Evaluation of Post-hoc Explanations for Malaria Detection

V. B. Araújo[1], L. B. Marinho[2]

Universidade Federal de Campina Grande, Brazil
vinicius.brandao.araujo@ccc.ufcg.edu.br
lbmarinho@dsc.ufcg.edu.br

Abstract. It has been advocated that post-hoc explanation techniques are crucial for increasing the trust in complex Machine Learning (ML) models. However, it is so far not well understood whether such explanation techniques are useful or easy for users to understand. In this work, we explore the extent to which SHAP's explanations, a state-of-the-art post-hoc explainer, help humans to make better decisions. In the malaria classification scenario, we have designed an experiment with 120 volunteers to understand whether humans, starting with zero knowledge about the classification mechanism, could replicate the complex ML classifier's performance after having access to the model explanations. Our results show that this is indeed the case, i.e., when presented with the ML model outcomes and the explanations, humans can improve their classification performance, indicating that they understood how the ML model makes its decisions.

CCS Concepts: • **Applied computing**;

Keywords: deep learning, explainability, explanation evaluation, shap

1. INTRODUÇÃO

A subárea da Inteligência Artificial (IA) conhecida como *Deep Neural Networks (DNNs)* tem impulsionado avanços surpreendentes, tanto na academia como na indústria, em domínios diversos. Como exemplos de aplicações habilitadas por essa tecnologia, podemos citar veículos autônomos, chatbots e diagnóstico automático de doenças em imagens médicas [LeCun et al. 2015].

Contudo, o uso de DNNs como modelos caixa-preta pode trazer riscos em cenários em que as decisões tomadas pelos modelos tem sérias consequências. Caruana et. al [Caruana et al. 2015] mencionam que um cenário onde modelos de Aprendizagem de Máquina (AM) são usados para ajudar na tomada de decisão sobre a internação ou não de pacientes em UTIs, é preferível não usar modelos muito complexos (e.g., DNNs). A razão para essa observação é que modelos de AM muito complexos não são facilmente interpretáveis e por isso são pouco transparentes.

A capacidade de fornecer explicações sobre as saídas de modelos de AM complexos é crucial para tornar o modelo subjacente mais transparente e confiável. Retomando o exemplo anterior, é importante que médicos e pacientes entendam as razões pelas quais o modelo está sugerindo ou não internar a internação em uma UTI. Além de aumentar a confiabilidade no modelo, as explicações ajudam a identificar possíveis erros e idiossincrasias.

Normalmente há um trade-off entre interpretabilidade e complexidade, i.e., quanto mais simples um modelo, mais interpretável ele é. Já para modelos complexos, ocorre o oposto. O problema é que modelos muito simples, tais como modelos aditivos lineares ou árvores de decisão, embora interpretáveis, não alcançam bons resultados em problemas complexos. Por outro lado, modelos complexos alcançam bons resultados mas não são interpretáveis. Para resolver esse problema, técnicas

conhecidas como explicações *post hoc* tem surgido. Essas técnicas se referem ao conjunto de abordagens que visam explicar as saídas de qualquer modelo de AM caixa-preta [Guidotti et al. 2018]. Ribeiro et al. [2016], por exemplo, propõe o LIME (Local Interpretable Model-agnostic Explanations), uma das primeiras técnicas *post hoc* para a explicação das saídas de qualquer modelo de AM. A ideia é construir um modelo local e interpretável que imite o modelo complexo ao redor da instância sendo classificada. Assim, o modelo local pode ser usado para realizar explicações de previsões individuais.

Embora a área de interpretabilidade em AM tenha avançado bastante em anos recentes, ainda há poucos estudos sobre o impacto das explicações na compreensão das pessoas acerca do modelo de AM responsável pelas previsões. Para contribuir nessa direção, propomos um experimento com humanos no domínio de classificação de imagens médicas. A pergunta de pesquisa que guia este trabalho é a seguinte:

—Pessoas expostas à explicações *post-hoc* de modelos de AM complexos conseguem tomar decisões melhores do que as que não são expostas às explicações?

Para responder essa pergunta, realizamos um experimento com 120 voluntários recrutados entre os estudantes de exatas na Universidade Federal de Campina Grande (UFCG)[1]. Para esse experimento, escolhemos o problema de detecção de malária em imagens de esfregaço de sangue (ver mais detalhes nas Seções 2 e 5). Escolhemos o modelo VGG-19 [Simonyan and Zisserman 2014], uma DNN que mostrou bons resultados para esse problema na literatura revisada. Como explicador *post-hoc*, usamos a técnica estado-da-arte conhecida como SHAP [Lundberg and Lee 2017].

Inicialmente, solicitamos aos voluntários que classificassem imagens como infectadas ou não com malária, sem nenhum conhecimento prévio acerca do método de classificação. Em etapas posteriores, são apresentadas aos sujeitos do experimento imagens classificadas pelo VGG19, com e sem as explicações do SHAP. A partir daí, os voluntários são solicitados a classificarem novas imagens de esfregaço de sangue. O que queremos medir com esse experimento, em essência, é o ganho de conhecimento dos voluntários a partir das saídas do modelo com e sem explicações *post-hoc*. Quanto mais parecidas forem as classificações subsequentes dos voluntários no experimento com às do modelo, mais evidências temos de que os voluntários compreenderam o racional da inferência do modelo subjacente.

Os resultados, de forma geral, apontam que quando expostos às explicações, os usuários tendem a compreender melhor o mecanismo de inferência do modelo caixa-preta subjacente e com isso realizar classificações mais precisas (i.e., tomarem melhores decisões).

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, introduziremos conceitos que são importantes para um melhor entendimento deste trabalho.

2.1 Malária

A detecção de malária consiste em um procedimento que envolve o exame intensivo do esfregaço de sangue, no qual a amostra é levada para um microscópio que amplia a imagem da célula em cem vezes. Após esse procedimento, especialistas contam visualmente quantas células vermelhas do sangue possuem o parasita.

A Biblioteca Nacional de Medicina (do inglês: *National Library of Medicine (NIH)* - [NLM]) dispõe de uma base de dados, cuidadosamente coletada e classificada de um determinado conjunto de imagens de esfregaços de sangue representando pessoas saudáveis e infectadas com malária. Pesquisadores do *NIH* desenvolveram um aplicativo móvel que roda em um smartphone Android conectado a um

[1] `https://portal.ufcg.edu.br/`

microscópio de luz convencional. As lâminas de esfregaço de sangue fino coradas com Giemsa de 150 doentes infectados com *P. falciparum* e 50 saudáveis foram recolhidas e fotografadas no Chittagong Medical College Hospital, Bangladesh. A câmera embutida do smartphone adquiriu imagens de células para cada campo microscópico de visão. As mesmas foram anotadas manualmente por um especialista que as classificava como infectadas ou não.

2.2 VGG-19

O modelo VGG-19 é uma DNN de 19 camadas treinada na tão conhecida base de dados ImageNet, para fins de classificação de imagens. Este modelo foi desenvolvido por Karen Simonyan e Andrew Zisserman [Simonyan and Zisserman 2014].

A arquitetura do modelo VGG-19 é composta por um total de 16 camadas usando 3 x 3 filtros de convolução, além de camadas de *max pooling* para *downsampling* e um total de duas camadas ocultas de 4.096 neurônios totalmente conectadas em cada camada, seguidas por uma camada densa de 1.000 neurônios, em que cada neurônio representa uma das categorias de imagem no banco de dados ImageNet [Deng et al. 2009]. Em seu artigo, Dipanjan [Sarkar 2019] demonstra que essa DNN consegue detectar malária em imagens de esfregaço de sangue com uma alta taxa de acurácia.

2.3 SHAP

SHAP(*Shapley Additive Explanations*) [Lundberg and Lee 2017] é uma abordagem para explicar a saída de qualquer modelo de aprendizado de máquina. O SHAP é um explicador *post-hoc* que usa a teoria dos jogos para gerar explicações de qualquee modelo de AM.

O SHAP atribui um valor de importância especifico para cada um dos atributos usados no modelo e esses valores são determinados de acordo com a contribuição do atributo para a predição [Trumbelj and Kononenko 2013].

3. TRABALHOS RELACIONADOS

A avaliação da explicação de modelos de AM, como as geradas pelo SHAP, é fundamental para a aplicação efetiva dessas técnicas [Gilpin et al. 2018]. Velez e Kim [Doshi-Velez and Kim 2017] discutem a importância da avaliação das explicações fornecidas como forma de aumentar a confiança e usabilidade dessas ferramentas.

O artigo que introduz o SHAP [Lundberg and Lee 2017] realiza um experimento com usuários comparando as saídas fornecidas de diferentes explicadores. Sayres et. al. [Sayres et al. 2019] realiza um processo de avaliação de explicações por meio de uma variação da técnica *post-hoc* denominada de *Gradiente Explain* onde o autor avaliou alguns cenários de explicação de um modelo capaz de identificar Retinopatia Diabética em imagens de córnea. Para avaliar essas explicações, os autores contaram com a participação de especialistas humanos.

Nosso trabalho difere dos apresentados acima em dois aspectos principais: ao contrário de [Lundberg and Lee 2017], estamos interessados em entender o impacto das explicações no entendimento das pessoas acerca do mecanismo decisório de métodos complexos de AM; e diferente de [Sayres et al. 2019], contamos com pessoas leigas acerca do problema de classificação escolhido.

4. BASE DE DADOS

A base de dados utilizada nesse trabalho é constituída de 27.558 imagens de células com presença ou ausência do parasita de malária. Para o experimento, foram coletadas doze imagens dessa base de maneira aleatória, divididas igualmente entre infectadas e não infectadas. Esse conjunto de imagens pode ser visto na Figura 1.

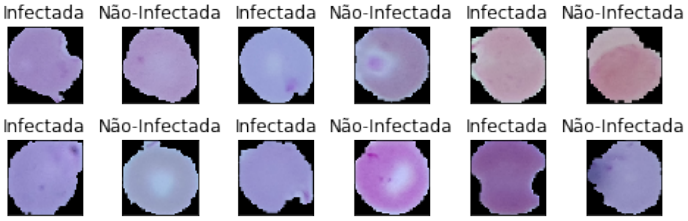


Fig. 1. Base de dados selecionada para o Experimento

5. METODOLOGIA

Nessa seção apresentamos os detalhes do experimento.

5.1 Configuração do Experimento

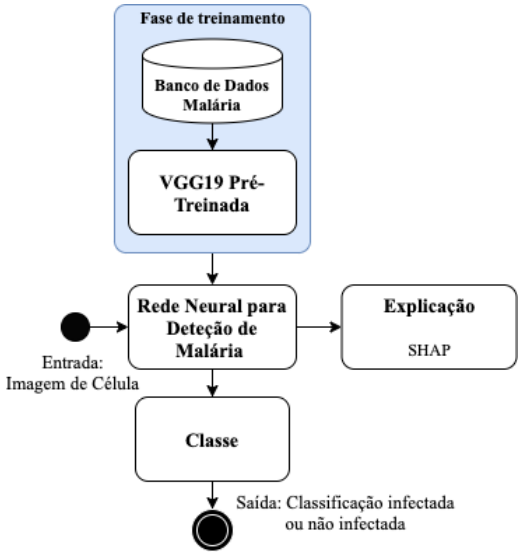


Fig. 2. Representação do desenvolvimento do modelo e explicação.

Como mostrado na Figura 2, o experimento consiste em treinamento do modelo e realização de predições/explicações para cada instância de entrada.

O treinamento do modelo foi inspirado no trabalho de Dipanjan [Sarkar 2019], que usa *Transfer Learning*. *Transfer Learning* consiste na reutilização de um modelo pré-treinado em um problema similar ao problema-alvo e reutilizado no problema-alvo treinando-se somente uma parte da rede. Nesse sentido, utilizamos o modelo VGG-19 pré-treinado na base de dados ImageNet [Deng et al. 2009].

Aplicamos o ajuste fino ao modelo VGG-19, onde liberamos somente os dois últimos blocos (Bloco 4 e Bloco 5) para treinamento.

Com o modelo definido, particionamos o conjunto de dados da malária em 63% para treino, 30% teste e 7% validação; utilizamos 25 épocas no treinamento e, como métricas de avaliação do modelo, utilizamos a acurácia.

Como podemos ver na Figura 3, o modelo chega a 96.5% de acurácia e assim demostra ter bons resultados.

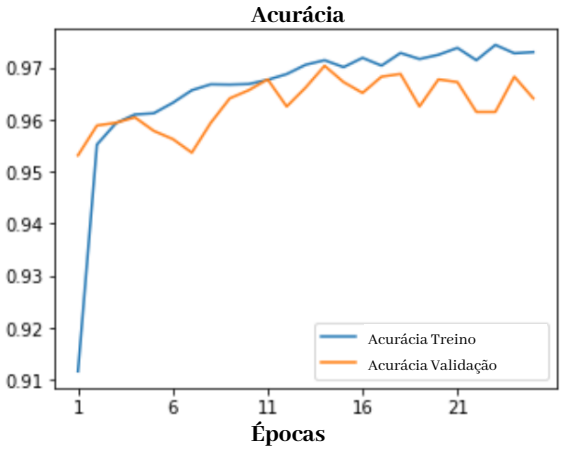
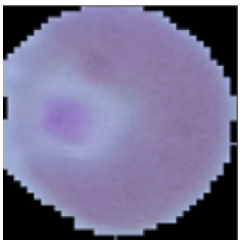
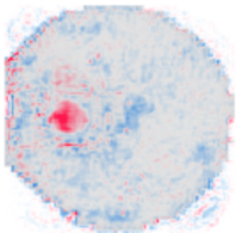


Fig. 3. Acurácia do modelo ao longo das Épocas.

Após a fase de treinamento, aplicamos o SHAP e geramos as explicações referentes às saídas do modelo. Na Figura 4, temos um exemplo de uma imagem do esfregaço de sangue e a saída fornecida pelo SHAP.



(a) Exemplo de entrada (figura coletada no microscópio)



(b) Exemplo de saída do SHAP em que a cor vermelha indica partes da imagem que tiveram influência para o modelo classificar essa célula como infectada.

Fig. 4. Exemplo de Entrada/Saida do *framework* SHAP

5.2 Experimento

Neste artigo, avaliamos as explicações do SHAP medindo se estas contribuem para o entendimento de pessoas acerca do mecanismo decisório do modelo de AM subjacente.

Para isso, projetamos um experimento com 120 voluntários que não tinham conhecimento sobre como é feita a classificação de células infectadas com o vírus da malária. Esses voluntários foram recrutados entre graduandos de cursos da área de Ciências Exatas da UFCG. Utilizando formulários para coleta dos dados, dividimos o experimento em três etapas como mostrado na Figura 5:

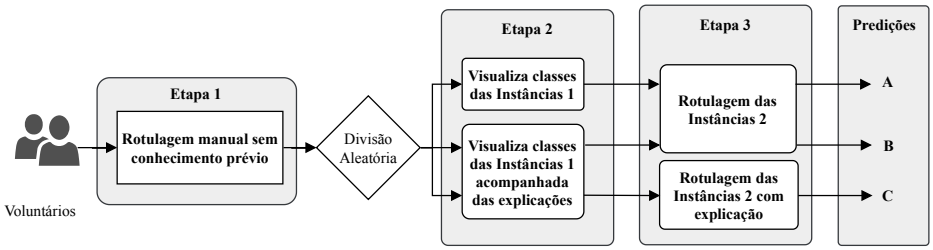


Fig. 5. Arquitetura do Experimento.

(1) Sem nenhum tipo de conhecimento prévio sobre o problema da detecção da malária em imagens de esfregaço de sangue, os voluntários fora solicitados a classificar seis imagens divididas igualmente entre as duas classes (Infectada/Não Infectada) escolhidas aleatoriamente na base de dados como mostra a Figura 6. Os resultados obtidos nessa etapa servem como controle, i.e., é esperado que as acurácias das classificações sejam próximas de aleatórias, já que os voluntários inicialmente não sabem como classificar as imagens.

(2) Nessa etapa, dividimos os participantes em três grupos doravante denominados A, B e C. Para cada uma das imagens da etapa anterior, os grupos receberam a classe predita pelo modelo VGG-19 juntamente com a confiança da predição. Adicionalmente, os grupos B e C receberam a explicação fornecida pelo SHAP. Na Figura 9, temos um exemplo de como essas informações foram repassados aos grupos.

(3) Nesta etapa, após serem apresentados às classificações (todos os grupos) e explicações do modelo (grupos B e C), solicitamos aos grupos para classificarem um novo conjunto de imagens 7. Porém ao grupo C, além das imagens, foram mostradas também as explicações do SHAP para a classificação do modelo VGG-19, como mostrado na Figura 8. Note que, nesse caso, somente as explicações são mostradas e não as classificações em si. Se os sujeitos do grupo C entenderam bem a lógica do classificador subjacente, isso deveria ajudá-los a classificar as imagens da forma como o classificador faria.

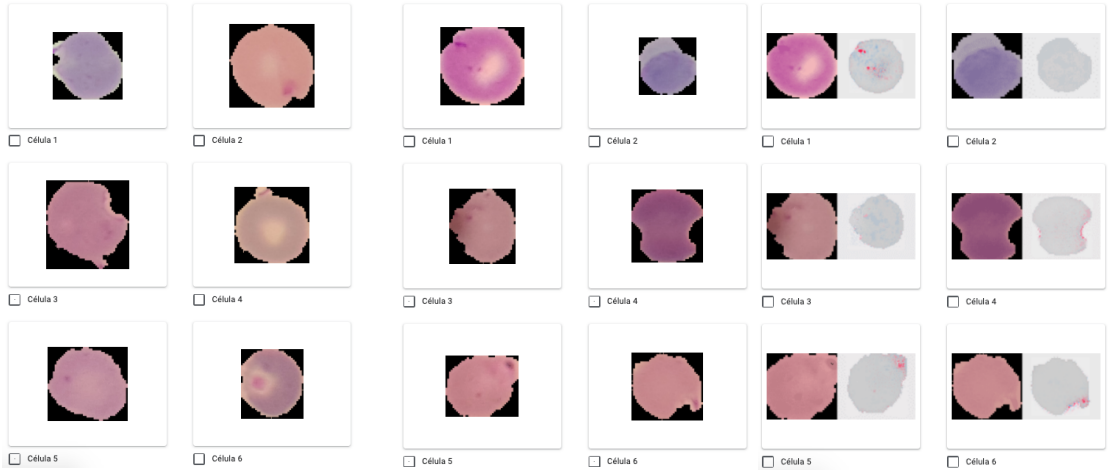
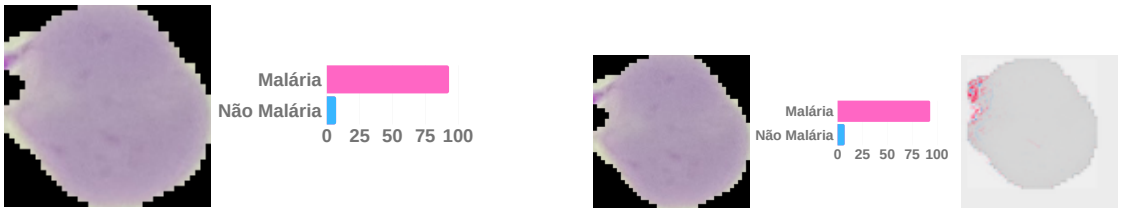


Fig. 6. Formulário aplicado na Etapa 1.

Fig. 7. Formulário aplicado na Etapa 3 para o grupos A e B.

Fig. 8. Formulário aplicado na Etapa 3 para o grupos C.



(a) Exemplo do resultado exibido ao grupo A

(b) Exemplo do resultado exibido aos grupos B e C

Fig. 9. Exemplos referentes a etapa 2 do experimento.

6. RESULTADOS

Para gerar os resultados referentes às classificações fornecidas pelos voluntários, usamos as métricas *precision* e *recall*. O intuito do uso dessas métricas é analisar a taxa de erros/acertos das classificações dos sujeitos em todas as etapas do experimento. A Figura 10 mostra os resultados.

Como podemos observar, na primeira etapa os resultados dos grupos possuem um comportamento parecido com valores baixos de *precision* e *recall*, o que é esperado já que os voluntários não sabem como realizar a classificação. É importante notar, entretanto, que as células infectadas apresentam um certo destaque por causa da coloração e portanto é razoável esperar que alguns voluntários identificaram esse padrão.

Quando comparamos os grupos na terceira etapa, notamos um padrão interessante. Primeiro, o grupo B tem um desempenho médio ligeiramente melhor que o grupo A, o que é esperado já que o grupo B teve acesso às explicações na etapa 2 do experimento. A diferença não parece ser significativa, entretanto. Já a diferença entre os grupos B e C, na etapa 3, é significativa. Isso mostra que as explicações, especialmente quando mostradas junto às imagens a serem classificados, ajudaram os voluntários do experimento a aumentarem a acurácia de classificação. Em outras palavras, os voluntários, do grupo C em especial, aprenderam a lógica de classificação do modelo de AM subjacente, que tem, por sua vez, alta acurácia de classificação (por volta de 96.5%).

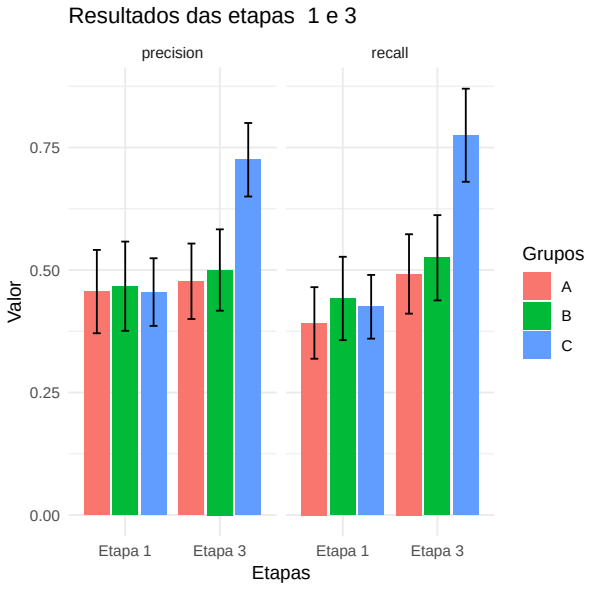


Fig. 10. Resultados *Precision* e *Recall*.

7. CONCLUSÃO

Nesse trabalho, investigamos a utilidade de explicações *post-hoc* de modelos de AM no contexto de classificação de imagens médicas. O protocolo de avaliação usado é novo haja vista a inexistência de protocolo similar na literatura revisada.

A ideia é primeiro solicitar que os voluntários realizem uma tarefa de classificação para a qual não foram treinados. Em etapas posteriores, são apresentados exemplos classificados por um modelo complexo de AM, com e sem as explicações. A partir daí, os voluntários são solicitados a classificarem novamente novos exemplos. Quanto mais parecidas forem as classificações subsequentes dos voluntários às classificações do modelo, mais evidência temos que os voluntários compreenderam o racional da inferência do modelo de AM subjacente.

Observamos que quando expostos à explicações, os voluntários conseguem de fato aumentar o nível de compreensão sobre o mecanismo decisório do modelo de AM. Isso implica que as explicações conseguem efetivamente informar humanos sobre a lógica de funcionamento de modelos complexos de AM.

Para trabalhos futuros, desejamos replicar esse experimento em outros domínios e com voluntários de diferentes perfis e níveis de conhecimento acerca do problema de classificação abordado.

REFERENCES

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. ACM, New York, NY, USA, pp. 1721–1730, 2015.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255, 2009.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning, 2017.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. pp. 80–89, 2018.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., and Giannotti, F. A survey of methods for explaining black box models, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature* 521 (7553): 436–444, 2015.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., pp. 4765–4774, 2017.

NLM. Malaria datasets.

Ribeiro, M. T., Singh, S., and Guestrin, C. "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, San Francisco, CA, USA, August 13-17, 2016*. pp. 1135–1144, 2016.

Sarkar, D. D. Detecting malaria with deep learning, 2019.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A. B., Corrado, G. S., Peng, L., and Webster, D. R. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 126 (4): 552 – 564, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR* vol. abs/1409.1556, 2014.

Trumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* vol. 41, pp. 647–665, 2013.