Título

Caracterização de dados de sintomas e demográficos/ambientais através da quantificação e qualificação destes para o diagnóstico alternativo de Malária

Resumo (2000 caracteres)

Doenças tropicais negligenciadas (DTNs) afetam mais de um bilhão de pessoas em todo o mundo, mas recebem recursos insuficientes para pesquisa, tratamento e vacinas. Neste contexto de DTN, a malária é uma doença febril de diagnóstico eficiente quando usados testes de análise microscópicas e testes de diagnóstico rápido. Entretanto, em regiões remotas, como reservas indígenas e comunidades de difícil acesso da Amazônia e com recursos limitados, estes testes precisos não estão amplamente disponíveis. Nestes casos, o tratamento presuntivo e a automedicação para a malária têm sido usados. No entanto, essas abordagens foram consideradas não confiáveis devido ao uso desnecessário de medicamentos contra a malária. Este projeto tem como objetivo demonstrar modelos de aprendizado de máquina supervisionado no diagnóstico da malária e subtipos, usando sintomas do paciente, características demográficas e ambientais. Para o estudo serão fundidos diversos datasets públicos de malária com o objetivo de prover um único dataset maior e mais representativo da doença. Serão considerados ainda dados regionais (estado do Amazonas) sobre diagnóstico de malária para enriquecimento do dataset final. Uma vez curado o dataset será criteriosamente avaliado para quantificar e determinar quais características são mais relevantes para o diagnóstico de malária. Nesta fase espera-se contribuir com novos algoritmos/variantes de agrupamento (do inglês clustering) que considerem tanto dados numéricos quanto categóricos de forma integrada. Finalmente, após a análise e caracterização quantificada do dataset, serão criados novos modelos de aprendizagem de máquina, baseados nesta caracterização, que funcionem como uma alternativa viável para diagnóstico de malária quando testes de análise microscópica e testes rápidos de diagnósticos não estão disponíveis.

Introdução (6000 caracteres)

Doenças tropicais negligenciadas (DTNs) afetam mais de um bilhão de pessoas em todo o mundo e são uma importante causa mortalidade em países de baixa renda [Mitra and Mawson 2017]. A denominação negligenciadas advém do fato de que essas doenças recebem investimentos insuficientes em pesquisas, produção de medicamentos e vacinas, embora sejam as que mais provocam mortes no mundo. Estas doenças atingem principalmente países tropicais em desenvolvimento, em particular, no Brasil, regiões como a Amazônia. Isso porque as DTNs são geralmente doenças infeciosas e desenvolvem-se rapidamente em clima quente e úmido [Camargo 2008] e as regiões tropicais sofrem fortes chuvas, altas temperaturas e alta umidade. Logo, é notável que essas condições forneçam cenário propício para infecções patogênicas, afetando a população destas regiões. Em geral, os agentes causadores dessas doenças são bactérias, vírus, vermes parasitas e protozoários, que são transmitidos aos seres humanos através de um humano infectado, um vetor, ou fazendo uso de um veículo contaminado como água, comida, plantas, solo [Rupali 2019]. Ademais, é importante ressaltar que o termo doença tropical"engloba doenças transmissíveis e não transmissíveis, doenças causadas por deficiências nutricionais ou condições ambientais e distúrbios genéticos nessas regiões [Zumla and Ustianowski 2012]. Os exemplos mais comuns destas enfermidades incluem malária, febre tifóide, sarampo, doença de chagas, malária, dengue, sarampo, hepatite, zika vírus e chikungunya. Além das condições ambientais comentadas, temos que a origem confusa das doenças tropicais e suas consequentes complicações em diagnosticá-las. A malária, em particular, é uma DTN mais conhecida e estudada, mas que ainda apresenta desafios importantes no seu diagnóstico e, principalmente, no alcance de métodos de diagnósticos a populações remotas da região Amazônica.

A malária compartilha sintomas semelhantes com outras doenças febris, como dengue, febre tifóide, resfriado comum, infecção do trato respiratório, dispepsia e pneumonia [Crump et al, 2017] . Testes parasitológicos microscópicos e de diagnóstico rápido, são as ferramentas padrão e recomendadas

para o diagnóstico da malária [OMS, 2020]. No entanto, em áreas onde os testes parasitológicos para a malária não estão prontamente disponíveis, a complexidade do diagnóstico da malária pode levar a erros de diagnóstico, sobrediagnóstico e tratamento presuntivo inadequado [UM, 2016]. Conforme especificado pela OMS, em situações como áreas rurais onde não há teste parasitológico disponível dentro de 2 horas após a apresentação para tratamento em centros médicos, os médicos podem fornecer um prognóstico usando um exame clínico e exame físico para tratar pacientes suspeitos [OMS, 2021]. Consequentemente, os pacientes suspeitos seriam tratados presuntivamente. Um diagnóstico clínico de malária é tradicional entre os médicos. Este método é o menos dispendioso e mais amplamente praticado. Um diagnóstico clínico chamado tratamento presuntivo é baseado nos sinais e sintomas dos pacientes e nos achados físicos no exame. Os primeiros sintomas da malária são muito inespecíficos e incluem febre, dor de cabeça, fraqueza corporal, calafrios, tontura, dor abdominal, diarréia, náusea, vômito, anorexia e prurido. Com o diagnóstico clínico, o diagnóstico incorreto é possível devido à falta de conhecimento suficiente sobre os sintomas significativos da malária (além de calafrios, febre e sudorese) e fatores não relacionados à malária para o diagnóstico clínico da malária [Bria et al, 2021]. O tratamento presuntivo pode aumentar o uso de medicamentos antimaláricos desnecessários, que têm efeitos colaterais e aumentam a disseminação da resistência aos medicamentos.

Na última década, a pesquisa sobre malária foi realizada nas áreas de teste diagnóstico e microscopia, especificamente a automação dessas ferramentas [Aqeel et al. 2023]. Esses estudos revelaram como o Aprendizado de Máquina (ML machine learning) [Kundu & Anguarj, 2023] e Aprendizagem Profunda (DL deep learning) [Hemachandran et al., 2023] podem ajudar no diagnóstico de malária através de imagens microscópicas de esfregaço de sangue e automatizar o hemograma completo, que é o teste que rastreia a infecção no sangue. Apesar dos resultados promissores desses estudos, a indisponibilidade de um microscópio e mRDT (do inglês, malaria rapid diagnostic tests) em algumas unidades de saúde em áreas restritas continuam sendo o principal desafio. Portanto, ainda é um desafio grande a compreensão de como dados clínicos podem ser melhor utilizados como ferramenta secundária para diagnóstico de malária, com aplicação em regiões mais remotas com acesso esporádico ou sem acesso a testes microscópicos e mRDT; em particular em áreas indígenas e com tensões provocadas por garimpos ilegais [Teixeira, 2014].

Ainda que o diagnóstico com base em dados sintomáticos seja menos preciso [Mariki et al., 2022; Lee et al. 2021] que mRDT e testes microscópicos (testes tipo ouro), o enriquecimento destes dados sintomáticos com dados geográficos, temporais e ambientais são capazes de aprimorar o resultado [Haddawy, 2018; Mbunge. 2022]. As características mais comumente associadas à malaria provêm da sintomatologia e incluem: área de residência de um paciente, febre, idade do paciente, mal-estar geral do corpo, data da visita, dor de cabeça, dor abdominal, dor nas costas, dor no peito, sexo do paciente, vômito, confusão, tontura, tosse e dor nas articulações. Nota-se que embora a maioria dos dados sejam categóricos, há dados numéricos também como temperatura, peso e altura do paciente que podem ser decisivos (febre) ou coadjuvantes na determinação de diagnóstico, prognóstico e tratamento. Consequentemente, se faz necessário o tratamento destes dados para uma análise que considere múltiplos tipos de dados [Dinh et al., 2021; Mousavi & Sehhati, 2023; Kar et al., 2023], representando novas oportunidades para ML.

Objetivo Geral (250 caracteres)

Prover um dataset consistente para avançar na caracterização multifatorial (informações sobre sintomas, informações demográficas/ambientais) e criação de novos modelos de ML para diagnóstico de subtipos de malária com base nessas informações.

Objetivo Específico (800 caracteres)

Considerando apenas sintomas e características demográficas e ambientais, os objetivos específicos incluem:

(1) Fusão de múltiplos datasets públicos de malária;
(2) Curadoria de um único dataset de malária limpo, consistente e homogêneo;
(3) Caracterização de coeficientes de determinação de diagnósticos de malária para as características consideradas;
(4) Criação de novos modelos e técnicas de ML para dados de múltiplos tipos que avancem no diagnóstico de malária.

Metodologia (6000 caracteres)

Na literatura são conhecidos pelos menos 231 datasets públicos de malária (https://data.world/datasets/malaria). Neste projeto, estes datasets serão avaliados e curados de forma a selecionar e fundir os dados coerentes em um único dataset público mais completo e mais representativo que permita o melhor entendimento da doença e determinação de diagnóstico, prognóstico e tratamento, quando testes de sangue e mRDT (testes tipo ouro) não são possíveis de serem conduzidos.

Após o dataset ser integrado e adequadamente curado, garantindo homogeneidade, consistência, densidade de dados, este será publicamente disponibilizado. Vale ressaltar que a este dataset, serão disponibilizados dados sobre malária provenientes da Fundação de Medicina Tropical Doutor Heitor Vieira Dourado que sejam públicos. A princípio, por se tratarem de dados públicos não há a necessidade prévia de aprovação da pesquisa em comitê de ética. Entretanto, na eventualidade novos dados precisarem de tratamento prévio para garantir que não sejam dados sensíveis, estes dados somente serão coletados após aprovação em comitê de ética. Nenhum tipo de experimento será feito com nenhum paciente e nenhum paciente pode ter sua identidade rastreada no dataset. O dataset curado garantirá aderência irrestrita à Lei Geral de Proteção de Dados.

Uma vez curado o dataset, o projeto seguirá para um estudo detalhado dos seus dados usando técnicas de mineração de dados, análise estatística e inteligência artificial explicável para quantificar a importância de cada característica/sintoma na determinação/diagnóstico da malária e de seu subtipo. Serão considerados pelo menos os subtipos P. Vivax e P. Falciparum, com possibilidade de incluir outros subtipos menos frequentes no Brasil. Nesta etapa, o projeto prevê a adaptação e criação de novos algoritmos de agrupamento aplicáveis a dados de múltiplos tipos (dados numéricos e dados categóricos) que sejam mais assertivo que os métodos do estado-da-arte.

Finalizada a etapa de análise e caracterização do dataset serão criados novos modelos de ML para quantificar a viabilidade do dataset e suas características para o diagnóstico de malária. Mais uma vez, ressalta-se que a eficácia de soluções desse tipo são menos precisas que testes microscópicos e mRDT, porém são alternativas quando estes testes não estão disponíveis (especialmente regiões remotas e de difícil acesso, tão características da região Amazônica). Todos os métodos serão avaliados com estratégias k-fold focadas em reduzir a exposição à fatores que levem a overfitting (sobreajuste, tradução livre).

O projeto, portanto, terá como subprodutos: (1) dataset curado de dados clínicos, sintomáticos e ambientais associados à malária; (2) novos algoritmos de agrupamento para dados de múltiplos tipos (numéricos e categóricos); e (3) novos modelos de ML para diagnóstico alternativo de malária e subtipos.

Referências (4000 caracteres)

Aqeel, S. et al. Towards digital diagnosis of malaria: How far have we reached? Journal of Microbiological Methods, v204, 2023.

Bria, Y. P. et al. Significant symptoms and nonsymptom-related factors for malaria diagnosis in endemic regions of Indonesia. International Journal of Infectious Diseases 103:194-200, 2021.

Crump, J.A. et al. Febrile Illness in Adolescents and Adults. Disease Control Priorities, 3rd ed, v 6: Major Infectious Diseases 365-85, 2017.

Dinh, D.-T. et al. Clustering mixed numerical and categorical data with missing values. Information Sciences, v571, 418-442, 2021.

Haddawy, P. et al. Spatiotemporal Bayesian networks for malaria prediction. Artificial Intelligence in Medicine, v. 84, 2018,127-138.

Hemachandran, K. et al.. Performance Analysis of Deep Learning Algorithms in Diagnosis of Malaria Disease. Diagnostics 2023, 13, 534.

Kar, A.K. et al. An efficient entropy based dissimilarity measure to cluster categorical data. Engineering Applications of Artificial Intelligence, v119, 2023.

Kundu, T. K. & Anguraj, D. K. A Performance Analysis of Machine Learning Algorithms for Malaria Parasite Detection using Microscopic Images, 5th IEEE Int'l Conf. on Smart Systems and Inventive Technology (ICSSIT), 2023, pp. 980-984.

Lee, Y.W. et al. Machine learning model for predicting malaria using clinical information, Computers in Biology and Medicine, v129, 2021.

Mariki, M. et al. Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach, Applied Artificial Intelligence, 36:1, 2022.

Mbunge, E. Application of machine learning models to predict malaria using malaria cases and environmental risk factors, IEEE Conf on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2022.

Mousavi, E. & Sehhati, M. A generalized multi-aspect distance metric for mixed-type data clustering, Pattern Recognition, v138, 2023.

OMS. World Malaria Report 2020. Geneva: World Health Organization. 2020.

OMS. INTRODUCTION WHO guidelines for malaria NCBI Bookshelf. Geneva: NCBI. 2021.

Teixeira, L.F. . A malária no estado do Amazonas de 2003 a 2011: distribuição espaço-temporal e correlação com populações indígena. Programa de Pós-Graduação em Medicina Tropical, Fiocruz, 2014.

UM, C. Malaria treatment in children based on presumptive diagnosis: A make or mar? Pediatric Infectious Diseases, v1:(2), 2016.