VERSÃO ON-LINE - ISSN: 2525-7293

USO IMAGENS MÉDICAS PARA CRIAÇÃO DE SOLUÇÕES DE APOIO A DECISÃO BASEADAS EM DEEP LEARNING

USING MEDICAL IMAGES TO DEVELOP DEEP LEARNING-BASED DECISION SUPPORT SOLUTIONS

Victor De Almeida Thomaz, docente do curso de Ciência da Computação – UNIFESO, e-mail: victorthomaz@unifeso.edu.br

Kleber Daniel Mattos Viana, discente do curso de Ciência da Computação – UNIFESO, e-mail: kleberdanielviana@unifeso.com.edu.br

Yuri Domingues Santos, discente do curso de Ciência da Computação – UNIFESO, e-mail: yuridomingues@unifeso.com.edu.br

RESUMO

Atualmente a análise de imagens médicas desempenha um papel indispensável tanto na pesquisa científica quanto no diagnóstico clínico. Nos últimos anos houve um aumento do interesse nas pesquisas voltadas para o desenvolvimento de sistemas para o apoio ao diagnóstico médico. As soluções baseadas em aprendizagem profunda (Deep Learning) surgem como uma técnica efetiva para abordar diversas tarefas de imagens médicas. No entanto, o desempenho destas abordagens está fortemente relacionado com a qualidade dos conjuntos de dados (datasets) empregados para o treinamento de modelos aprendizado de máquina. Neste contexto, o objetivo deste projeto é coletar e analisar imagens médicas mais relevantes para criação de conjuntos de dados de treinamento para modelos aprendizagem profunda. Tais modelos, quando adequadamente treinados, superam em muito abordagens tradicionais, em geral, alcançando níveis de precisão elevados, motivando sua aplicação em sistemas de análise de imagens médicas assistida por computador. Este projeto faz uso de imagens obtidas a partir de base de dados públicas, mais especificamente das modalidades de exames de colonoscopia CVC-ClinicDB e mamografia VinDrMammo. Sobre estes datasets exploramos abordagens baseadas em métodos de similaridade de imagens Structural Similariy Index (SSIM) e Normalized Mutual Information (NMI). Além disso, aplicamos os algoritmos de clustering K-Means e HDBSCAN com o objetivo agrupar imagens semelhantes. Obtivemos resultados promissores utilizando os métodos HDBSCAN e K-Means nos dados CVC-ClinicDB. No entanto, para o dataset VinDrMammo estes métodos não geraram clusters adequados. A partir desta metodologia de seleção será possível guiar a criação de datasets otimizados para aprendizado de máquina.

Palavras-chave: Imagens Médicas; Aprendizado de Máquina; Conjunto de Dados; Similaridade de Imagens; Clustering.

ABSTRACT

Currently, medical image analysis plays an indispensable role both in scientific research and in clinical diagnosis. In recent years, there has been growing interest in research focused on the development of systems to support medical diagnosis. Deep Learning-based solutions have emerged as an effective technique to address various medical imaging tasks. However, the performance of these approaches is strongly related to the quality of the datasets used for training machine learning models.

In this context, the objective of this project is to collect and analyze the most relevant medical images for the creation of training datasets for deep learning models. Such models, when properly trained, significantly outperform traditional approaches, generally achieving high levels of accuracy, which motivates their application in computer-assisted medical image analysis systems. This project makes use of images obtained from public databases, specifically from colonoscopy exams (CVC-ClinicDB) and mammography exams (VinDrMammo). On these datasets, we explored approaches based on image similarity methods such as the Structural Similarity Index (SSIM) and Normalized Mutual Information (NMI). In addition, we applied clustering algorithms such as K-Means and HDBSCAN with the aim of grouping similar images. We obtained promising results using HDBSCAN and K-Means on the CVC-ClinicDB dataset. However, for the VinDrMammo dataset, these methods did not generate suitable clusters. Through this selection methodology, it will be possible to guide the creation of optimized datasets for machine learning.

Keywords: Medical Imaging; Machine Learning; Datasets; Image Similarity; Clustering.

INTRODUÇÃO

A análise de imagens médicas é indispensável na atualidade. Tanto a pesquisa médica de ponta realizada em laboratório quanto o diagnóstico feito pelos médicos, exigem uma grande quantidade de evidências fornecidas pela análise de imagens médicas para fazer conjecturas ou diagnósticos (WANG, 2021). Com o avanço tecnológico, equipamentos de exames médicos são capazes de fornecer imagens de alta qualidade. Isto favoreceu o aumento do interesse em pesquisas voltadas para apoio ao diagnóstico em modalidades como, por exemplo, colonoscopia (GIL et al., 2015) e mamografia (SILVA et al., 2023).

> O termo aprendizado de máquina se refere a algoritmos que são capazes de aprender uma tarefa específica examinando dados fornecidos como entrada (GIGER, 2018). Uma definição para algoritmos de aprendizado de máquina foi apresentada por Mitchell (MITCHELL, 1997) como:

"Um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T com desempenho medido por P, se seu desempenho sobre as tarefas T, mensurado por P, melhora de acordo com a experiência E." (tradução nossa).

Uma tarefa T é a previsão que deve ser feita a partir da análise dos dados de entrada. Os dados de entrada definem a experiência E, sendo efetuado um aprendizado de padrões sobre estes dados para melhorar previsão. Os resultados da previsão são avaliados por P, que indica se será necessária uma nova análise dos dados, até que se obtenha o nível considerado adequado de previsão de acordo com o contexto do problema (GROUP, 2017).

O processo de aprendizado é geralmente dividido em duas etapas. A primeira é o treinamento, onde os dados de entrada analisados pelo algoritmo na busca de padrões. A segunda etapa é chamada de teste. Neste caso, novos dados de entrada (que não foram usados no treinamento) são recebidos pelo algoritmo que irá utilizar os padrões aprendidos na fase de treinamento para computar uma previsão para cada dado novo.

Quando o aprendizado é chamado de supervisionado (SHALEV-SHWARTZ; BEN-DAVID, 2014), significa que os dados usados no treinamento trazem alguma indicação sobre o que se deseja prever. Por exemplo, se o objetivo do sistema é prever a localização de pessoas em imagens, os dados de treinamento serão compostos das imagens de fato e de respectivas anotações que indiquem a localização de cada pessoa em cada imagem. Deste modo, o sistema tem sempre uma referência (chamado de anotação ou etiqueta ou *ground truth*) para efetuar uma validação do aprendizado antes da etapa de testes, por exemplo.

Os dados são cruciais no aprendizado supervisionado. No entanto, estes devem ser cuidadosamente selecionados pois, frequentemente, as fontes de dados podem trazer informações que não contribuem e que ainda prejudicam a etapa de aprendizado, levando o modelo a funcionar abaixo do esperado.

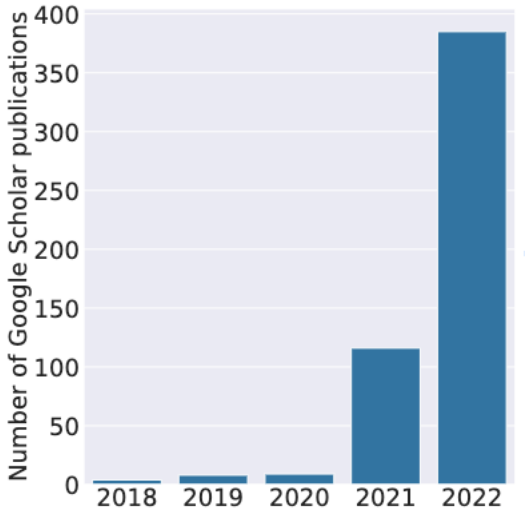
Um dos desafios relacionados aos conjuntos de dados na área médica são as restrições devido a privacidade dos pacientes e o alto custo de anotação, visto que um especialista precisa avaliar as imagens. Isto implica em uma baixa disponibilidade de dados médicos para o desenvolvimento de pesquisas. Para fins de comparação, pode-se citar o *dataset* ImageNet, composto por imagens (não médicas) extraídas da internet, para treinamento de modelos de aprendizagem profunda. Este conjunto de dados possui 10.000.000 (dez milhões) de imagens no total (DENG et al., 2009). Em contraste, o *dataset* de imagens de colonoscopia disponibilizado publicamente, CVC-ClinicDB, possui apenas 612 imagens (BERNAL et al., 2015). Os conjuntos de dados de imagens em outras áreas da visão computacional, como rostos, objetos, carros, animais etc., são abundantes na atualidade, no entanto, as imagens médicas, especialmente sua anotação, ainda são escassas e insuficientes (WANG et al., 2021).

Devido a esta baixa disponibilidade de dados na área médica é preciso estabelecer processos com o objetivo de otimizar a seleção de imagens, de modo que contribuam para melhores resultados dos modelos de aprendizagem.

JUSTIFICATIVA

Na atualidade existe uma alta disponibilidade de dados, porém, à medida que a quantidade de dados aumenta, mais desafiadora se torna a seleção dos melhores dados para uso na etapa de treinamento de modelos de aprendizado de máquina. Durante anos, a pesquisa e desenvolvimento de modelos para aprendizado de máquina foi focada em otimizar as respectivas arquiteturas. Nos últimos anos uma abordagem direcionada aos dados, chamada de *Data-Centric AI* (NG; LAIRD; HE, 2024), tem se tornado popular. Neste contexto, a preocupação maior está na qualidade dos dados e não no aprimoramento das arquiteturas dos modelos em si. A publicação de trabalhos que utilizam a abordagem *Data-Centric* aumentou nos últimos anos, como apresentado na Figura 1.

Figura 1: Quantidade de publicações que citam a abordagem centrada em dados (Data Centric AI) no *site* Google Acadêmico. Foi utilizada a frase para busca a seguir: "*data-centric AI*".



Fonte: Artigo ZHA et al. (2023).

Neste contexto, o presente trabalho está direcionado a uma abordagem que prioriza a qualidade dos dados para o treinamento de modelos. Os resultados apresentados em trabalhos recentes que aplicam a abordagem *Data-Centric* tem contribuído para aumentar a qualidade dos modelos de aprendizados de máquina, justificando o aumento do interesse e publicações.

OBJETIVOS

Objetivo geral

O objetivo geral deste projeto é criar um conjunto de dados de imagens médicas que contribua para aprimorar os resultados de análise de imagens médicas em sistemas baseados em aprendizagem de máquina, mais especificamente nas redes neurais convolucionais profundas (CNNs).

Objetivos específicos

Os objetivos específicos deste projeto são:

- Coletar imagens médicas de conjuntos de dados públicos;
- Estabelecer processos e/ou métodos para seleção de imagens médicas mais relevantes;

O restante deste trabalho está dividido conforme descrito a seguir. A seção de Revisão bibliográfica apresenta de forma breve trabalhos que tratam de métodos e técnicas que serão utilizados para análise de similaridade e agrupamento de imagens. A seção de Metodologia descreve brevemente os datasets utilizados, as etapas empregadas para seleção de imagens e as tecnologias aplicadas no desenvolvimento. A seção de Resultados e Discussão mostra os resultados do processo de seleção de imagens obtidos pelos métodos de similaridade e agrupamento, discute os resultados e disponibiliza o código de implementação do projeto, utilizado para obtenção dos resultados, em repositório público. Por fim, na seção de Considerações Finais, é apresentada a síntese dos principais resultados e os trabalhos futuros.

REVISÃO BIBLIOGRÁFICA

As etapas deste trabalho se concentram em explorar técnicas ou métodos para seleção de imagens para o conjunto de dados de treinamento do modelo de aprendizado de máquina. Uma forma de realizar esta tarefa é empregar métodos de análise de similaridade de imagens. Nesta seção serão apresentados uma breve descrição dos trabalhos da literatura referentes aos métodos e técnicas utilizados no desenvolvimento do projeto.

O método SSIM (*Structural Similariy Index*) (WANG et al., 2004) (WANG; BOVIK, 2009) é uma métrica amplamente utilizada que avalia a similaridade estrutural entre duas imagens. Ela considera luminância, contraste e estrutura, dando uma pontuação entre -1 (diferente) e 1 (idêntico) as imagens comparadas.

Outra forma de analisar similaridade de imagens é com o uso da medida estatística NMI (*Normalized Mutual Information*) (STUDHOLME; HILL; HAWKES, 1999). Este método avalia a distribuição de probabilidade de uma amostragem de pixels de duas imagens. Imagens semelhantes irão compartilhar os mesmos dados, assim, esta informação mútua será medida. Pode ser entendido

como uma amostragem de pixels obtida em duas imagens para medir a certeza de que os valores de um conjunto de pixels mapeiam para valores semelhantes na outra imagem.

K-Means é um método baseado em aprendizado de máquina não supervisionado muito popular para agrupamento (*clustering*) de dados (CHONG et at., 2021). O objetivo é particionar os dados semelhantes em K grupos distintos, onde k é quantidade de grupos que precisa ser definida previamente. O K-Means tende a apresentar melhores resultados se os dados são distintos pois produzirá clusters mais separados. No contexto de imagens, isto significa que imagens com aparência variada serão mais facilmente identificadas pelo K-Means. Neste método é fundamental a escolha adequada da quantidade de *clusters*, i.e., o valor de K. Existem métodos que auxiliam na escolha do K, como o *Silhouette Score* (ROUSSEEUW, 1987). Esta é uma métrica usada para avaliar a qualidade dos resultados do agrupamento fornecido pelo K-Means. Ela mede o quão bem cada ponto de dados foi escolhido para determinado cluster em comparação aos outros agrupamentos.

HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo de aprendizado de máquina não supervisionado utilizado para agrupar dados (CAMPELLO; MOULAVI; SANDER, 2013). Ele se baseia nos conceitos do DBSCAN, mas que não requer a definição prévia do número de *clusters*. Este método se baseia na criação de uma hierarquia de agrupamentos sendo mais eficaz em dados com ruído.

METODOLOGIA

A metodologia deste projeto está baseada em processos de análise de imagens para seleção das mais relevantes para treinamento de um modelo de aprendizado de máquina especializado em detectar e/ou segmentar áreas de interesse na imagem em questão. Por exemplo, no caso de imagens de colonoscopia a área de interesse é a região de uma lesão, como o pólipo. Já nas imagens de mamografia a área de interesse pode ser uma calcificação no tecido fibroglandular. Uma vez definido este conjunto de imagens mais relevante, o mesmo pode ser utilizado para treinar um modelo de CNN (i.e., um sistema aprende automaticamente, por meio de exemplos, as características das imagens). Este modelo pode ser incorporado em um protótipo que recebe uma nova imagem e efetua uma análise automática, destacando a área de interesse.

Muitas pesquisas de análise de imagens baseadas em aprendizado de máquina estão direcionadas a melhorias na arquitetura das redes neurais convolucionais. Em contraste, trabalhos mais recentes estão focados na melhoria dos dados utilizados no treinamento, i.e., centrada em dados. Isto se deve ao fato de que dados não adequados prejudicam o aprendizado da rede. Uma mesma arquitetura pode apresentar resultados inferiores ou elevados, dependendo da seleção dos dados.

Neste contexto, estabelecer processos para seleção adequada de dados é fundamental. É importante destacar que uma metodologia obtenção dos melhores dados é geralmente dependente do domínio da aplicação. Um processo ou método no domínio de imagens médicas pode não ser eficiente para imagens aéreas.

Inicialmente, pretende-se utilizar balanceamento de classes, por meio do algoritmo da mochila (SALKIN; DE KLUYVER, 1975) ou outros. Será investigado a análise de imagem por medidas de similaridade (WANG et al., 2004). A seleção imagens mais diversas tende a melhorar a etapa de treinamento.

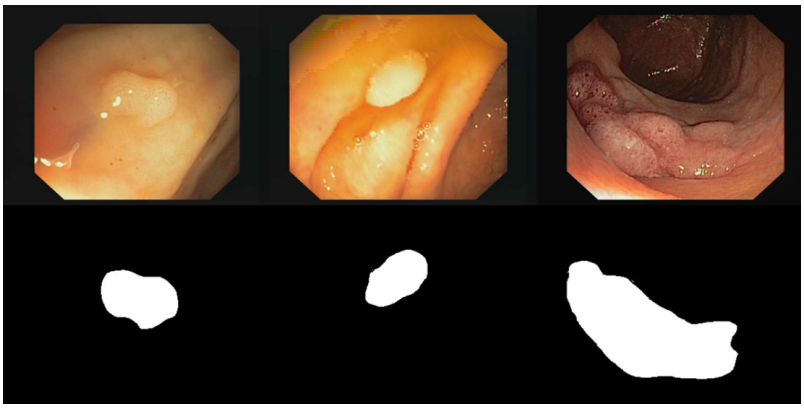
Cada conjunto de dado resultante da etapa de seleção, será empregado no treinamento de modelo de aprendizado de máquina (etapa de treinamento). Durante o treinamento serão obtidas métricas de avaliação desta etapa. Posteriormente, será definido um conjunto de dados de teste, nunca utilizado na etapa de treinamento do modelo, para simular um caso de uso real, em que o modelo deve encontrar os padrões aprendidos em uma imagem do mesmo domínio, porém inédita. De igual modo, nesta etapa de teste do modelo, serão obtidas métricas de avaliação.

Os melhor processo ou método de seleção de imagens, em cada domínio, será definido a partir da avaliação das métricas. Esta etapa de análise de resultados é fundamentalmente baseada na análise estatística, considerando como modelo preferencial aquele com maior probabilidade de acerto, de acordo com objetivo do sistema.

Para o desenvolvimento do projeto foi utilizada a linguagem de programação Python. Esta linguagem é amplamente empregada para criação de aplicações de processamento de imagem e aprendizado de máquina, que se enquadra no contexto deste projeto. Dentre as bibliotecas principais pode-se citar NumPy, Pandas, Scikit-Learn, TensorFlow/Keras, PyTorch, OpenCV, Scikit-Image.

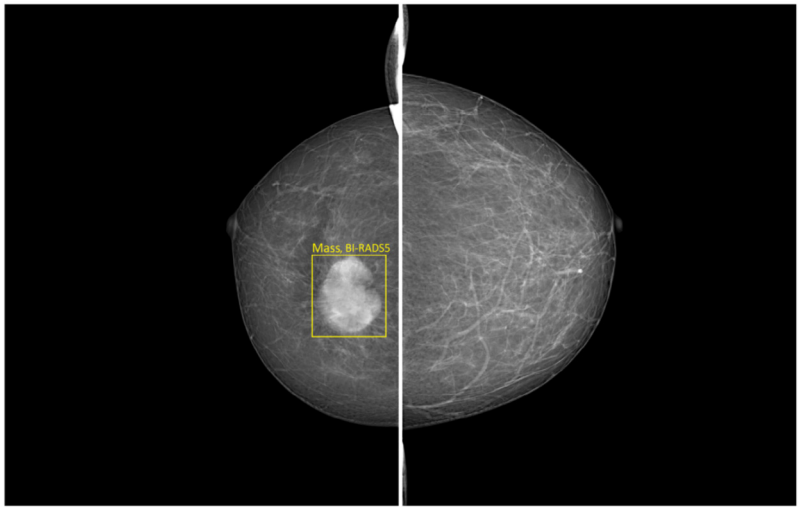
No presente projeto serão utilizados dois conjuntos de dados de imagens médicas disponibilizados publicamente. Na modalidade de colonoscopia serão empregadas as imagens do CVC-ClinicDB (BERNAL et al., 2015). A modalidade de mamografia será investigada com o dataset VinDr-Mammo (NGUYEN et al., 2023). A Figura 2 apresenta três exemplos de pólipos e suas respectivas imagens *ground truth*. Um exemplo de imagem do exame de mamografia das mamas direita e esquerda da vista crânio-caudal (CC) pode ser vista na Figura 3.

Figura 2: Exemplos de imagens do *dataset* CVC-ClinicDB. A primeira linha mostra imagens originais, enquanto a segunda linha mostra as imagens *ground truth* correspondentes.



Fonte: Artigo BERNAL et al. (2023).

Figura 3: Imagem de mamografia *dataset* VinDr-Mammo. Vista Crânio Caudal (CC) das mamas esquerda e direita.



Fonte: Artigo NGUYEN et al. (2023).

A partir destes dados, o objetivo é criar um conjunto de dados para treinamento do modelo de aprendizado que máquina aprimorado, no sentido de uma seleção de imagens otimizada. Tal característica faz com que a arquitetura obtenha mais exemplos diferenciados e aumentando a qualidade do resultado do treinamento. O treinamento em si será realizado em projetos futuros.

RESULTADOS E DISCUSSÃO

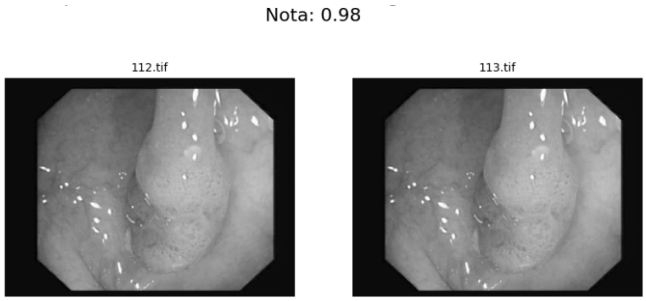
Nesta seção serão apresentados exemplos dos resultados obtidos utilizando os métodos de similaridade nas imagens de colonoscopia do *dataset* CVC-ClinicDB e VinDrMammo. Importante destacar que as imagens neste no *dataset* CVC-ClinicDB foram extraídas do vídeo original do exame, logo, uma sequência de quadros (*frames*) próximos tendem a registrar imagens muito parecidas. O *dataset* VinDrMammo apresenta um grau de semelhança de imagens desafiador, pois é composto por imagens em tons de cinza.

A seguir são apresentados os resultados nos métodos SSIM e NMI aplicados aos dados CVC-ClinicDB.

No método SSIM o índice de similaridade é indicado pela faixa de valores entre -1.0 e 1.0, onde mais próximo de 1.0 indica maior similaridade, enquanto valores próximos a -1.0 indicam que as imagens são mais distintas. Exemplos dos resultados com o método SSIM podem ser vistos nas Figuras 4 e 5. No método NMI, os valores próximos a 2.0 indicam imagens semelhantes e valores próximos a 1.0 denotam imagens distintas. Exemplos dos resultados com o método NMI podem ser vistos nas Figuras 6 e 7.

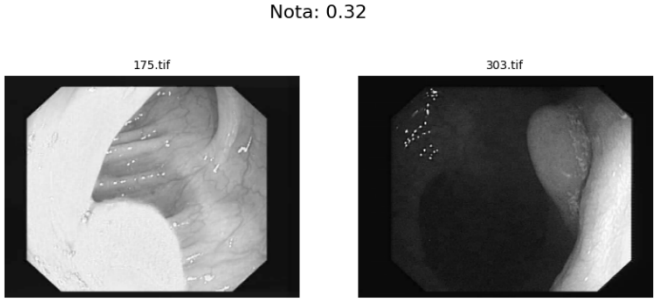
Com método SSIM foi possível observar que este método conseguiu identificar imagens semelhantes, como mostrado na Figura 4. As imagens são de fato frames de vídeo que mostram o mesmo pólipo de pontos de vista diferentes, com índice de similaridade de 0.98. Já a Figura 5 apresenta imagens exemplo de imagens diferentes com índice de similaridade de 0.32.

Figura 4: Par de imagens semelhantes. Índice de similaridade 0.98 de acordo com o método SSIM.



Fonte: Elaborado pelo autor.

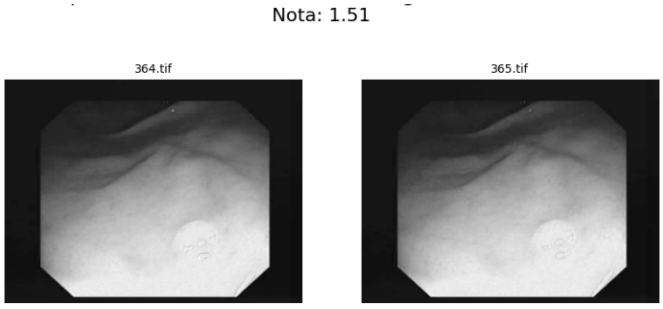
Figura 5: Par de imagens distintas. Índice de similaridade 0.32 de acordo com o método SSIM.



Fonte: Elaborado pelo autor.

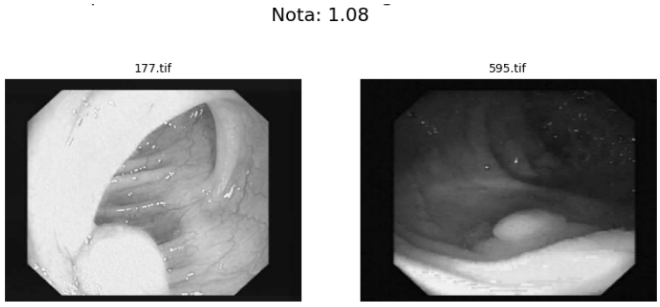
Observando os resultados do método NMI, vemos que imagens de frames próximos do vídeo de colonoscopia apresentaram valores de similaridade altos, conforme vista na Figura 6. A Figura 7 é um exemplo de imagens diferentes, com o índice de similaridade mais baixo 1.08.

Figura 6: Par de imagens semelhantes. Índice de similaridade 1.51 de acordo com o método NMI.



Fonte: Elaborado pelo autor.

Figura 7: Par de imagens distintas. Índice de similaridade 1.08 de acordo com o método NMI.



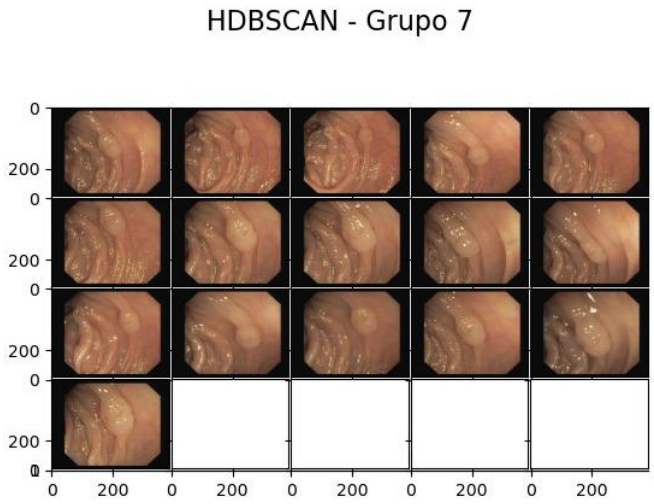
Fonte: Elaborado pelo autor.

Avaliando todo o conjunto de imagens de colonoscopia e os respectivos resultados dos métodos SSIM e NMI, observou-se a partir de uma análise visual empírica que o método SSIM é mais adequado para categorizar as imagens semelhantes ou diferentes.

As abordagens baseadas nos métodos K-Means e HDBSCAN aplicados aos dados CVC-ClinicDB são apresentadas a seguir. Estes métodos procuram adicionar cada imagem em um grupo que mais esta se assemelha.

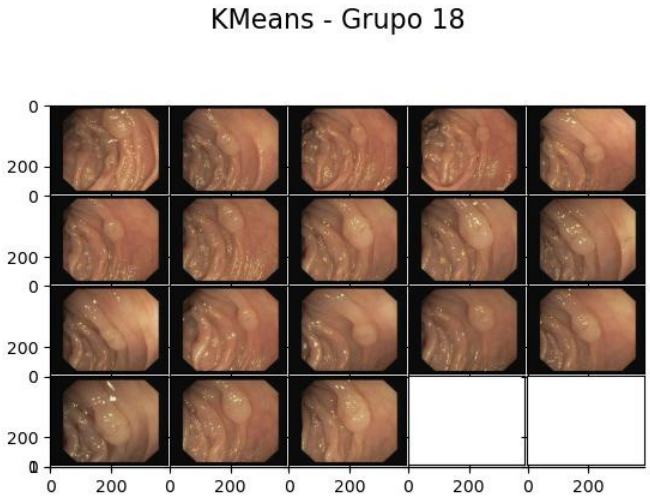
O método HDBSCAN gerou 25 agrupamentos (*clusters*) e o K-Means 27. A Figura 8 apresenta um exemplo com imagens semelhantes. Imagens de pólipos agrupados pelo K-Means podem ser vistas na Figura 9. O *cluster* gerado pelo K-Means é formado por 18 imagens e o HDBSCAN inclui 16 imagens.

Figura 8: Agrupamento de imagens do *dataset* CVC-Clinic obtidos pelo método HDBSCAN.



Fonte: Elaborado pelo autor.

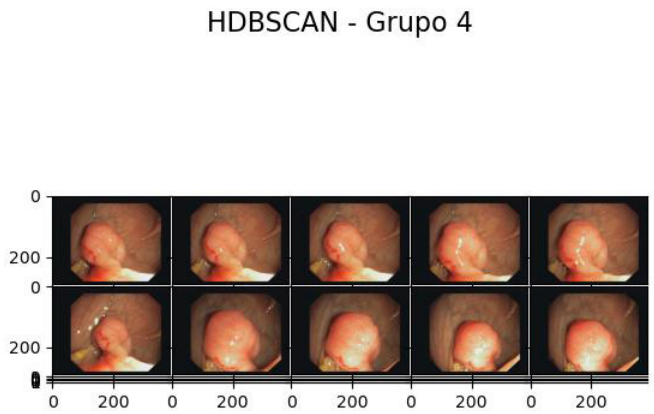
Figura 9: Agrupamento de imagens do *dataset* CVC-Clinic obtidos pelo método K-Means.



Fonte: Elaborado pelo autor.

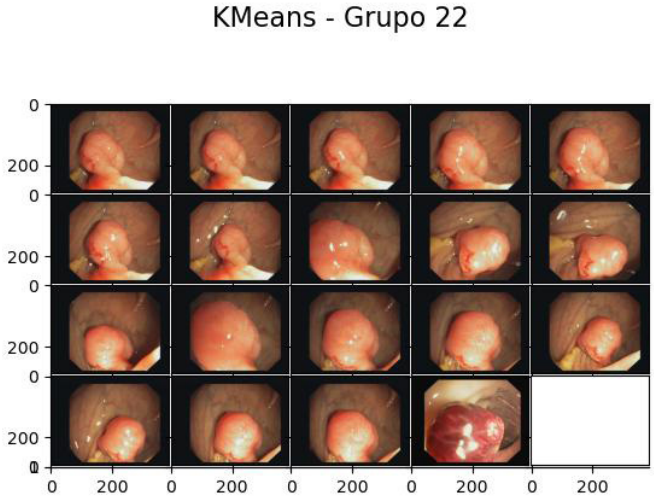
Outro exemplo de agrupamento pelo método HDBSCAN pode ser visto na Figura 10, formado por 10 imagens. Já o método K-Means (Figura 11) incluiu 19 imagens ao cluster, porém a última imagem (quarta linha) parece ser diferente das demais. O método K-Means depende da escolha correta da quantidade de agrupamentos (*clusters*) K. O processo de escolha do valor K ocorre por meio do cálculo do *Silhouette Score* (SC) para cada valor de K. Por exemplo, dada uma faixa de valores de K de 5 a 30, é computado uma pontuação do SC para cada K. O valor de K com maior pontuação de *Silhouette Score* é escolhido como melhor K para os dados em questão. O agrupamento apresentado na Figura 11 provavelmente pode ser melhorado aumentando o valor de K.

Figura 10: Agrupamento de imagens do *dataset* CVC-Clinic obtidos pelo método HDBSCAN.



Fonte: Elaborado pelo autor.

Figura 11: Agrupamento de imagens do *dataset* CVC-ClinicDB obtidos pelo método K-Means.

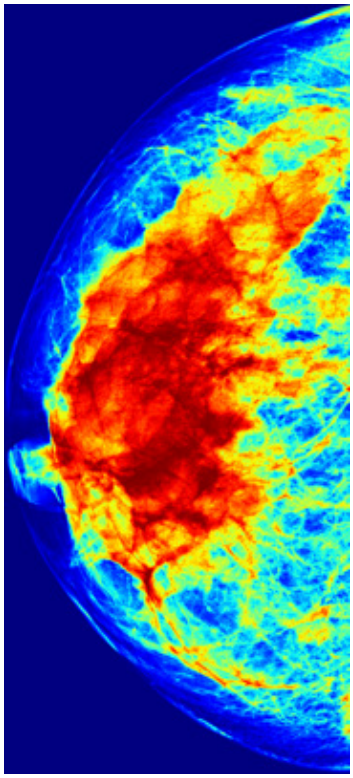


Fonte: Elaborado pelo autor.

Os métodos mais promissores com os dados de colonoscopia CVC-ClinicDB foram HDBSCAN e K-Means. Comparando estes métodos de forma empírica, visualmente observamos que o K-Means apresenta resultados de agrupamentos de imagens semelhantes mais adequadamente para os dados CVC-ClinicDB. No entanto, os resultados do HDBSCAN são aceitáveis e em alguns casos até equivalente ao K-Means. Outra forma de avaliação numérica comparativa precisa ser implementada para ser obtido outro parâmetro de comparação destes métodos.

A partir desta informação resolvemos investigar se estas metodologias são efetivas para as imagens de mamografia do *dataset* VinDrMammo. No entanto observamos que estes métodos geram apenas dois grupos de imagens. As imagens de mamografia são tradicionalmente capturadas em equipamentos que geram imagens em tons de cinza, tornando mais desafiadora a distinção de estruturas semelhantes para fundamentar o agrupamento dos métodos HDBSCAN e K-Means. Exploramos alternativas superar este problema, sendo a principal a colorização (*color map*). Aplicar um mapa de cores a uma imagem em tons de cinza pode melhorar significativamente a visibilidade de características ou diferenças na intensidade que, de outra forma, seriam difíceis de perceber. A ideia é criar uma versão mais adequada da imagem original para ser utilizada como entrada para os métodos de clusterização citados. Esta nova imagem é uma versão colorida como um mapa de calor conforme apresentado na Figura 12.

Figura 12: Versão da imagem de mamografia após a colorização (*color map*).



Fonte: Elaborado pelo autor.

No entanto, os métodos HDBSCAN e K-Means não foram capazes de agrupar as imagens de colonoscopia de forma satisfatória, gerando poucos grupos, mesmo após o processo de colorização. Para as imagens da modalidade de colonoscopia será necessária uma análise adicional para entender melhor como tratar estas imagens antes de aplicar os métodos HDBSCAN e K-Means.

Para alcançarmos estes resultados foi desenvolvido um projeto implementado na linguagem de programação Python e disponibilizado em um repositório no GitHub[1].

CONSIDERAÇÕES FINAIS

Soluções baseadas em Aprendizado de Máquina são altamente dependentes da disponibilidade e qualidade dos dados utilizados para treinamento. No contexto de imagens médicas, estas soluções empregam modelos que recebem uma imagem como entrada e executam algum tipo de análise, indicando áreas de interesse (e.g., localização de pólipos). Tais modelos, quando adequadamente treinados, superam em muito abordagens tradicionais.

Os dados têm papel fundamental para um treinamento adequado. Nem sempre aumentar a quantidade de dados garante um melhor resultado. Assim, no presente trabalho, foram utilizados os datasets públicos de colonoscopia CVC-ClinicDB e de mamografia VinDrMammo. No dataset de colonoscopia empregamos métodos de similaridade de imagens baseados nas métricas (SSIM e NMI) e os métodos de agrupamento baseados em aprendizado de máquina (HDBSCAN e K-Means) com o objetivo agrupar imagens semelhantes.

Nos resultados para os dados CVC-ClinicDB foram apresentadas as imagens semelhantes e distintas com os métodos SSIM e NMI. A partir desta informação é possível conduzir a criação de

1 https://github.com/VictorThomaz/feso_picpq2024

um conjunto de dados com imagens de maior similaridade. No entanto, os métodos baseados em aprendizado de máquina HDBSCAN e K-Means apresentaram resultados superiores, em especial os clusters produzidos pelo K-Means.

No contexto do dataset de mamografia, reproduzimos os métodos baseados em aprendizado de máquina HDBSCAN e K-Means para o agrupamento de imagens semelhantes. Porém, estas abordagens não foram capazes criar clusters adequados. Entendemos que isso pode estar relacionado ao alto grau de semelhança entre as imagens de mamografia capturadas em tons de cinza, diferentemente das imagens de colonoscopia que são coloridas.

Para os dados CVC-ClinicDB concluímos que os grupos de imagens semelhantes gerados pelo K-Means serão mais adequados para guiar a criação do dataset otimizado de treinamento, validação e teste em novos projetos que poderão dar continuidade a esta pesquisa.

Os dados VinDrMammo são mais desafiadores pois as imagens apresentam alto grau de semelhança (imagens tons de cinza). As abordagens mais promissoras sobre os dados CVC-ClinicDB (imagens coloridas), isto é, HDBSCAN e K-Means mostraram resultados limitados sobre as imagens de mamografia pois não conseguiram agrupar as imagens corretamente.

Como trabalhos futuros serão será empregado um método para gerar o dataset otimizado para treinamento a partir dos dados agrupados pelo método K-Means, que apresentou melhor resultado com dados CVC-ClinicDB. Uma possível estratégia é aplicar o algoritmo da mochila (Knapsack) onde cada item a ser adicionado na mochila é um grupo completo de imagens semelhantes. Características dos objetos nas imagens podem ser consideradas como dimensões da mochila (critérios de escolha). Deste modo, um grupo de imagens semelhantes pode ser adicionado ao conjunto de treinamento e nenhuma imagem deste grupo aparecerá nos conjuntos de validação ou teste, evitando o vazamento de dados.

Metodologias mais adequadas as imagens de mamografia em tons de cinza serão conduzidas nos trabalhos futuros. Uma alternativa possível é aplicar pré-processamentos sobre as imagens originais que ajudem a diferenciar estas imagens, por exemplo, operações morfológicas. Deste modo, o formato das regiões internas da mama pode ser destacado para facilitar o agrupamento pelo método K-Means.

REFERÊNCIAS

BERNAL, Jorge et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, v. 43, p. 99-111, 2015.

CAMPELLO, Ricardo JGB; MOULAVI, Davoud; SANDER, Jörg. Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia conference on knowledge discovery and data mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160-172.

CHONG, Bao et al. K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, v. 4, n. 5, p. 37-40, 2021.

DENG, J. et al. ImageNet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248-255, 2009, doi: 10.1109/CVPR.2009.5206848.

GIGER, M. L. Machine learning in medical imaging. *Journal of the American College of Radiology* 15, 3 (2018), 512–520.

GIL, Debora et al. 3D Stable Spatio-Temporal Polyp Localization in Colonoscopy Videos. In: *International Workshop on Computer-Assisted and Robotic Endoscopy*. Springer International Publishing, 2015. p. 140-152.

GROUP, R. S. W., et al. *Machine learning: the power and promise of computers that learn by example*. Technical report, 2017.

HASSAN, Noor Salah et al. Medical images breast cancer segmentation based on K-means clustering algorithm: a review. *Asian Journal of Research in Computer Science*, v. 9, n. 1, p. 23-38, 2021.

MITCHELL, T. M. *Machine learning*. Mcgraw-hill higher education. New York (1997).

NG, A.; LAIRD, D.; HE, L., *Data-centricai competition*, DeepLearning AI, 2024. Disponível em: < https://https-deeplearning-ai. github. io/data-centric-comp>. Acesso em: 29 jul. 2024.

NGUYEN, H.T. et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*. May 2023 12;10(1):277.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, v. 20, p. 53-65, 1987.

SALKIN H.M.; DE KLUYVER, C.A. The knapsack problem: a survey. *Naval Research Logistics Quarterly*. 1975 Mar;22(1):127-44.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

STUDHOLME, C; HILL, D.L.G.; HAWKES, D.J., (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition* 32(1):71-86 DOI:10.1016/S0031-3203(98)00091-0

SILVA, S.V. et al. (2023). A Data-Centric Approach for Pectoral Muscle Deep Learning Segmentation Enhancements in Mammography Images. In: Bebis, G., et al. *Advances in Visual Computing*. ISVC 2023. Lecture Notes in Computer Science, vol 14361, Springer, Cham. 2023 https://doi.org/10.1007/978-3-031-47969-4_5.

WANG, Z; BOVIK, A.C.; Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures, *Signal Processing Magazine*, IEEE, vol. 26, no. 1, pp. 98-117, Jan. 2009.

WANG, Z; BOVIK, A.C.; SHEIKH, H. R.; SIMONCELLI, E. P., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

WANG, J. et al. A Review of Deep Learning on Medical Image Analysis. *Mobile Netw Appl* 26, 351–380 (2021). https://doi.org/10.1007/s11036-020-01672-7.

ZHA, Daochen et al. Data-centric ai: Perspectives and challenges. In: *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 2023. p. 945-948.