UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VINÍCIUS BRANDÃO ARAÚJO

AVALIAÇÃO DA UTILIDADE DAS EXPLICAÇÕES DOS MODELOS DE INTERPRETABILIDADE PÓS-HOC PARA A DETEÇÃO DE MALÁRIA

CAMPINA GRANDE - PB

2019

VINÍCIUS BRANDÃO ARAÚJO

AVALIAÇÃO DA UTILIDADE DAS EXPLICAÇÕES DOS MODELOS DE INTERPRETABILIDADE PÓS-HOC PARA A DETEÇÃO DE MALÁRIA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Leandro Balby Marinho.

CAMPINA GRANDE - PB

2019

A663a Araújo, Vinícius Brandão.
Avaliação da utilidade das explicações dos modelos de interpretabilidade pós-hoc para a detecção de malária. / Vinícius Brandão Araújo. – 2019.

9 f.

Orientador: Prof. Dr. Leandro Balby Marinho.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Método de interpretação pós-hoc. 2. Deep learning. 3. Classificação de malária. 4. Modelos de interpretabilidade pós-hoc. 5. Tecnologia aplicada à saúde. 6. Malária. I. Marinho, Leandro Balby. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

VINÍCIUS BRANDÃO ARAÚJO

AVALIAÇÃO DA UTILIDADE DAS EXPLICAÇÕES DOS MODELOS DE INTERPRETABILIDADE PÓS-HOC PARA A DETEÇÃO DE MALÁRIA

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Leandro Balby Marinho
Orientador – UASC/CEEI/UFCG

Professora Dra. Melina Mongiovi Cunha Lima Sabino
Examinadora – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de novembro de 2019.

CAMPINA GRANDE - PB

Avaliação da utilidade das explicações dos modelos de interpretabilidade pós-hoc para a deteção de malária

Trabalho de Conclusão de Curso

Vinicius Araujo
vinicius.brandao.araujo@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba

Leandro Marinho
lbmarinho@dsc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba

RESUMO

Explicações têm sido usadas como um meio de interpretar o raciocínio por trás das decisões de um modelo complexo (caixa preta) de Machine Learning. No entanto, uma crítica bem conhecida é que não há garantia de que as explicações produzidas sejam úteis ou fáceis de entender pelos usuários. Neste trabalho, exploramos a extensão da utilidade das explicações geradas por um método de interpretação pós-hoc estado-da-arte. Apresentamos um experimento com 120 usuários com o objetivo de avaliar explicações, baseado na simulação correta da saída do modelo de *Deep Learning* utilizado para a classificação de Malária. Utilizamos o método de explicação pós-hoc SHAP para gerar as explicações. Finalmente, mostramos que as explicações podem aumentar a compreensão de um modelo complexo pelos utilizadores.

1 INTRODUÇÃO

Impulsionado pela grande quantidade de dados, o uso de Inteligência Artificial vem crescendo consideravelmente no domínio médico. Esse forte crescimento pode ser associado ao conjunto de técnicas conhecidas coletivamente como Deep Learning, que promove algoritmos que permitem analisar e interpretar grandes conjuntos de dados com eficiência, e assim promover, de maneira ágil, diagnósticos precisos [9].

Contudo, alguns cuidados precisam ser tomados nesse domínio. Caruana et. al [3] em seu artigo nos mostra um exemplo de uma Rede Neural Artificial (RNA) treinada para prever quais pacientes com pneumonia deveriam ser internados em hospitais e quais deveriam ser tratados em regime ambulatorial. Os achados iniciais indicaram que a rede neural era mais precisa que métodos estatísticos clássicos. No entanto, após um extenso teste, verificou-se que a RNA tinha inferido que pacientes com pneumonia e asma detinham um risco clínico menor. Do ponto de vista médico, isso é contra-intuitivo. Foi então decidido abandonar o sistema de IA alegando que poderia ser muito perigoso usá-lo clinicamente. Sendo assim, a transparência do modelo torna-se fundamental para a descoberta de problemas cruciais e estratégias para evitá-los.

Sendo assim, a capacidade de fornecer explicações sobre as saídas de modelos complexos pode tornar o modelo subjacente mais transparente e mais confiável. Retomando o exemplo anterior, o modelo deveria ser capaz de explicar quais características são consideradas relevantes para classificar o paciente em relação ao tratamento sugerido e, assim, promover maior confiança para o usuário consumidor das predições, além de deixar mais evidente quando houve erros.

A necessidade de modelos interpretáveis leva ao surgimento de diversas técnicas, tais como *intrinsic* e *post hoc*. A técnica *intrinsic* está ligada a modelos considerados inerentemente transparentes como regressões lineares ou baseados em árvore em que os próprios pesos aprendidos servem como explicação, enquanto a *post hoc* refere-se ao conjunto de abordagens que visam interpretar um determinado modelo caixa preta treinado, treinando um modelo inerentemente interpretável sobre seus resultados [6]. Dessa forma, o desenvolvedor de ML é livre para usar qualquer modelo que queira, independente de quão complexo ele seja, uma vez que as explicações são extraídas do modelo interpretável treinado sobre ele. Marco Tulio [13] propõe uma das primeiras técnicas de *post hoc* demostrando uma aplicação no contexto real.

Existe um *trade-off* entre o desempenho do modelo e a interpretabilidade, considerando o domínio médico a resolução de problemas de alta complexidade com conjunto de dados extenso, o uso de modelos *intrinsic* acaba não sendo o suficiente para problemas desse domínio, desse modo, os desenvolvedores acabam optando por modelos de natureza não linear e de alta complexidade e assim o uso da técnica *post hoc* acaba sendo inevitável.

Um ponto importante sobre interpretabilidade de modelo está relacionado as explicações que são fornecidas, uma vez que precisam ser compreensíveis para os seres humanos [2]. Lipton [8] defende que interpretabilidade de modelos usando a técnica *post hoc* consegue chegar a um nível de compreensão humana. Um ponto importante sobre *post hoc* é a abordagem de explicação por meio de exemplos, onde essa se assemelha ao entendimento humano sobre explicações [7].

Considerando todos esses aspectos sobre interpretabilidade de modelos e as explicações fornecidas, o seguinte questionamento pode ser definido:

- Usuários conseguem tomar as melhores decisões ao serem expostos à explicação? Isto é, ao serem expostos às explicações, os usuários conseguem tomar decisões melhores do que quem não é exposto a essa informação?

Para responder essa pergunta, realizamos um experimento com 120 estudantes universitários da área de exatas, com o intuito de avaliar o impacto das explicações. Para esse experimento, escolhemos

o problema de detecção de malária na qual é notável a facilidade de classificação por usuários sem conhecimento prévio em distinguir entre uma célula infectada e não-infectada. Depois, escolhemos uma Rede Neural VGG-19 considerando a extensa base de dados na qual foi treinada, essa rede atua como um bom extractor de características para novas imagens sendo adequadas para problemas de visão computacional, tal como a detecção de vírus da malária, porém, essa Rede Neural é considerada caixa preta, por isso, aplicamos o explicador estado-da-arte SHAP em um determinado conjunto de dados utilizados para o experimento.

O experimento consiste em três etapas, onde na **primeira etapa**, todos os usuários foram submetidos a um teste de validação, sem possuir conhecimento prévio algum foi solicitado que classificassem imagens de células como infectada com o vírus da malária. Após esse processo, dividimos os usuários em três grupos iguais denominados de *A*, *B* e *C* e aplicamos a **segunda etapa** em que o grupo *A* obteve o gabarito da etapa anterior, já *B* e *C* obtiveram o gabarito e as explicações. Por fim, na **terceira etapa** os usuários foram apresentados a um novo grupo de imagens, foi solicitado que cada grupo classificassem as novas imagens como infectadas, porém o grupo *C* agora estava auxiliado de explicações, nessa etapa o intuito foi levantar dados para responder aos seguintes pontos:

- **PQ1:** O usuário que tem acesso à explicação do modelo consegue ter mais precisão na classificação de novos exemplos em relação aqueles que não tem acesso a essa informação?
- **PQ2:** Auxiliados por explicações, os voluntários conseguem tomar melhores decisões?

Os resultados referentes a etapa três do experimento sugerem com um intervalo de confiança de 95%, que os usuários expostos às classes preditas pelo modelo, não conseguem generalizar as classificações tão bem quanto pessoas que recebem explicações como auxílio, ou seja, os resultados apontaram que explicações trazem uma maior confiança aos modelos de inteligência artificial.

Esse documento está organizado como se segue: na Seção 2 apresentamos a Fundamentação Teórica, na Seção 3 os Trabalhos Relacionados, na Seção 4 a Metodologia de desenvolvimento do experimento, na Seção 5 a Base de Dados, na Seção 6 os Resultados obtidos e, por fim, na Seção 7 a Conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Malária

A malária é uma doença infecciosa, transmitida por mosquitos, causada por parasitas do gênero *Plasmodium*. Esses parasitas são transmitidos pelas picadas de mosquitos *Anopheles*. Ao picar uma pessoa, os parasitas transportados pelo mosquito ficarão no seu sangue e começarão a destruir as hemácias que transportam oxigênio (glóbulos vermelhos). Normalmente, os primeiros sintomas da malária são semelhantes aos da gripe.

De acordo com a Organização Mundial de Saúde (OMS) [1], quase metade da população mundial está em risco de contrair malária no próximo ano, e existem mais de 200 milhões de pessoas infectadas e aproximadamente 400.000 mortes devido à malária todos os anos. Isto proporciona ainda mais motivação para tornar a detecção e o diagnóstico da malária rápido, confiável e eficaz.

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. Assim, observamos que a detecção é um trabalho árduo, manual que pode potencialmente ser automatizado com Deep Learning [12].

A Biblioteca Nacional de Medicina (do inglês: *National Library of Medicine* - NLM) [11] dispõe de uma base de dados, cuidadosamente coletada e classificada de um determinado um conjunto de imagens de esfregaços de sangue saudáveis e infectados. Os pesquisadores desenvolveram um aplicativo móvel que roda em um smartphone Android padrão conectado a um 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 classificava como infectadas ou não infectadas.

2.2 VGG-19

O modelo VGG-19 é uma rede de aprendizagem profunda de 19 camadas (convolucionais e totalmente conectadas) construída no banco de dados ImageNet, para fins de reconhecimento e classificação de imagens. Este modelo foi desenvolvido por Karen Simonyan e Andrew Zisserman [16].

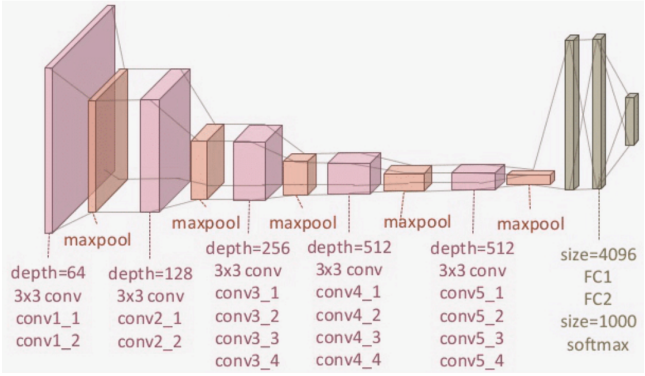


Figura 1: Arquitetura do modelo VGG-19.

Como podemos vê na Figura 1, arquitetura do modelo VGG-19 é composta por um total de 16 camadas de convolução usando 3 x 3 filtros de convolução, além de camadas máximas de *pooling* para *downsampling* e um total de duas camadas ocultas de 4096 unidades totalmente conectadas em cada camada, seguidas por uma camada densa de 1000 unidades, em que cada unidade representa uma das categorias de imagem no banco de dados ImageNet [4].

2.3 SHAP

SHAP(*Shapley Additive Explanations*) [10] é uma abordagem unificada para explicar a saída de qualquer modelo de aprendizado de máquina. O SHAP conecta a teoria dos jogos a explicações locais, unindo vários métodos anteriores e representando o único método de atribuição de recursos aditivos consistente e preciso localmente.

O SHAP atribui a cada recurso um valor de importância para uma previsão específica. Seus novos componentes incluem: a identificação de uma nova classe de medidas de importância de características aditivas e resultados teóricos mostrando que há uma solução única nessa classe com um conjunto de propriedades desejáveis. Normalmente, os valores SHAP tentam explicar a saída de um modelo (função) como uma soma dos efeitos de cada recurso sendo introduzido em uma expectativa condicional. É importante ressaltar que, para funções não lineares, importa a ordem em que os recursos são introduzidos. Os valores SHAP resultam da média de todos os pedidos possíveis. Provas da teoria dos jogos mostram que esta é a única abordagem consistente possível.

Uma maneira intuitiva de entender o valor de *Shapley* [17] é o seguinte: os valores dos recursos entram em uma sala em ordem aleatória. Todos os valores dos recursos na sala participam do jogo (= contribuem para a previsão).

Os valores de *Shapley* nos dizem como distribuir de forma justa a predição entre os recursos. Por exemplo, para explicar uma imagem, os pixels podem ser agrupados em super pixels e a previsão distribuída entre eles.

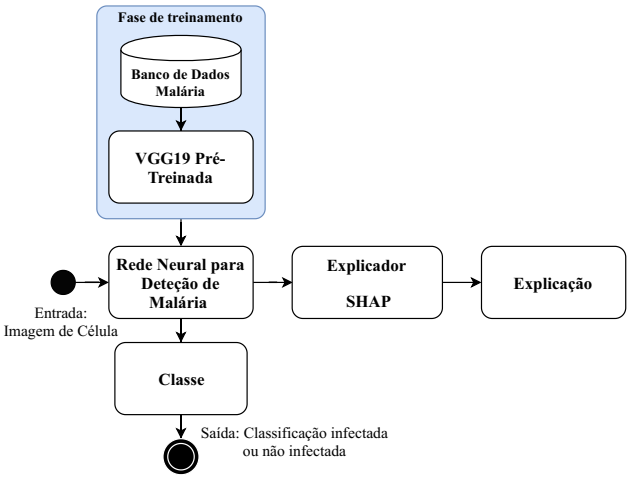


Figura 2: Representação do desenvolvimento modelo e explicação.

3 TRABALHOS RELACIONADOS

O processo de avaliação de explicações é uma abordagem defendida pela comunidade, no próprio artigo do SHAP [10] observamos que os autores realizaram um processo de validação com usuários, avaliando os resultados das explicações fornecidas em comparativo com outros *frameworks*.

Sayres et. al. [15] realizaram um experimento abordando a técnica de *Gradiente Explain* no contexto de retinopatia diabética, e aplicando cenários de avaliação de impacto de explicações em diagnósticos.

Velez e Kim [5] relatam em seu trabalho da importância da avaliação de explicações, trazendo pontos relevantes de como a qualidade de uma explicação pode impactar no uso de modelos transparentes, nesse mesmo trabalho, é exposto cenários de experimento necessários para a validação de explicações.

Amina Adadi et. al [2], nos mostra a defasagem de trabalhos que envolvem alocação de humanos como avaliadores de explicação, mostrando isso pode trazer consequências para colocar essas novas técnicas em produção no contexto real.

Esses trabalhos nos mostram da importância de um processo de avaliação se as explicações fornecidas conseguem atender com o seu objetivo final de tornar os modelos interpretáveis, nesse trabalho propomos um complemento com o estudo de caso da malária e uma abordagem de experimento com humanos como avaliadores de explicações.

4 METODOLOGIA

4.1 Background do Experimento

Como mostrado na Figura 2 o *background* do experimento consiste em fase de treinamento e desenvolvimento do modelo utilizado, aplicação do *framework* SHAP e geração das explicações e classe predita de acordo com o exemplo de entrada.

O treinamento do modelo foi inspirado no trabalho de Dipanjan [14], desse modo, para o treinamento do modelo utilizamos a técnica de *Transfer Learning* que consiste na reutilização de um modelo pré-treinado em um novo problema. Isto é, usar uma rede neural treinada em outro conjunto de dados, geralmente maior, para resolver um novo problema. A partir disso, utilizamos o modelo de *Deep Learning* pré-treinado VGG-19 na base de dados ImageNet [4].

Retiramos as últimas três camadas, pois usamos nossas próprias camadas densas totalmente conectadas para prever a malária. Aplicamos o ajuste fino ao modelo VGG, onde liberamos os dois últimos blocos (Bloco 4 e Bloco 5) para que seus pesos sejam atualizados em cada iteração (por lote de dados) enquanto treinamos o modelo.

Com o modelo definido, particionamos o conjunto de dados da malária em 63% para treino, 30% em teste e 7% validação, e utilizamos 25 épocas no treinamento e como métricas de avaliação do modelo utilizamos acurácia. Como podemos ver na Figura 3 o modelo chega a 96.5% de acurácia e assim demostra ter bons resultados.

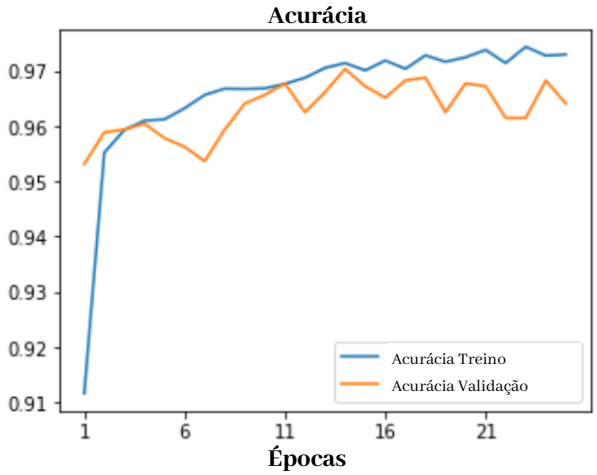
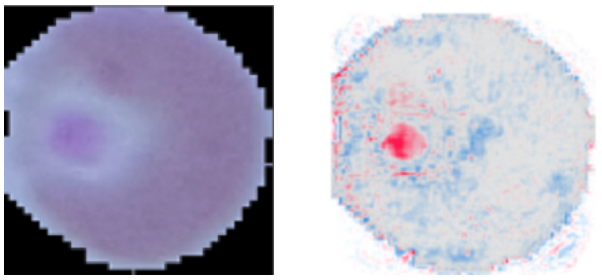


Figura 3: Acurácia do modelo ao longo das Épocas.

Após a fase de treinamento, aplicamos o *framework* SHAP e geramos as explicações referentes às saídas do modelo. Na Figura 4 temos um exemplo de entrada e de saída.



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

(b) Exemplo de saída do SHAP em que o vermelho indica partes da imagem que tiveram influência para o modelo classificar essa célula como infectada pelo vírus da malária.

Figura 4: Exemplo de Entrada/Saida do *framework* SHAP

4.2 Experimento

Para responder os questionamentos levantados nesse trabalho, aplicamos um experimento com 120 voluntários que não tinham conhecimento sobre como é feita a classificação de células infectadas com o vírus da malária. Utilizando formulários para coleta dos dados dividimos o experimento em três etapas como mostrado na Figura 5, são elas:

(1) Sem nenhum tipo de conhecimento prévio, todos os voluntários tiveram que rotular dentre 6 imagens quais eles julgavam estar infectada com malária. Os resultados obtidos nessa etapa devem servir como controle.
(2) Nessa etapa, os participantes passam pelo processo de obter conhecimento sobre a classificação de células infectadas, para isso, dividimos os participantes em três grupos iguais denominados como *A*, *B* e *C*. Os grupos receberam o resultado real do conjunto classificado anteriormente, porém o grupo *A* teve acesso ao gráfico de porcentagem de classificação do modelo para os exemplos, enquanto os grupos *B* e *C* receberam a resposta do modelo junto com a explicação fornecida pelo *framework* SHAP. Na Figura 6 temos um exemplo de como o gabarito foi fornecido para os grupos.
(3) Colocamos os grupos para classificar um novo conjunto, porém o grupo *C* tinha o auxílio das explicações para tomada de decisão.

O intuito da última etapa é criar um cenário de Simulação/Previsão [5], onde, considerando que na etapa 2 os usuários adquiriram conhecimento de como o modelo se comporta, isto nos possibilitará uma avaliação caso a explicação consiga tornar o modelo transparente o suficiente para que o usuário simule as possíveis saídas do mesmo.

5 BASE DE DADOS

A base de dados utilizada nesse trabalho é constituída de 27.558 imagens das células com presença ou ausência do parasita de malária de maneira igualitária para ambas as classes, na Figura 7 vemos alguns exemplos das imagens de células da base.

Figura 7: Imagens da Base de Dados.

Para o experimento, foram coletadas 12 imagens dessa base de maneira aleatória, divididas igualmente entre infectadas e não infectadas, esse conjunto utilizado pode ser visto na Figura 8.

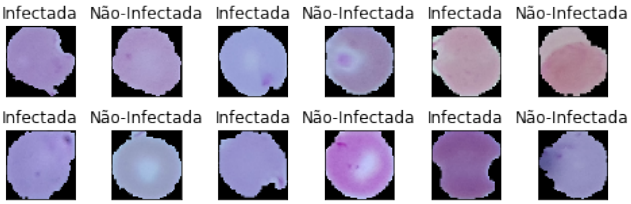


Figura 8: Base de dados selecionada para o Experimento

6 RESULTADOS

Para medir os resultados referente as classificações fornecidas pelos voluntários usamos às métricas *precision* e *recall*. O intuito do uso dessas métricas é analisar a taxa de acertos dos usuários perante a classificação e também a quantidade de erros no total, considerando todas as classificações dos usuários. Desse modo, obtivemos os resultados demostrados na Figura 9.

Como podemos observar, na primeira etapa os resultados dos grupos possuem um comportamento parecido e assim supomos que os voluntários classificaram as células de maneira aleatória.

Tendo como referência as perguntas levantadas nesse trabalho, para responder a PQ1 usamos o grupo *A* e *B* considerando que o cenário do grupo *A* refere-se ao conhecimento apenas da classe predita e do grupo *B* a explicação, um comparativo entre esses dois grupos considerando *precision* não temos uma diferença significativa entre eles em relação à terceira etapa e também um comparativo individual em relação primeira e terceira etapa do experimento não temos uma diferença considerável em relação ao ganho de conhecimento por partes dos voluntários.

Considerando a PQ2, temos o grupo *C* aplicado ao cenário de auxílio de explicações no processo de classificação da terceira etapa, tendo como referência a métrica *precision* temos um acerto significativo perante às células infectadas, porém o *recall* nos mostra que existiram casos que mesmo com a explicação auxiliando o usuário acaba errando na hora da classificação causando falsos-positivos.

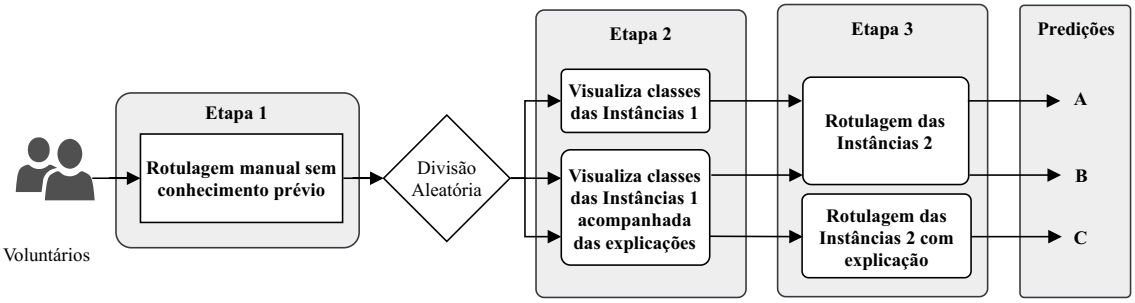
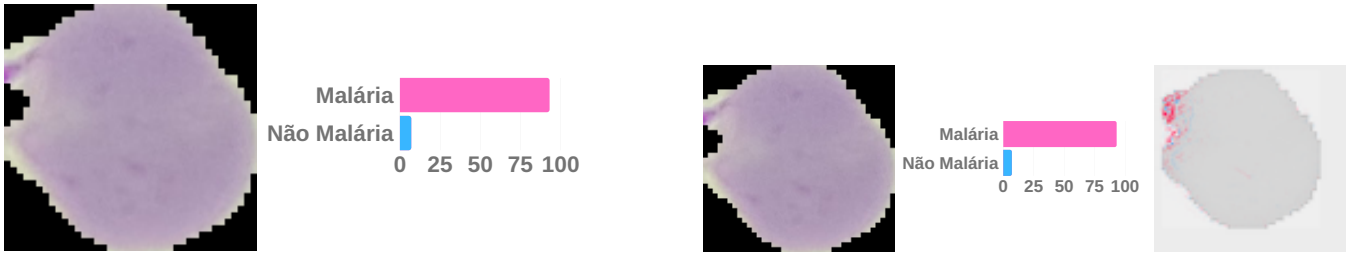


Figura 5: Arquitetura do Experimento.



(a) Exemplo do resultado exibido ao grupo *A*

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

Figura 6: Exemplos referentes a etapa 2 do experimento.

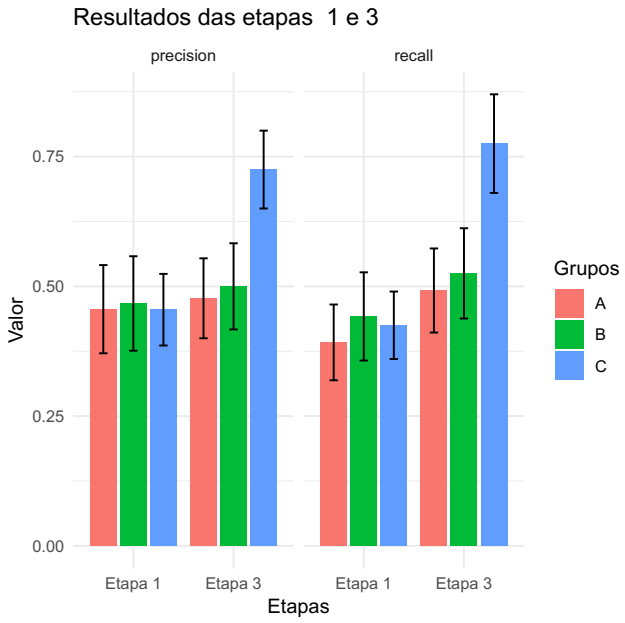


Figura 9: Resultados *Precision* e *Recall*.

Observando a terceira etapa do experimento, temos as células 4, 5 e 6 como infectadas com o vírus da malária, desse modo, observamos no gráfico da Figura 10 que as células 1 e 3 causaram para o grupo *C* os falsos-positivos, nas quais se refere às células classificadas de maneira errada como infectadas pelo vírus da malária pelos participantes deste grupo. Tomando como referência a célula 1, temos que o grupo *C* teve um maior número de participantes que classificaram de maneira errada em comparativo com os demais grupos.

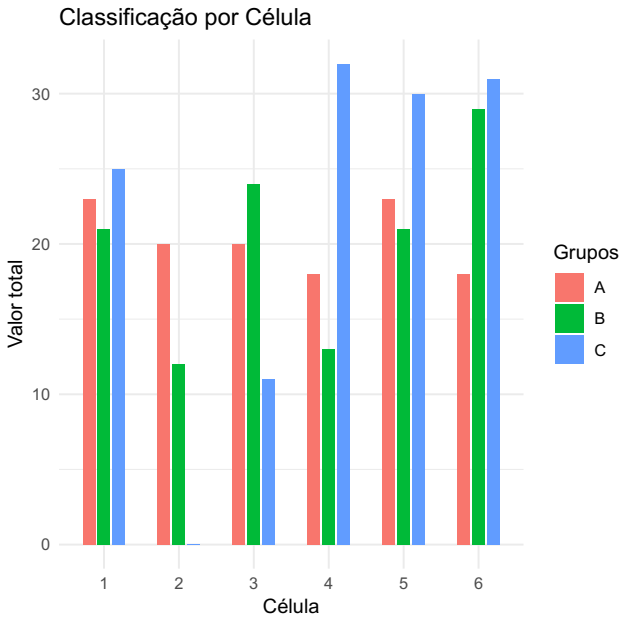


Figura 10: Classificação de células por usuário na etapa 3.

Retomando os casos de falsos-positivos encontrados em relação ao grupo *C*, podemos observar na Figura 11 que o grupo recebeu essas explicações na classificação da terceira etapa, essas explicações têm a presença de pontos vermelhos no quais pode ter influenciado alguns participantes a classificar as células com a presença do vírus da malária. A presença dos pontos vermelhos na explicação pode ser devido ao fato de que o modelo fornece uma aproximação percentual de quanto porcento aquele exemplo pode ter de presença do vírus, e assim, como a técnica se baseia no modelo, ela pode replicar exatamente a pequena porcentagem de classificação e rotular qual parte influenciou. Para esses exemplos o modelo classificou com a presença do vírus da malária com 10%, no quais essa porcentagem influenciou os pontos vermelhos da explicação.

(a) Célula 1 (b) Célula 3

Figura 11: Células acompanhadas das explicações.

Observamos que para alguns usuários a presença de explicação acabou influenciando decisões erradas. Esse problema envolvendo falsos-positivos em relação às explicações também foi apresentado no artigo de Sayres et. al. [15].

7 CONCLUSÃO

Nesse trabalho, investigamos a utilidade de explicações no contexto de avaliação dessas por humanos. Para isso, propusemos uma abordagem de experimento visando avaliar se as explicações estão coerentes com o principal objetivo, que é ser entendida por humanos, levando em consideração que essas pessoas não têm conhecimento em determinado assunto, assim como os modelos, e fazem com que as mesmas possam compreender o funcionamento, é algo extremamente relevante para as explicações.

Observamos com os resultados do experimento referentes ao grupo *C* que as explicações conseguem aumentar o nível de compreensão de humanos para determinados problemas e assim promover mais transparência para os usuários finais que irão consumir as informações providas pelos modelos. Entretanto, a pesquisa mostrou um problema que as explicações podem causar e que já foi abordado em outros trabalhos: os falsos-positivos. Em se tratando de domínio médico, isso pode ter graves consequências.

Para trabalhos futuros, desejamos aplicar esse experimento com especialistas na área, com objetivo de verificar se, caso o modelo possua uma eficiência semelhante ao especialista, então boas explicações levariam humanos leigos no assunto adquirir o mesmo nível de conhecimento, isso por meio de um comparativo de resultados da classificação da terceira etapa.

AGRADECIMENTOS

Agradeço aos meus familiares, professores e colegas que me ajudaram imensamente durante todo o período em que estive na graduação. Agradeço ao professor Leandro pela orientação e paciência. Por fim, agradeço a Ítalo Pontes, Allan Sales e Ricardo Oliveira por toda paciência em ajudar na correção desse trabalho.

REFERÊNCIAS

[1] [n. d.]. Malaria. https://www.who.int/news-room/facts-in-pictures/detail/malaria

[2] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. 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, 1721–1730. https://doi.org/10.1145/2783258.2788613

[4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[5] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models. arXiv:cs.CY/1802.01933

[7] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2280–2288. http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf

[8] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[9] Luiz Carlos Lobo. 2018. InteligÃartificial, o Futuro da Medicina e a EducaÃ§ÃMÃ. *Revista Brasileira de EducaÃ§ÃMÃ* 42 (09 2018), 3 – 8. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022018000300003&nrm=iso

[10] Scott M Lundberg and Su-In Lee. 2017. 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., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[11] National Library of Medicine. [n. d.]. Malaria Datasets. https://lhncbc.nlm.nih.gov/publication/pub9932

[12] Poostchi M Silamut K Hossain MA Maude RJ Jaeger S Thoma GR Rajaraman S, Antani SK. 2018. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. https://doi.org/10.7717/peerj.4568 (2018). http://arxiv.org/abs/1409.1556

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[14] Dipanjan (DJ) Sarkar. 2019. Detecting Malaria with Deep Learning. (2019).

[15] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* 126, 4 (2019), 552 – 564. https://doi.org/10.1016/j.ophtha.2018.11.016

[16] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[17] Erik trumbelj and Igor Kononenko. 2013. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41 (2013), 647–665.