ser suficiente para segmentar as regiões de maior interesse das imagens pertencentes à base de dados aqui analisada.

2.6. *SquareGrid*

A dificuldade descrita na seção anterior, relacionada à geração de segmentações úteis para as explicações, foi discutida em [Palatnik de Sousa et al. 2019], que propuseram um

método alternativo independente de parâmetros de segmentação. Esse método, denominado *SquareGrid*, consiste em dividir a imagem em grades quadradas progressivamente menores. A Figura 3 mostra um exemplo com grades de 16, 64 e 256 quadrados respectivamente.

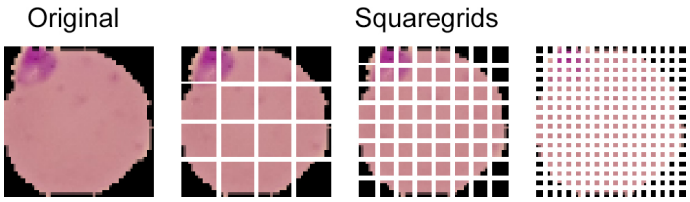

Figura 3. *SquareGrid* aplicado sobre uma imagem do dataset. Grades de 16, 64 e 256 quadrados, respectivamente, à direita da figura original.

Deve-se notar que, embora as grades não necessariamente sigam os contornos das diferentes cores e texturas da imagem, usar as grades de diferentes tamanhos para gerar explicações com o LIME permite criar um *heatmap* grosseiro, mas aproximado das regiões mais relevantes. Ademais, apesar de menos precisa em comparação com o FHA, esta técnica possibilita uma significativa redução do tempo de processamento computacional demandado para geração das explicações. O *SquareGrid* foi demonstrado como uma aproximação viável quando anteriormente avaliado para o caso de um problema de histopatologia [Palatnik de Sousa et al. 2019]. Por esse motivo, decidiu-se também avaliar o emprego desta metodologia neste trabalho.

3. Resultados e Discussão

O procedimento de treino descrito na seção anterior resultou em uma ResNet50 com acurácia de aproximadamente 96% para as imagens do conjunto de teste. Este valor é comparável àqueles apresentados na literatura [Rajaraman et al. 2018] [Sayyed et al. 2019], parecendo ser próximo ao máximo esperado para o desempenho desta arquitetura na classificação de imagens contidas na base de dados analisada. Por sua vez, a Figura 4 mostra a matriz de confusão resultante da aplicação do modelo implementado para os dados do conjunto de teste.

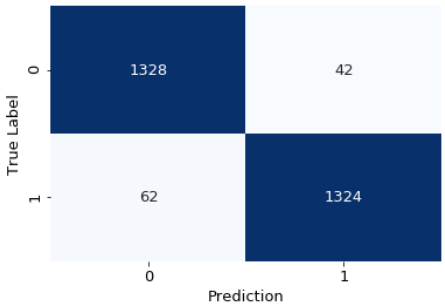

Figura 4. Matriz de confusão resultante da aplicação da ResNet50 treinada às imagens do conjunto de teste.

A figura 5 mostra o comportamento das acurácias de validação e teste ao longo do treino. Notavelmente, a época 16 teve a melhor acurácia de validação, e portanto os pesos dessa época foram utilizados.

Em seguida, foram geradas explicações utilizando as técnicas LIME e *SquareGrid* para diversas imagens do conjunto de teste, conforme descrito na seção anterior. A Figura 6 ilustra alguns exemplos dos resultados obtidos para 6 imagens classificadas como infectadas, mostrando que as regiões com manchas roxas de fato correspondem a superpixels com valor explicativo mais alto, conforme esperado. É possível perceber claramente que a coloração nessas regiões do *heatmap* apresenta tons mais escuros de azul.

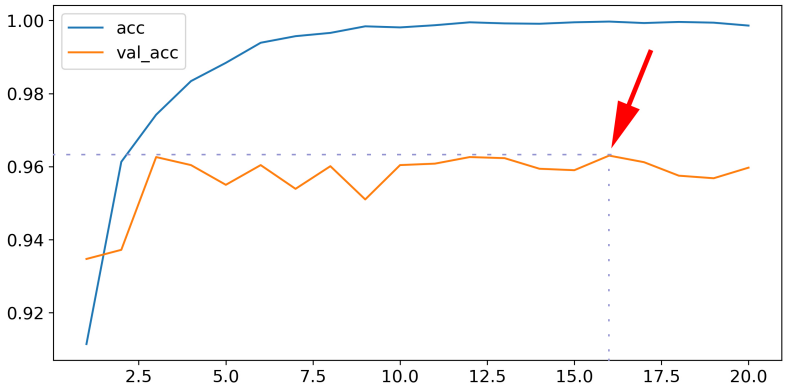

Figura 5. Treino da ResNet50. Seta em vermelho indica a época 16, onde a validação de acurácia atingiu seu maior valor. Vale notar que a partir da época 3 os valores já passam a oscilar ao redor de 0,95 e 0,96.

O fato das explicações realçarem justamente as estruturas mais características das células infectadas parece confirmar que a heurística usada para a segmentação FHA é útil também para este caso. Adicionalmente, é encorajador notar que, na grande maioria dos casos, as explicações geradas através da metodologia *SquareGrid* parecem destacar as mesmas regiões realçadas pelo FHA com pesos explicativos mais altos. Dessa forma, isso também parece confirmar, assim como na aplicação anterior do *SquareGrid* [Palatnik de Sousa et al. 2019] para uma base de dados de histopatologia, que as explicações via *SquareGrid* de fato conseguem aproximar, mesmo que de forma grosseira, os algoritmos mais sofisticados de segmentação, ou ao menos dar uma ideia das regiões mais relevantes da imagem original.

A terceira coluna da Figura 6 circunda com contornos brancos as regiões da imagem original onde os *heatmaps* gerados pela aplicação do LIME, explicitados na segunda coluna, possuem pesos explicativos superiores a um dado *threshold* arbitrário, que neste caso foi feito igual a 0,2. Esta coluna apresenta uma maneira alternativa de visualizar os resultados, que potencialmente seria interessante para um profissional médico usando esta técnica. Ademais, ao visualizar diretamente quais são as áreas de maior peso explicativo na imagem original, percebe-se que se tratam das áreas de coloração roxa mais acentuada, conforme esperado. Ainda nesse sentido, trabalhos futuros podem focar em comparar estas análises com a opinião de especialistas.

Outro aspecto importante sobre definir *thresholds* de pesos explicativos é que, como já notado na literatura [Palatnik de Sousa et al. 2019], áreas com pesos explicativos muito baixos apresentam maior variabilidade entre execuções diferentes do LIME, para uma mesma imagem. Dessa forma, são menos confiáveis para interpretação dos resultados. Por outro lado, é encorajador que, conforme mostrado na Figura 6, justamente as

áreas de pesos mais altos da imagem correspondam a estruturas biológicas claramente demarcadas nas imagens de infecções.

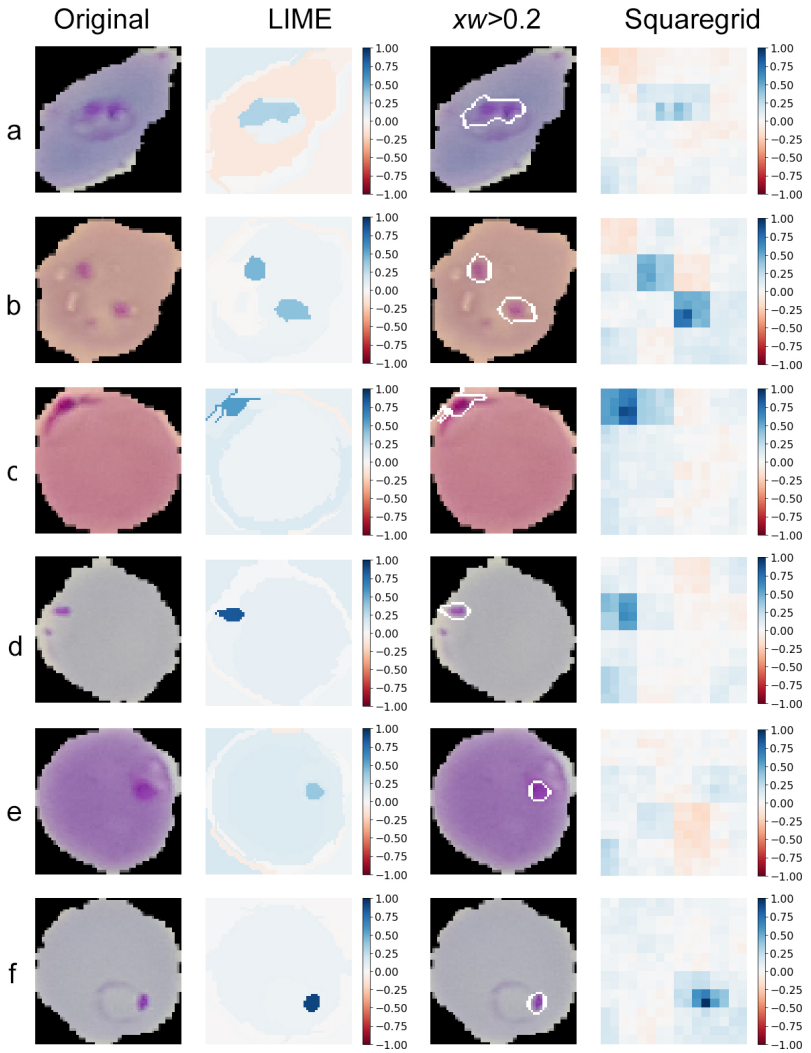

Figura 6. Linhas (a) até (f): Explicações para diversas imagens da classe 1 (infectadas) do conjunto de teste. Da esquerda para a direita as colunas representam, respectivamente, as imagens originais, as explicações por LIME, as regiões da imagem original onde os pesos explicativos ficaram acima de 0,2 e a explicação obtida pelo *SquareGrid*. A escala de cor dos heatmaps se encontra ao lado de cada mapa, e vai de -1 a 1 (vermelho à azul).

Observando-se especificamente os resultados obtidos pelo *SquareGrid*, nota-se que a imagem na linha (e) da Figura 6 parece ser a única para a qual a explicação obtida por esta técnica falhou em encontrar regiões de peso explicativo maior. Destaca-se ainda que, para a imagem da linha (e), apesar da explicação via LIME ter sido exitosa, pode-se perceber que ela atribuiu pesos explicativos consideravelmente menores para a mancha roxa, quando comparados com os pesos presentes nas demais explicações.

A maior dificuldade observada para a geração de explicações para a imagem da linha (e) pode estar associada à mancha roxa nela presente ser mais difusa, apresentando menor contraste em relação ao restante da imagem. Ademais, no caso específico do *SquareGrid*, o problema também pode ser atribuído ao fato da mancha localizar-se justamente

entre alguns dos quadrados que compõem a grade, de modo que nenhum deles foi capaz de capturar totalmente o peso explicativo daquela região. Para casos como este, poderia ser potencialmente benéfica uma varredura de parâmetros mais minuciosa, objetivando gerar segmentações mais aprimoradas. Essas varreduras que requerem maior custo computacional, sendo consideravelmente mais demoradas, ficarão como tópicos para trabalhos futuros.

Especificamente, a geração de uma explicação via LIME com FHA levou cerca de 16 segundos para uma dada combinação de parâmetros. É possível notar como o tempo computacional aumentaria caso centenas ou milhares de combinações de parâmetros fossem testadas. As explicações *SquareGrid*, por sua vez, são geradas em cerca de 80 segundos. Embora seja um tempo maior comparado com uma única explicação LIME comum, frequentemente é necessário fazer vários testes de parâmetros com a técnica LIME. Não foi diferente neste caso, no qual vários testes foram realizados a fim de encontrar uma heurística boa para o parâmetro *scale*.

Os resultados obtidos parecem sugerir uma possível estratégia para geração de explicações em aplicações clínicas de XAI, onde um sistema caixa-preta de classificação, como a ResNet50 aqui estudada, precisasse ser auditada. Inicialmente, as explicações poderiam ser geradas com a técnica *SquareGrid*, que é rápida e não requer nenhum ajuste de parâmetros por especialista. Em seguida, caso explicações mais detalhadas fossem necessárias, poderia ser utilizada alguma técnica de segmentação mais sofisticada, como o algoritmo FHA com alguma heurística tal como a encontrada neste trabalho, para esta base de dados específica.

Por fim, caso ainda mais detalhes sejam necessários, é possível realizar uma investigação de mais algoritmos de segmentação, com o intuito de encontrar as segmentações que gerem explicações de maior peso explicativo e relevância. No entanto, como mencionado anteriormente, esta última etapa estava fora do escopo deste trabalho, e fica como ideia para trabalhos futuros.

4. Conclusão

Um sistema de classificação explicável para imagens de células infectadas por malária foi apresentado neste trabalho. Através do uso de segmentação Felzenszwalb com uma heurística para o parâmetro scale, foi possível obter explicações que identificavam manchas roxas características das imagens de classe 1 (infectadas) como tendo pesos explicativos mais altos.

Como ideias para trabalhos futuros, estes promissores resultados iniciais podem ser comparados com as opiniões de especialistas da área médica, e podem ser implementadas varreduras de parâmetros mais custosas computacionalmente, a fim de se obter segmentações mais detalhadas.

Os resultados aqui apresentados contribuem para aumentar a transparência e confiabilidade de sistemas de classificação por IA, que normalmente são tidos como caixas-pretas (black-boxes). A análise dos aspectos considerados no processo de tomada de decisão destes sistemas de classificação é de especial importância para inúmeras aplicações na área médica, como por exemplo o estudo de caso aqui tratado.

5. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores agradecem ainda ao CNPq e a FAPERJ pelos fundos que financiaram esta pesquisa.

Referências

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.

Keskar, N. S. and Socher, R. (2017). Improving generalization performance by switching from adam to sgd.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19:1236–1246.

OPAS (2019). Folha informativa - malária.

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., and Costa da Silva, E. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors*, 19(13):2969.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Sayyed, A. Q. M. S., Saha, D., Hossain, A. R., and Shahnaz, C. (2019). Effectiveness of convolutional and capsule network in malaria parasite detection. IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON).

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018). Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. International Conference on Medical image computing and computer-assisted intervention.

WHO (2019). World malaria report 2019. Technical report.