SHAP-Driven Explicability of CNN-Based Computer-Aided Diagnosis for Malaria

Matheus Silva dos Santos[1], **Fagner Cunha**[1],
Rafael Giusti[1], **Juan Gabriel Colonna**[1]

[1]Instituto de Computação (IComp), Universidade Federal do Amazonas (UFAM),
Manaus, Amazonas (AM), Brasil.

{matheus.silva, fagner.cunha, rgiusti, juancolonna}@icomp.ufam.edu.br

Abstract. *Convolutional Neural Networks (CNN) hold great promise for medical image classification, but their clinical adoption depends on model reliability. In this study, we leverage Shapley Additive Explanations (SHAP) to explain a CNN-based Computer-Aided Diagnosis (CAD) system for malaria detection from microscopy images. We employ SHAP's Gradient Explainer to highlight key pixel regions driving the CAD model predictions. By providing transparent, pixel-level insights, this approach empowers healthcare professionals to understand system decisions and enhances trust in automated malaria diagnosis.*

Resumo. *Redes Neurais Convolucionais (CNNs) têm grande potencial para classificação de imagens médicas, mas sua adoção clínica depende da confiabilidade dos modelos. Neste estudo, utilizamos Shapley Additive Explanations (SHAP) para explicar um sistema de CAD (Computer-Aided Diagnosis) baseado em CNN para diagnóstico de malária em imagens de microscopia. Aplicamos o Gradient Explainer para destacar as regiões de pixel mais relevantes para as previsões do CAD. Ao oferecer explicações transparentes em nível de pixel, esta abordagem capacita profissionais de saúde a compreenderem as decisões do sistema e fortalece a confiança no diagnóstico automatizado de malária.*

1. Introdução

A malária é uma doença transmitida através da picada de fêmeas dos mosquitos do gênero *Anopheles*. Essa doença é comum em regiões tropicais e subtropicais do mundo, incluindo parte da América, Ásia e África, uma vez que a disseminação desses mosquitos está relacionada à altitude, clima e vegetação adequados [Guerra et al. 2008]. Os surtos desta doença estão diretamente ligados à pobreza e desastres naturais [Glowac et al. 2024], além de representarem um desafio significativo em um cenário de mudanças climáticas globais [Megersa and Luo 2025].

Essa doença é causada por parasitas protozoários do gênero *Plasmodium*, onde cinco tipos de protozoários são conhecidos por infectar humanos: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* e *P. knowlesi* [Shapiro et al. 2013]. Dentre eles, *P. falciparum* e *P. vivax* são os mais prevalentes, sendo o primeiro o mais letal, especialmente em crianças com baixa imunidade e mulheres grávidas [Das et al. 2022].

A malária é dividida em duas manifestações clínicas: não complicada e grave. O diagnóstico clínico acaba se tornando não confiável porque os sintomas da malária não

complicada são altamente inespecíficos e incluem febre, calafrios, dores no corpo, dor de cabeça, tosse e diarreia [Ashley et al. 2018]. As manifestações mais comuns da malária grave são malária cerebral, lesão pulmonar aguda que pode evoluir para síndrome do desconforto respiratório agudo (até 25% dos casos), lesão renal aguda que normalmente se manifesta como necrose tubular aguda e acidose [Das et al. 2022].

O método padrão para detectar a malária é a microscopia óptica de esfregaços sanguíneos corados. Esse método é realizado com filmes de glóbulos vermelhos, que podem ser espessos ou finos, e que permitem a especiação e quantificação dos parasitas [Ashley et al. 2018]. Para isso, uma gota de sangue é espalhada sobre uma lâmina de vidro e misturada com coloração de *Giemsa*, facilitando a visualização dos parasitas dentro dos glóbulos vermelhos sob o microscópio, como mostra a Figura 1a. Embora também existam testes de diagnóstico rápido para realizar o diagnóstico inicial da doença, a microscopia ainda é considerada o padrão ouro e o teste rápido é propenso a falsos negativos [Chiodini 2014, Yadav et al. 2024].

O método de detecção por microscopia de luz demanda muito tempo e treinamento humano [Moody 2002]. Com o avanço das técnicas de aprendizado de máquina profundo (DL–*Deep Learning*), podemos desenvolver modelos que realizam diagnóstico automatizado de malária. Nos modelos DL, as camadas ocultas extraem características das imagens automaticamente, eliminando a necessidade de fazer isso manualmente. Redes Neurais Convolucionais (CNNs) são amplamente utilizadas para tarefas de classificação de imagens por serem muito eficazes [Berezsky et al. 2024], mas o seu desempenho computacional pode variar dependendo do tamanho do modelo, do conjunto de dados utilizado para o treinamento e da complexidade das imagens.

Um ponto negativo das CNNs é que são modelos "caixa preta", pois mesmo um usuário treinado não consegue explicar o modelo [Guidotti et al. 2018]. Para contornar isso, técnicas de Inteligência Artificial explicável (XAI–*Explainable Artificial Intelligence*) podem ser empregadas. Uma delas é o SHAP (Shapley Additive Explanations), que fornece uma melhor ideia sobre como um modelo CNN toma decisões, mostrando quais pixels das imagens são os mais importantes.

Este artigo tem como objetivo investigar o método de explicação SHAP no contexto da detecção da malária em esfregaços de sangue. Em diagnósticos clínicos, é fundamental entender como um modelo "caixa preta" toma decisões, já que a precisão e a confiabilidade dos resultados são essenciais para auxiliar o diagnóstico médico. Para demonstrar a aplicação do SHAP, utilizamos uma arquitetura CNN de autoria própria para o diagnóstico de imagens de microscopia, a fim de avaliar o que o modelo está aprendendo e se está realizando corretamente sua tarefa. Por fim, exploramos a saída desse modelo para identificar possíveis erros na classificação de células como saudáveis ou infectadas pelo parasita. Especificamente, analisamos casos em que o modelo erra na classificação devido a possíveis ruídos de rotulagem ou erros de segmentação da base de dados.

O restante deste artigo está organizado da seguinte forma: na Seção 2, são apresentados os trabalhos relacionados; na Seção 3, descrevem-se a base de dados utilizada, o procedimento de pré-processamento das imagens e o método SHAP; na Seção 4, apresenta-se a CNN desenvolvida especificamente para este estudo; a Seção 5 apresenta e discute os resultados obtidos; por fim, as considerações finais são apresentadas na Seção 6.

2. Trabalhos Relacionados

O diagnóstico da malária por sistemas CAD (*Computer-Aided Diagnosis*) que utilizam imagens de microscopia e DL foi explorado em estudos anteriores, porém a aplicação de XAI para entender os resultados dos modelos ainda permanece pouco explorada.

O primeiro trabalho a utilizar um modelo CNN para classificar células infectadas e não infectadas em esfregaços sanguíneos foi proposto por [Liang et al. 2016]. Outro trabalho relevante foi conduzido por [Boit and Patil 2024], que apresentou o modelo híbrido EDRI, integrando múltiplas arquiteturas de DL para detectar malária em imagens de glóbulos vermelhos. No entanto, nenhum desses trabalhos utilizou XAI para explicar as decisões do modelo, o que torna difícil garantir que os modelos tenham aprendido corretamente características das imagens relacionadas aos protozoários da malária. É importante destacar que ambos os trabalhos utilizaram a mesma base de dados, que é considerada padrão para esse tipo de aplicação e foi adotada neste estudo.

Outro trabalho que foi base para este artigo é o de [Rajaraman et al. 2018a], que descreve a utilização de modelos CNN personalizados para a detecção de parasitas da malária em imagens de esfregaços sanguíneos. Os autores realizaram experimentos para avaliar o desempenho das redes CNN personalizadas e analisaram a relação entre as características das imagens e as saídas dos modelos. Além disso, eles propõem uma metodologia para a explicação das decisões do modelo, o que pode levar a uma melhor compreensão do comportamento das CNNs personalizadas na detecção de parasitas da malária.

A revisão realizada por [Eke and Shuib 2025] investiga o papel das técnicas de XAI na construção de confiança em sistemas de IA para a saúde. Os autores argumentam que o fato de esses modelos funcionarem como "caixa preta" representa uma barreira significativa à sua adoção, levantando preocupações sobre vieses e robustez. Concluem que a XAI é essencial para promover transparência e responsabilidade, viabilizando o desenvolvimento de sistemas de IA mais confiáveis para o setor.

No artigo de [Gaouar et al. 2025], é apresentado um novo modelo baseado em Stacked-LSTM que superou outras arquiteturas, como CNNs e Vision Transformers. Reconhecendo a importância da explicabilidade em aplicações de saúde, os autores aplicaram as técnicas de XAI Grad-CAM e LIME para visualizar e explicar as decisões dos modelos. Destacam, ainda, que é fundamental desenvolver sistemas de diagnóstico clínico que sejam não apenas precisos, mas também confiáveis e transparentes.

As evidências apresentadas pelos trabalhos relacionados destacam a crescente preocupação com a aplicação de modelos de Inteligência Artificial com explicabilidade (XAI) para o diagnóstico de malária usando aprendizado profundo. É crucial ressaltar que a explicação desses modelos é fundamental para garantir a confiabilidade das decisões tomadas pelo modelo e aumentar a transparência do processo de diagnóstico para os médicos. Nesse contexto, a principal contribuição deste artigo é abordar a explicabilidade em um modelo CNN para o diagnóstico de malária, fornecendo aos médicos e pesquisadores uma compreensão clara do processo de tomada de decisão do modelo e melhorando a confiabilidade das conclusões alcançadas.

3. Materiais e Métodos

Nesta seção apresentamos a base de dados utilizada, bem como o pré-processamento aplicado para a preparação dos dados. Além disso, será descrito o método SHAP utilizado para explicar o modelo CNN. Por fim, serão apresentadas as métricas utilizadas para avaliar o desempenho do modelo no diagnóstico da malária.

3.1. Base de Dados

Neste trabalho utilizamos a base de imagens NIH Malaria [Rajaraman et al. 2018b] para treinar um modelo CNN. Essa base está disponível publicamente no repositório do *National Institute of Health* (NIH) [NIH 2022]. As imagens foram coletadas ao aplicar um *smartphone* em um microscópio de luz convencional, resultando na captura de imagens similares com a Figura 1(a). O objetivo foi automatizar o diagnóstico da malária utilizando um aplicativo Android para reduzir a sobrecarga de trabalho dos microscopistas e aprimorar a precisão do diagnóstico [Poostchi et al. 2018]. Assim, foram coletados esfregaços finos de sangue de 150 pacientes infectados com a espécie de protozoário *P. falciparum* e mais 50 pacientes saudáveis. Durante a captura das imagens foi aplicada uma coloração *Giemsa* para aumentar a visibilidade dos parasitas nas células. Posteriormente, as imagens foram segmentadas e rotuladas manualmente por um especialista.

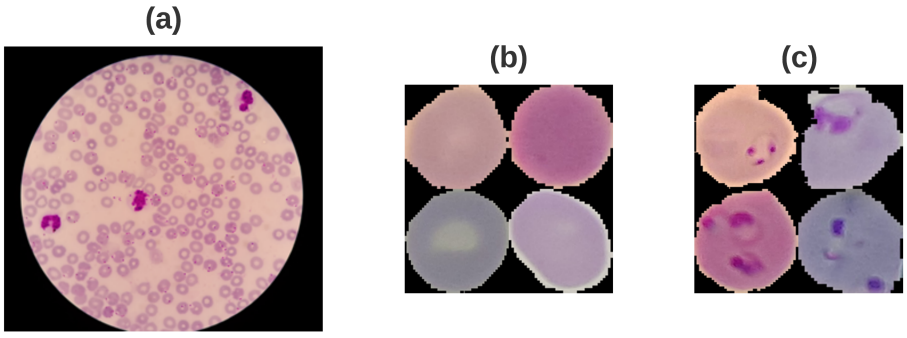

Figura 1. Exemplos de imagens da base de dados do NIH Malaria. Imagem do campo de visão do microscópio (a), células não infectadas (b) e células infectadas com o parasita *P. falciparum* (c).

As imagens maiores (Figura 1a) foram segmentadas em imagens menores, contendo unicamente uma célula por imagem [Ersoy et al. 2012], com auxílio de um algoritmo algoritmo de detecção e segmentação proposto por [Parvin et al. 2007]. Consequentemente, as imagens obtidas não possuem um tamanho padrão (altura e largura) porque os glóbulos vermelhos segmentados possuem tamanhos diferentes. A base obtida contém 27.558 imagens de células individuais distribuídas igualmente em imagens de células infectadas e imagens de células não infectadas. As imagens possuem mudanças na distribuição das cores devido a diferenças durante a coloração empregada no processo de captura do microscópio. As figuras 1b e 1c contêm amostras das células segmentadas.

3.2. Pré-processamento dos dados

Foi realizado um redimensionamento nas imagens segmentadas para o tamanho de 100x100 pixels com três camadas de cores RGB. Para treinar e avaliar o desempenho do modelo CNN, as 27.558 imagens foram distribuídas em três subconjuntos: treino, validação e

teste, sendo que 60% foram para o subconjunto de treinamento (16.534 imagens), 10% para validação (2.754 imagens) e 30% para teste (8.270 imagens). A reamostragem da base foi feita com estratificação de classes e todos os subconjuntos são balanceados.

Adicionalmente, utilizamos a técnica de aumento aleatório de dados por meio da biblioteca Keras [Keras 2020]. Essa técnica aumenta artificialmente a quantidade de dados, adicionando cópias levemente modificadas sem a necessidade de coletar novos dados de treinamento [Ali et al. 2021]. Inversões horizontais e verticais foram aplicadas ao subconjunto de treinamento para simular variações naturais das imagens, como mudanças de perspectiva ou orientação que podem ocorrer quando um especialista analisa as células com um microscópio. Essas mudanças aleatórias nas imagens ajudam também a aumentar a variabilidade dos dados e evitar um possível sobreajuste do modelo CNN.

3.3. Shapley Additive Explanations (SHAP)

O método de explicação SHAP calcula os valores Shapley da teoria dos jogos de coalizão [Shapley 1988]. O SHAP tem por objetivo prover uma explicação quantificável sobre as previsões de modelos CNN determinando a contribuição individual de cada pixel na previsão de uma amostra específica. O SHAP possui propriedades matemáticas únicas de explicadores do tipo "local", que permite entender a contribuição dos pixels para cada amostra da base de teste individualmente. A equação de explicação do SHAP é descrita como:

$$\mathrm{g}\left(z'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j, \tag{1}$$

onde g é o modelo de explicação, $z' \in \{0,1\}^M$ é o vetor de coalizão, onde significa a existência ou não de cada característica, M é o tamanho máximo do vetor de coalizão e $\phi_j \in \mathbb{R}$ corresponde à contribuição da característica j para a predição (valores de Shapley). O vetor de coalizão representa a contribuição de cada característica para a previsão de uma amostra específica. Ele é calculado usando a teoria dos valores Shapley, que fornece uma forma justa de distribuir o "pagamento" (que significa a previsão) entre as características. Uma entrada igual a 1 no vetor de coalizão significa que a característica está "presente" e igual a 0 que está "ausente" [Molnar 2022].

Uma análise mais rigorosa sobre as propriedades consideradas desejáveis (precisão local, falta e consistência) do SHAP é detalhada em [Lundberg and Lee 2017] e [Molnar 2022]. A biblioteca SHAP [SHAP 2020] oferece diversos tipos de explicadores para modelos de aprendizado de máquina. Neste artigo, o explicador utilizado foi o *Gradient Explainer*. Esse método serve para modelos diferenciáveis e estima os valores de Shapley calculando a expectativa dos gradientes do modelo em relação às entradas. Ele atribui eficientemente a contribuição de cada característica à predição ao integrar esses gradientes ao longo de trajetórias de referência. Na seção 5 apresentamos os resultados obtidos juntamente às visualizações que permitem entender o funcionamento do modelo treinado.

3.4. Métricas de Desempenho

A avaliação do desempenho de modelos é uma etapa crucial na construção de modelos de aprendizado de máquina. Uma das técnicas mais comuns para avaliar a eficácia de um modelo é o uso da matriz de confusão, que apresenta o desempenho do modelo em

termos de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN) [Sokolova et al. 2006]. Neste artigo, os elementos positivos são as células infectadas com o parasita da malária, enquanto os elementos negativos são as células saudáveis.

A precisão (P), a revocação (R), o F1-score (F1) e a acurácia (Acc) são métricas bastante utilizadas para avaliar o desempenho de um modelo . A precisão mede a proporção de previsões corretas da classe positiva em relação ao número total de amostras classificadas como positivas (Equação 2), enquanto a revocação mede a proporção de amostras positivas verdadeiras que foram corretamente classificadas pelo modelo (Equação 3). O F1-score é uma média harmônica entre a precisão e a revocação (Equação 4). Já a acurácia mede a proporção de previsões corretas feitas pelo modelo em relação ao número total de amostras (Equação 5).

$$\mathrm{P} = \frac{VP}{VP + FP} \quad (2) \qquad \mathrm{R} = \frac{VP}{VP + FN} \quad (3)$$

$$\mathrm{F1} = \frac{2PR}{P + R} \quad (4) \qquad \mathrm{Acc} = \frac{VP + VN}{VP + VN + FP + FN} \quad (5)$$

4. Arquitetura do Modelo CNN

A arquitetura CNN utilizada neste trabalho, de autoria própria, foi desenvolvida especificamente para se adequar à base de imagens descrita na Seção 3.1, considerando tanto a quantidade de instâncias disponíveis quanto as características das imagens, as quais não possuem alta diversidade de formas. Um modelo personalizado e mais simples oferece um cenário mais controlado para explorar o foco principal deste trabalho em relação à explicabilidade por meio do método SHAP, evitando assim complexidades adicionais que poderiam ser trazidas por arquiteturas mais profundas, tipicamente pré-treinadas em bases de dados massivas, como ResNet ou ViT. Portanto, embora o desenvolvimento de uma arquitetura otimizada para o problema não seja o escopo central, procuramos desenvolver um modelo com desempenho razoável, evitando *underfitting* ou *overfitting*.

Utilizamos imagens de células segmentadas com resolução de 100x100x3 pixels para o treinamento do nosso modelo CNN. Para extrair as características mais relevantes, cada imagem de entrada passa por quatro camadas de convolução 2D com kernel 3x3. Em seguida, os vetores de características (*embeddigns*) são passados para duas camadas densas totalmente conectadas e um último neurônio *sigmoid* realiza a classificação. A arquitetura do modelo pode ser visualizada na Figura 2.

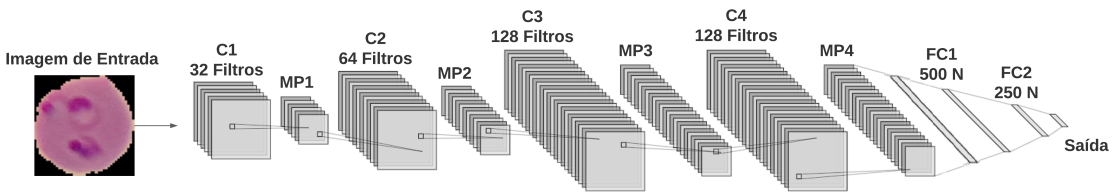

Figura 2. Arquitetura do nosso modelo CNN. As camadas de convolução são representadas pela sigla "C", as camadas de MaxPooling pela sigla "MP" e as camadas totalmente conectada pela sigla "FC" (*Fully Connected*).

A primeira camada de convolução gera 32 mapas de características, enquanto a segunda gera 64 e as terceira e última camadas geram 128. Após cada camada de convolução, existe uma camada de *pooling* máximo que é usada para reduzir a dimensionalidade dos mapas. Após a última camada de convolução, há uma camada de concatenação ("*flatten*"), para transformar os mapas de características em vetores unidimensionais, tornando-os compatíveis com as camadas totalmente conectadas. A primeira camada totalmente conectada contém 500 neurônios e a segunda 250. Foi utilizado o algoritmo de otimização ADAM e a função de perda *Binary Categorical Crossentropy*. Todas as camadas do modelo têm função de ativação ReLU exceto a última camada, responsável por prever se a célula está infectada ou não, que utiliza a função sigmoide. O tamanho do lote de dados (*batch size*) utilizado no treinamento foi de 512 imagens.

5. Resultados e Discussões

O modelo foi treinado por 20 épocas, porém o treinamento foi detido em 15 épocas para evitar sobreajuste (*overfitting*), de acordo com a comparação do desempenho do modelo entre os conjuntos de treino e de validação. O aumento de dados permitiu que o modelo melhorasse em até 2% em comparação com um experimento sem aumento de dados.

A Figura 3 apresenta as curvas de perda e acurácia do modelo durante as etapas de treinamento e validação, respectivamente. Notamos que as curvas se estabilizam durante o treinamento, indicando que o modelo está aprendendo os dados corretamente e não está apenas "decorando" os padrões do conjunto de treinamento.

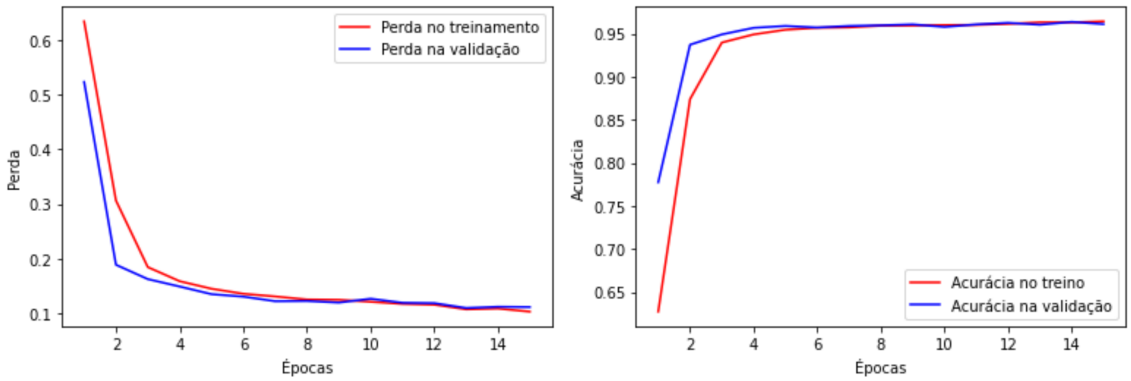

Figura 3. Evolução do treinamento do modelo CNN.

Na Figura 4, é apresentada a matriz de confusão nos dados de teste. Observe que o número de amostras é o mesmo para ambas as classes (4.135).

Classe verdadeira \ Classe prevista	Infectada	Saudável
Infectada	**3.940**	195
Saudável	147	**3.988**

Figura 4. Matriz de confusão no conjunto de teste.

Com base na matriz de confusão, calculamos as métricas de desempenho do modelo, que são apresentadas na Tabela 1. A precisão e a revocação são calculadas para cada classe e a média macro também é dada.

Tabela 1. Métricas no conjunto de teste (%).

	Precisão	Revocação	F1	Acurácia
Infectada	96,40	95,28	95,84	-
Saudável	95,34	96,44	95,89	-
Média	95,87	95,86	95,86	95,86

5.1. Explicando o modelo CNN

Após analisarmos o desempenho obtido pelo modelo, aplicamos o método *Gradient Explainer* para identificar os pixels importantes que levam o modelo a classificar as imagens entre infectada ou saudável. A Figura 5 apresenta os resultados para alguns exemplos selecionados aleatoriamente do conjunto de teste, mas contemplando os quatro casos de classificação (VP, VN, FP e FN). Para cada exemplo, é exibida a imagem de explicação gerada pelo SHAP à direita da imagem original. Os pixels em vermelho representam uma contribuição positiva para a classificação da imagem como malária, enquanto os pixels em azul representam uma contribuição negativa.

Os exemplos das figuras 5a e 5d correspondem a imagens do conjunto de teste que foram corretamente classificadas pelo modelo—o exemplo em 5a possui rótulo verdadeiro "infectada" (VP) e o exemplo em 5d tem rótulo verdadeiro "saudável" (VN). Observamos que os pixels que correspondem à presença do parasita da malária contribuem positivamente para a classificação da figura 5a como infectada. Por outro lado, os pixels que correspondem à ausência do parasita, ou seja, o meio vazio da célula, contribuem para a classificação da figura 5a como saudável. Além disso, verificamos que as bordas das células também têm sua relevância na classificação de uma imagem, principalmente para a classe saudável.

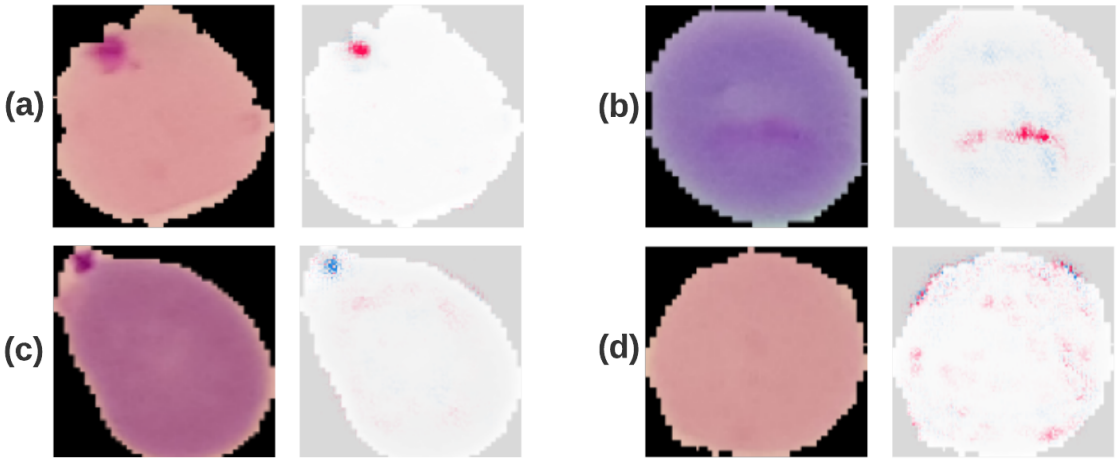

Figura 5. Método SHAP aplicado em quatro imagens selecionadas aleatoriamente do conjunto de teste. Cada exemplo corresponde a um dos termos da matriz de confusão: (a) VP, (b) FP, (c) FN e (d) VN.

As figuras 5b e 5c ilustram casos em que o modelo classificou de maneira incorreta. O exemplo 5b é uma imagem rotulada como “infectada”, mas o modelo a classificou como saudável (FN). Uma possível explicação para este erro é que a imagem não apresenta um parasita, constituindo um erro de rotulagem possivelmente causado pelo algoritmo de segmentação. Esse é o pior tipo de erro para um sistema CAD, uma vez que esse tipo de diagnóstico pode levar o paciente a complicações por ignorar o tratamento. Já o exemplo 5c é uma imagem rotulada como “saudável”, mas o modelo a classificou como “infectada” (FP). Observa-se nessa figura que o modelo não deu importância positiva aos pixels mais escuros, indicando assim que poderia ter um viés no treinamento do modelo. Contudo, pela inspeção visual podemos constatar que trata-se de um erro de rotulagem.

5.2. Identificação de erros de rotulagem na base de dados

Focamos também exclusivamente na análise SHAP de exemplos classificados incorretamente pelo modelo. Por meio dessa análise, foi possível identificar diversos erros de rotulagem na base de dados utilizada. Embora o modelo CNN tenha apresentado bom desempenho médio nas métricas, identificando corretamente em 95,86% dos casos se uma célula tinha ou não parasitas, houve 195 casos de falsos negativos (células infectadas que não foram identificadas pelo modelo). A hipótese principal aponta erros de rotulagem, possivelmente causados pelo algoritmo de segmentação e falta de verificação humana. Essa hipótese se sustenta pelo fato de que, visualmente, essas imagens quase não possuem pixels com o parasita em destaque, como mostrado na Figura 6.

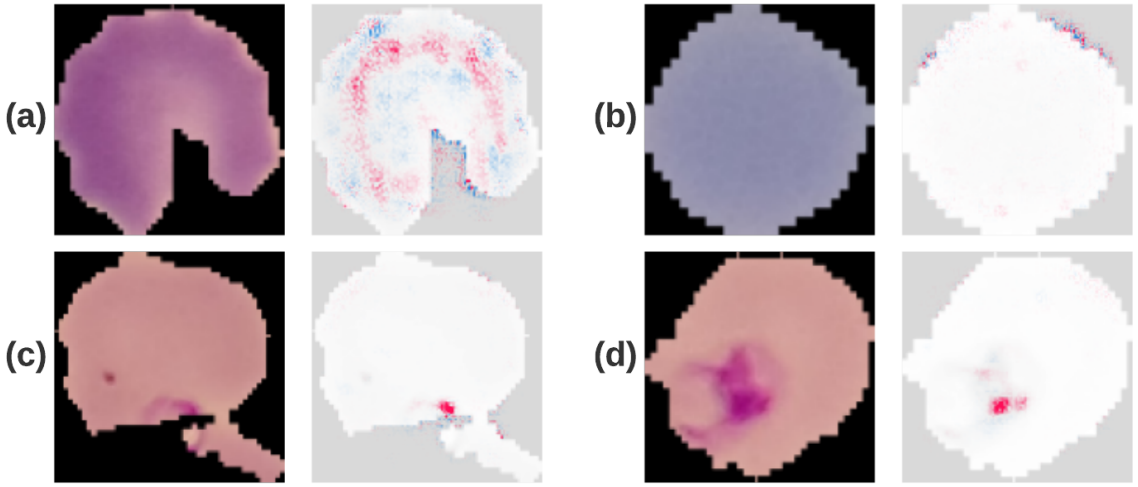

Figura 6. Exemplos de erros de classificação causados por problemas na base de dados. Os exemplos em (a) e em (b) foram incorretamente rotulados como positivos, enquanto (c) e (d) foram incorretamente rotulados como negativos.

A Figura 6 ilustra casos de erros do modelo juntamente com a saída do método SHAP. As imagens apresentadas nas figuras 6a e 6b são imagens rotuladas como “infectadas”, mas que o modelo classificou como “saudáveis”. Observamos que não há parasitas presentes nessas imagens, indicando erros de rotulagem na base. Por sua vez, as figuras 6c e 6d são imagens rotuladas como “saudáveis”, mas classificadas como “infectadas”. A característica mais importante para a classificação como infectada é a presença dos pixels com o parasita. Na Figura 6c, possivelmente a célula se estendeu para uma célula

infectada no momento da segmentação, o que pode ter levado o modelo a classificá-la erroneamente. Já na Figura 6d, é evidente que a célula está infectada, apesar de ter sido rotulada como "saudável" na base de dados.

6. Conclusão

Neste artigo, apresentamos uma abordagem para o diagnóstico da malária utilizando imagens de microscopia e um modelo CNN. Para avaliar a confiabilidade do modelo, aplicamos o método SHAP com o *Gradient Explainer*. Os resultados sugerem que o modelo foi treinado adequadamente, identificando relações entre as características das imagens e suas respectivas classes (infectada e saudável). Além disso, a aplicação do SHAP permitiu detectar erros de rotulagem humanos na base de dados—inconsistências que, de outra forma, não seriam identificáveis apenas pelas métricas de desempenho. Portanto, o uso de XAI é fundamental para garantir a confiabilidade e a qualidade das decisões dos sistemas CAD na área médica.

Embora o método SHAP possua propriedades e cálculos matemáticos complexos, sua saída é simples e intuitiva quando aplicada a imagens. A interface mostra uma representação visual da contribuição de cada pixel na classificação da imagem em uma determinada classe: os pixels vermelhos indicam uma contribuição positiva e os pixels azuis, uma contribuição negativa. Essa visualização ajuda a compreender como o modelo CNN toma suas decisões, permitindo que profissionais de saúde expliquem e validem o diagnóstico de forma simples e confiável.

Portanto, é crucial que, além de desenvolver e avaliar modelos de aprendizado de máquina para aplicações médicas, seja dada atenção à explicabilidade do modelo. Isso significa entender quais características foram aprendidas e como elas contribuem para a tomada de decisão do modelo. Somente com essa compreensão será possível confiar nas conclusões do modelo e aplicá-lo com segurança em situações reais. A explicabilidade é uma área de pesquisa em constante evolução e é fundamental para garantir que os modelos de inteligência artificial possam ser adotados em larga escala na prática clínica.

Entre as limitações deste estudo, destaca-se a ausência de análise comparativa com outras técnicas de XAI e com diferentes arquiteturas de redes neurais. Como trabalho futuro, pretendemos aplicar a técnica *label denoising* com o objetivo de reduzir o ruído presente nos rótulos, que podem ter sido corrompidos por fontes de imprecisão, como erros humanos de rotulagem. A aplicação dessa técnica pode produzir dados mais confiáveis, permitindo que reavaliemos o desempenho do modelo em relação às métricas anteriores. Após a mitigação do ruído, podemos aplicar novamente o método SHAP e incluir outras técnicas de explicabilidade, como LIME e Grad-CAM, para explicar o modelo e verificar se a melhoria na qualidade dos dados resultou em um modelo ainda mais confiável e preciso para as aplicações clínicas, além de comparar o desempenho das diferentes ferramentas de XAI.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (AUXPE-CAPES-PROEX) - Código de Financiamento 001 e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Adi-

cionalmente, este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM - por meio do projeto PDPG-CAPES.

Referências

Ali, M. S., Miah, M. S., Haque, J., Rahman, M. M., and Islam, M. K. (2021). An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications*, 5:100036.

Ashley, E. A., Pyae Phyo, A., and Woodrow, C. J. (2018). Malaria. *The Lancet*, 391(10130):1608–1621.

Berezsky, O., Liashchynskyi, P., Pitsun, O., and Izonin, I. (2024). Synthesis of convolutional neural network architectures for biomedical image classification. *Biomedical Signal Processing and Control*, 95:106325.

Boit, S. and Patil, R. (2024). An efficient deep learning approach for malaria parasite detection in microscopic images. *Diagnostics*, 14(23):2738.

Chiodini, P. L. (2014). Malaria diagnostics: now and the future. *Parasitology*, 141(14):1873–1879.

Das, A., Sahu, W., Ojha, D. K., Reddy, K. S., and Suar, M. (2022). Comparative analysis of host metabolic alterations in murine malaria models with uncomplicated or severe malaria. *Journal of Proteome Research*, 21(10):2261–2276.

Eke, C. I. and Shuib, L. (2025). The role of explainability and transparency in fostering trust in ai healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Computing and Applications*, 37(4):1999–2034.

Ersoy, I., Bunyak, F., Higgins, J. M., and Palaniappan, K. (2012). Coupled edge profile active contours for red blood cell flow analysis. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 748–751. IEEE.

Gaouar, A., Cherif, S. H., Rahmoun, A., and Daho, M. E. H. (2025). Explainable ai for early malaria detection using stacked-lstm and attention mechanisms. *Informatics in Medicine Unlocked*, 57:101667.

Glowac, C., Ferrão, J. L., and Searle, K. M. (2024). The association between infrastructure damage in the aftermath of cyclone idai and malaria risk in sofala province, mozambique: an ecological study. *Malaria Journal*, 23(1):355.

Guerra, C. A., Gikandi, P. W., Tatem, A. J., Noor, A. M., Smith, D. L., Hay, S. I., and Snow, R. W. (2008). The limits and intensity of plasmodium falciparum transmission: implications for malaria control and elimination worldwide. *PLoS medicine*, 5(2):e38.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Keras (2020). Keras: the Python deep learning API. Disponível em `https://keras.io/`. Acessado em 29 de setembro de 2022.

Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M. A., Sameer, A., Maude, R. J., et al. (2016). Cnn-based image analysis

for malaria diagnosis. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 493–496. IEEE.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Megersa, D. M. and Luo, X.-S. (2025). Effects of climate change on malaria risk to human health: A review. *Atmosphere*, 16(1):71.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Moody, A. (2002). Rapid diagnostic tests for malaria parasites. *Clinical microbiology reviews*, 15(1):66–78.

NIH (2022). Malaria Datasheet. Disponível em `https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html`. Acessado em 10 de setembro de 2022.

Parvin, B., Yang, Q., Han, J., Chang, H., Rydberg, B., and Barcellos-Hoff, M. H. (2007). Iterative voting for inference of structural saliency and characterization of subcellular events. *IEEE Transactions on Image Processing*, 16(3):615–623.

Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., and Thoma, G. (2018). Image analysis and machine learning for detecting malaria. *Translational Research*, 194:36–55.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018a). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Rajaraman, S., Silamut, K., Hossain, M. A., Ersoy, I., Maude, R. J., Jaeger, S., Thoma, G. R., and Antani, S. K. (2018b). Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. *Journal of Medical Imaging*, 5(3):034501–034501.

SHAP (2020). Welcome to the SHAP documentation. Disponível em `https://shap.readthedocs.io/en/latest/index.html`. Acessado em 18 de junho de 2022.

Shapiro, H. M., Apte, S. H., Chojnowski, G. M., Hänscheid, T., Rebelo, M., and Grimberg, B. T. (2013). Cytometry in malaria—a practical replacement for microscopy? *Current Protocols in Cytometry*, 65(1):11–20.

Shapley, L. S. (1988). A value for n-person games. *The Shapley value: essays in honor of Lloyd S. Shapley*, page 31.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Yadav, A., Verma, K., Singh, K., Tyagi, S., Kori, L., and Bharti, P. K. (2024). Analysis of diagnostic biomarkers for malaria: Prospects on rapid diagnostic test (rdt) development. *Microbial Pathogenesis*, 196:106978.