Automatização Total do Diagnóstico de Malária - Fase 1: Teste de Viabilidade - Modelo Classificatório Fuzzy - @yurisa2

1. INTRODUÇÃO

A malária é um problema de saúde global. A doença está em crescimento em 16 países (CNN, 2018). Um diagnóstico rápido é chave para seu controle. O método mais confiável e aceito é um técnico/médico analisar uma amostra de sangue (gota e esfregaço) através de microscopia com coloração (Brasil, 2018).

Por se tratar de uma doença cujo diagnóstico é visual e depende de um profissional qualificado e experiente para que os resultados sejam minimamente confiáveis, os modelos computacionais modernos começaram a criar soluções teóricas para seu diagnóstico automatizado. Este problema se tornou um exemplo de processamento de imagens digitais, com inúmeros trabalhos publicados nos últimos anos sobre o tema. Porém, na literatura relevante, não existe um trabalho que busque o diagnóstico real a partir de amostras - e com equipamentos de baixo custo (processamento digital limitado, *offline*). O mais próximo que chegou-se de automatização real, em campo, foi a aplicação descrita pelo Departamento de Saúde Norte-Americano (NIH, 2018), para *Smartphones Android*, porém sua utilização ainda é manual e depende de vários recursos para sua operação (microscópio, operador, celular, etc.).

A análise utilizada aqui foi feita puramente através das características da imagem e seus atributos estatísticos. Este primeiro passo para que o modelo seja criado dentro das capacidades de plataformas digitais sem co-processador numérico, ex.: *Arduino, NodeMCU,* etc. Este trabalho descreve o primeiro passo para a criação de um equipamento de baixo custo que, a partir da amostra de sangue, o diagnóstico seja apresentado com o mínimo de confiabilidade.

2. FUNDAMENTAÇÃO TEÓRICA

Devido ao advento da redução de custo das CPUs e GPUs experimentado nos últimos anos, inúmeros modelos e bibliotecas para processamento de enormes quantidades de dados foram criados, possibilitando um grande desenvolvimento de métodos e técnicas de processamento de imagem. Os modelos recentes são capazes de processar volumes muito grandes de dados e manipulá-los de maneiras inimagináveis anteriormente. Todos os últimos modelos desenvolvidos utilizam uma ou mais formas de redes neurais, uma técnica que aumenta demais a complexidade do modelo e o custo de processamento (CPU). Esses modelos exigem treinamento e uma vasta coleção de bibliotecas para funcionar, sendo que o modelo que, até hoje, tem a maior precisão (teórica) , se utiliza de Redes Neurais Convolucionais com até 16 camadas para que sejam efetivamente utilizados e, mesmo assim, se valeram de recortes e subutilização do corpo amostral devido ao seu alto custo computacional. A tendência a melhores resultados é a utilização deste tipo de redes neurais como melhor caminho para a excelência do resultado. (Rajaraman, 2018)

Como referência, a Tabela 1 exibe os trabalhos sobre classificação de amostras de sangue contaminadas com o plasmódio da malária na literatura recente.

Método	**Precisão**
Rajaraman et al. (2018)	0.959
Gopakumar et al. (2018)	0.977
Bibin, Nair & Punitha (2017)	0.963
Dong et al. (2017)	0.981
Liang et al. (2017)	0.973
Das et al. (2013)	0.840
Proposição deste trabalho	**0.769**
Ross et al. (2006)	0.730

Tabela 1 - Evolução da precisão dos modelos de classificação

3. METODOLOGIA

3.1 Aquisição dos dados

Os dados foram obtidos através do Instituto Nacional de Saúde estadunidense (NIH, 2018), no qual já existe um *dataset* contendo 26.558 imagens, classificadas como sadias e infectadas. As imagens contidas possuem 3 canais RGB e resoluções de 100x100 a 299x299 pixels (Rajaraman, 2018). Para que estas imagens possam ser utilizadas seguindo um modelo de visão computacional, todas foram importadas no software estatístico R e trabalhadas como uma Matriz. O ajuste da entrada (*Feature Engineering*) é o processo de normalização, adequação, agrupamento, combinação e decomposição das variáveis, para que o modelo preditivo/classificatório (GUYON, 2003) as compreenda melhor. De posse da imagem, houve a separação de cada canal e a remoção de pixels mortos (Valor igual 0 em todos os canais). Então, para cada canal, foi feita a média dos valores de cada pixel (variando de 0 a 255), que busca identificar marcações que são escuras, baixando o valor geral dos pixels (Figura 1); o valor mínimo de um pixel individual na imagem que busca verificar se possui alguma região muito escura, por fim; o número de pixels cujo valor está abaixo de 25% do máximo encontrado no canal, que busca encontrar concentrações de marcações escuras na imagem.

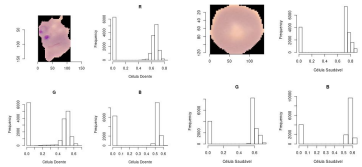

Figura 1 - Uma amostra infectada e os histogramas de cada canal, é importante observar a redução da média devido aos pontos escuros na imagem.

Figura 2 - Uma amostra saudável é mais uniforme e tende a ser mais clara, por isso existe uma elevação e concentração dos pixels mais claros.

O total de *features* ficou em 6, com mais uma coluna de resultado (0;1) para infectado ou não. Cada imagem gera então uma assinatura única, que pode ser descrita em 7 colunas.

3.2 - A Lógica Fuzzy

A lógica *fuzzy* busca determinar um número que descreve um conjunto elevado de variáveis incertas e vagas associando intervalos a variáveis linguísticas, flexibilizando a aplicação da lógica clássica, estendendo e facilitando seu uso em condições multivariadas e multidimensionais; agregando valores a variáveis além do conceito binário [0,1] de modo que sua implementação possibilita o tratamento de variáveis que antes fugiam ao controle (SHAW; SIMÕES, 1999).

A primeira ocorrência da expressão "lógica *fuzzy*" (nebulosa) para ser utilizada com base na teoria de conjuntos *fuzzy* foi usada no artigo *Fuzzy Sets* (ZADEH, 1965). A característica de lidar com a ambiguidade da informação e a incerteza do mundo real fez com que este raciocínio fosse aplicado em pesquisas de diversas áreas, tais como sistemas de controle e inteligência artificial devido sua capacidade de imitar o raciocínio humano, que considera verdades parciais ou graus de verdade. (GIGCH; PIPINO, 1980)

Na lógica *fuzzy*, as variáveis são associadas a termos que definem seu pertencimento e as operações são feitas a partir da linguagem natural, cujo maior benefício é a codificação de conhecimentos inexatos, se aproximando de um modelo cognitivo, característicos da mente humana (RUHOFF et al., 2005).

3.3 - A Lógica *Fuzzy* na visão computacional

Devido ao volume enorme de dados que uma única imagem nos trás (uma imagem VGA (640x480px) contém 307200 por canal, se for colorida 921600 pixels no total), cada imagem representa um problema de dados estatisticamente muito rico e, portanto, passível de utilização de diferentes técnicas de Inteligência Artificial.

A lógica *fuzzy*, com gradientes de verdade e pertencimento apresenta traços diferentes das técnicas já utilizadas pois diferentemente da evolução que analisamos nos algoritmos de classificação de imagens, como a própria *CNN (Convolutional Neural Networks)*, a lógica *fuzzy* não requer treino (otimização) ou grande custo computacional.

3.4 A separação do *dataset* com base no resultado

3.4.1 Definindo resultado

O *dataset* descrito (NIH, 2018), já é classificado por um técnico em microscopia, que contém imagens de lâminas segmentadas de 150

pacientes infectados pelo plasmódio da malária e 50 pacientes saudáveis, apesar do número de pacientes ser diferentes, o número de lâminas infectadas e saudáveis é o mesmo.

3.4.2 Separando os resultados em *datasets* distintos

O sistema estudado é um classificador binário, portanto, o melhor método para atingir tal resultado utilizando lógica Fuzzy é avaliar a probabilidade de cada resultado possível (0 - Células sadias, 1 - células doente (com plasmódio da malária)).

Cada *dataset* terá um sistema *Fuzzy* completo individualizado (entrada, funções de pertencimento, regras e saída próprios), e que o resultado de cada um seja a probabilidade de cada um acontecer, calculada pelo *FIS*.

3.4.3 Criação de sistemas *Fuzzy* (*FIS*) paralelos

Para que a classificação seja feita, dois sistemas *fuzzy (FIS)* são criados, e calculados independentemente, utilizando as mesmas variáveis, porém separadas pelo resultado. Conforme diagrama da Figura 3.

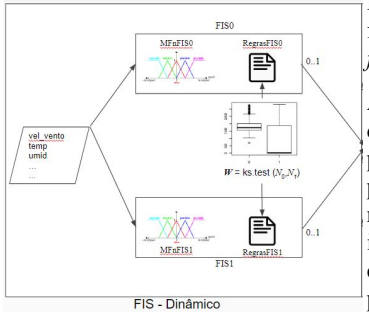

Figura 3 - Descreve o funcionamento dos dois sistemas *fuzzy* paralelos.

Ao final, cada consulta e execução dos sistemas *fuzzy* (um sistema para o resultado positivo e outro para o resultado negativo) devem retornar dois resultados independentes e simultâneos (não complementares). Onde o maior prevalece, o resultado obtido de cada sistema é estipulado de 0 a 1, e o maior é considerado o resultado.

Toda a implementação fuzzy foi feita no pacote *FuzzyR* (GARIBALDI, J, 2017), que contém todas as funções básicas do cálculo *fuzzy* baseado em regras e uma interface muito similar ao padrão da indústria, com arquivos de extensão .fis e importação das regras via matriz, tornando a experimentação rápida e prototipagem muito mais rápida e eficiente.

3.5 A formação das funções de pertinência

O sistema deve ser o mais dinâmico possível para que atenda o maior número de datasets e variáveis possível, para tanto, as funções de pertinência devem ser dinâmicas e auto-calculadas.

Dentre os muitos métodos possíveis, os mais populares utilizam redes neurais e algoritmos genéticos para definir as funções de pertinência, embora estejam bem estabelecidas na literatura, consomem muito tempo e recursos computacionais (ASANKA, 2017).

O método utilizado no trabalho, descrito na Figura 4 abaixo, define os limites através dos valores obtidos nos *box-plots* das colunas (Quartis e mediana de cada *feature*), criando 3 estágios de probabilidade sendo a mediana a maior probabilidade e quanto mais afastado, menor. Os dois conjuntos mais externos (distantes da mediana) tem a pertinência BAIXA, os dois intermediários (mais próximos da mediana) tem a pertinência MÉDIA e o conjunto (somente um) da média tem a pertinência alta. Os prováveis outliers são excluídos na camada de alimentação do algoritmo para reduzir o ruído da amostra.

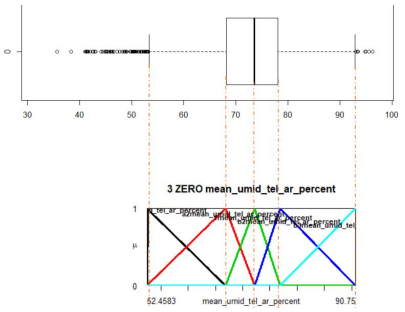

Figura 4 - *Boxplots* gerando funções de pertinência

Os dados obtidos a partir das primeiras 100 imagens de cada dataset (infectados e saudáveis) e podem ser gravadas e transferidas para um modelo de produção. Embora o próprio sistema, quando concluído, possa facilmente criar suas próprias funções de pertinência, o aumento do corpo amostral não tem relação com um aumento da precisão do modelo.

3.7 - Criação das regras

O modelo de Mamdani, utiliza-se de conjuntos que recebem uma atribuição linguística que são associadas por uma base de regras (Zadeh, 1965), cada regra deve associar uma entrada com uma saída..

Diferentemente de outros métodos de Inteligência Artificial /*Machine Learning* que dependem de aprendizado (otimização), o Sistema *Fuzzy* não requer treino, porém sofre do tamanho da lista de regras.

Uma estratégia foi divisada para superar este obstáculo, que mesmo alterando o resultado numérico do sistema Fuzzy, ainda dá a direção correta do cálculo, foram feitas então as regras incompletas do sistema, não relacionando as variáveis de entrada entre si, cada variável de entrada somente tem relação com o resultado.

Em uma relação com a Estatística, isso torna o sistema Fuzzy Univariado, sendo que o resultado final é um acumulado de todas as variáveis calculadas individualmente.

Para atingir essa tabela, o sistema cria um modelo preenchido com "0" (Regra vazia) e escreve em cima de cada regra somente, a partir de um modelo. Utilizando este método, invés de E^V (Entrada elevada a variações), convertemos o número de regras para E * V (Entrada multiplicada pelas variações). Porém exclui-se a influência de uma variável sobre a outra e suas correlações.

3.8 - A aplicação de pesos

O conceito de regras limitadas empregado neste trabalho facilita a aplicação de pesos a cada variável, portanto, é possível criar pesos diferentes para cada entrada no sistema, fazendo com que o resultado penda para as variáveis com maior peso.

O peso das regras de cada variável é calculado a partir da diferença da distribuição normal entre a mesma variável no dataset com resultado 0 e o dataset com resultado 1, utilizando o Teste Kolmogorov-Smirnov. Para uma maior otimização dos pesos (maximização dos resultados), todos os pesos foram colocados em uma matriz e então normalizados, ou seja, a maior diferença gera um peso de valor 1, e a menor de valor 0 (o que não é calculado no sistema Fuzzy resultante), conforme exemplificado na Figura 5 abaixo.

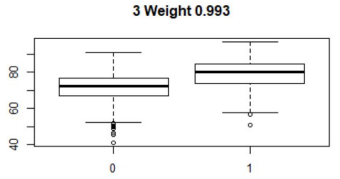

Figura 5 - *Boxplots* representando a distribuição normal da mesma variável nos *datasets* 0 e 1, quanto mais diferente, maior o peso (*weight*)

3.9 - Funções de resultado

O resultado dos dois sistemas é compilado em uma tabela junto ao resultado classificado do dataset (resultado original) e compila o cálculo em uma Matriz confusão.

4 - APLICAÇÃO E RESULTADOS

Devido à natureza homogênea do objeto (análise de imagens não é cíclica ou acontece em períodos e não existe autocorrelação como em séries temporais), os experimentos e seus resultados também foram desenhados desconsiderando períodos.

Uma vez que este trabalho é baseado em conhecimento de domínio, a escolha das variáveis também foi feita utilizando este raciocínio. A matriz confusão e suas características principais estão listadas abaixo:

```
          Reference
Prediction     0     1
         0 10520  3722
         1  3158  12256

               Accuracy : 0.7680

            Sensivity : 0.7671
            Specificity : 0.7691
         Pos Pred Value : 0.7951
         Neg Pred Value : 0.7387
                F1 Score : 0.7808
```

5. CONCLUSÕES

Diante do resultado, fica claro que o sistema proposto tem capacidade de classificar células doentes com um mínimo de informação.

É necessário, porém, testes e ensaios para que mais variáveis sejam criadas e auto-selecionadas, com base estatística.

O sistema e seu script R, como está posto também apresenta um consumo muito baixo de recursos computacionais, tornando sua utilização e teste extremamente ágeis em um computador pessoal intermediário.

Embora o modelo esteja muito aquém de experimentos recentes envolvendo tecnologias avançadas (*CNN*), ainda assim apresenta acertos de 79.51% no resultado positivo, tornando-o qualificado para desenvolvimentos futuros. Sendo a próxima fase da pesquisa o ensaio do sistema embarcado em plataformas digitais.