Catalogação na publicação
Seção de Catalogação e Classificação

C331aa Carvalho Neto, Vidal Elias de.
Aplicação de modelos de linguagem visual na classificação de lesões tumorais: um estudo com PubMedCLIP / Vidal Elias de Carvalho Neto. - João Pessoa, 2024.
16 f. : il.

Orientação: Clauirton Siebra.
TCC (Graduação) - UFPB/CI.

1. Inteligência artificial. 2. VLM. 3. Odontologia. 4. Medicina. I. Siebra, Clauirton. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

Aplicação de Modelos de Linguagem Visual na Classificação de Lesões Tumorais: Um Estudo com PubMedCLIP

Vidal Elias de Carvalho Neto[1]

[1]Centro de Informática - Universidade Federal da Paraíba (UFPB)

João Pessoa - PB - Brasil

vidalneto@academico.ufpb.br

Abstract. *The application of artificial intelligence (AI) in dentistry faces challenges such as data limitations and a lack of standardization in clinical information, hindering progress, particularly in the diagnosis of oral lesions. With oral cancer representing a significant portion of neoplasia cases in Brazil, early detection is essential, and AI can play a crucial role in this process. This study investigated the capacity of the pre-trained visual language model, PubMedCLIP, to classify malignant and benign tumor lesions, achieving AUC-ROC and F1-Score values of 0.9900 and 0.9665, respectively. While the results are promising, the scarcity of clinical data limits the full utilization of the model, which was designed for multimodal tasks.*

Resumo. *A aplicação de inteligência artificial (IA) na odontologia enfrenta desafios como a limitação de dados e a falta de padronização nas informações clínicas, dificultando avanços, especialmente no diagnóstico de lesões orais. Com o câncer de boca representando uma parte significativa das neoplasias no Brasil, a detecção precoce é essencial, e a IA pode ser crucial nesse processo. Este trabalho investigou a capacidade do modelo de linguagem visual pré-treinado, PubMedCLIP, em classificar lesões tumorais malignas e benignas, alcançando AUC-ROC e F1-Score de 0.9900 e 0.9665, respectivamente. Embora os resultados sejam promissores, a escassez de dados clínicos limita a utilização plena do modelo, que foi projetado para tarefas multimodais.*

1. Introdução

A utilização de inteligência artificial na odontologia tem mostrado potencial, mas ainda é um ramo com escassez de trabalhos. Enquanto a IA revolucionou áreas como a medicina e a radiologia, a odontologia apresenta dificuldade em ser contemplada pelas novas tecnologias. De acordo com Schwendicke et al. (2020), a limitação de dados em odontologia impede que algoritmos de IA alcancem o nível de precisão observado em outras áreas da saúde. Além disso, a falta de padronização de imagens e de informações clínicas em bancos de dados disponíveis dificulta a realização de trabalhos na área. Apesar das dificuldades, algumas áreas da odontologia apresentam progresso considerável ao relacionar-se com a IA, como é o caso da análise de radiografia e diagnóstico de cáries, enquanto em áreas como o diagnóstico de lesões ainda há uma escassez de estudos.

De acordo com o Instituto Nacional de Câncer (INCA), em 2023 o câncer de boca representou 4,6% do total de casos de neoplasias no Brasil. Além disso, segundo o Ministério da Saúde, 50% dos casos de câncer de boca são diagnosticados em estágios avançados da doença. A inteligência artificial tem a capacidade de auxiliar no processo de diagnóstico dessa lesão, facilitando o diagnóstico precoce. Estudos recentes, como o de Lee et al. (2021), demonstram que modelos de IA podem identificar lesões orais em imagens fotográficas com alta acurácia, melhorando a qualidade dos diagnósticos e a escolha dos tratamentos. Uma das ferramentas que podem ser utilizadas para a identificação e classificação dessas lesões são os Modelos de linguagem visual (VLM, Visual Language Models).

Modelos de linguagem visual utilizam processamento de linguagem natural (NLP) combinado com visão computacional para realizar uma interpretação de uma imagem juntamente com as descrições relacionadas a mesma imagem. Sendo assim uma boa opção como ferramenta no diagnóstico de lesões bucais, pois pode utilizar tanto a imagem da lesão em si quanto os detalhes do prontuário odontológico e realizar uma predição do possível diagnóstico. Estudos como o de Li et al. (2022) mostram que os VLMs podem alcançar alta precisão em tarefas multimodais e apresentar resultados satisfatórios no diagnóstico de lesões.

O trabalho tem como principal objetivo verificar a capacidade de um modelo VLM pré-treinado em diferenciar lesões tumorais entre malignas e benignas após a realização de um *fine-tuning*.

2. Trabalhos relacionados

Um dos estudos essenciais que viabilizam este trabalho é o de [Eslami et al. 2022] uma vez que é o artigo que descreve o treinamento do VLM utilizado neste estudo. O PubMedCLIP desenvolvido pelos autores é um modelo que realiza *fine-tuning* do domínio do modelo CLIP (Contrastive Language-Image Pre-training), alimentado com pares de imagem-texto extraídos da base de dados do PubMed, uma das maiores e mais relevantes biblioteca de artigos da área da saúde. Após expandir o domínio do modelo original, foi realizado o treinamento com o dataset ROCO (Radiology Objects in COntext). Todas as imagens do dataset foram extraídas do PubMed e todos os textos foram extraídos de suas respectivas legendas, sendo elas curtas e de até 20 palavras.

Os resultados observados no trabalho de [Eslami et al. 2022] em aperfeiçoar o modelo CLIP para o ambiente biomédico são promissores quando comparados com os outras conclusões observados na área de MedVQA (Medical Visual Question Answering). Tendo o *visual encoder* do modelo, apresentado um resultado 3% melhor do que outros *visual encoders* pré-treinados, estando assim em um patamar próximo do que possa ser considerado estado da arte.

Dessa forma, este trabalho tem como objetivo utilizar o *visual enconder* do modelo PubMedCLIP para realizar a classificação de lesões bucais. Sendo assim, se distanciando do objetivo inicial do CLIP de ser multimodal e trabalhar com a interpretação simultânea de imagem-texto, utilizando-o apenas como uma ferramenta de classificação de imagem.

Os resultados serão comparados com os obtidos pelo trabalho de [Tanriver et al. 2021] cujo objetivo era realizar a detecção e classificação de lesões orais utilizando deep learning. Durante o estudo vários modelos são utilizados, porém para a classificação os melhores resultados foram obtidos pelo modelo EfficientNet-b4, um modelo da família EfficientNet, desenvolvida pela Google. Apesar de utilizar dataset diferentes, é um dos poucos trabalhos que realiza classificação de imagens médicas não-laboratoriais, por isso será utilizado como objeto de comparação.

3. Metodologia

Esta seção apresenta os procedimentos implementados, incluindo a descrição da natureza da base de dados, do ambiente de desenvolvimento, das métricas de avaliação e dos modelos utilizados.

3.1. Base de dados

As imagens utilizadas neste trabalho foram retiradas do *Oral Images Dataset*, um dataset de imagens de lesões orais que contêm fotos obtidas por câmeras de celular e imagens intra-orais. As imagens foram obtidas por profissionais de diferentes hospitais e universidades de Karnataka na Índia. O dataset também contém *augmenting images*, imagens criadas a partir das imagens originais através de rotação, redimensionamento e inversão com a finalidade de aumentar a quantidade de dados disponíveis utilizando imagens reais. Os dados originais são compostos por 165 imagens de lesões benignas e 158 imagens de lesões malignas. Após o processo de *data augmentation* temos 1115 imagens de lesões malignas e 1115 imagens de lesões benignas.

Figura 1: Exemplos de imagens do Oral Image Dataset

3.1.1. Transformação do Banco de Dados

Para facilitar o processo de treinamento do modelo, o banco de dados de imagens foi transformado em um arquivo CSV contendo informações essenciais sobre cada imagem. Inicialmente, todas as imagens foram organizadas em diretórios separados por classes (benignas e malignas). Em seguida, um script foi desenvolvido para percorrer essas pastas, extraindo o caminho de cada imagem e associando-a ao respectivo rótulo (classe). Cada linha do arquivo CSV gerado contém dois campos principais:

1. Caminho da Imagem: O caminho completo da imagem no sistema de arquivos.
2. Rótulo: A classe correspondente da imagem (0 para benigno e 1 para maligno).

Essa abordagem possibilitou a integração eficiente do dataset com as bibliotecas que seriam utilizadas durante o desenvolvimento.

3.2. Ambiente de desenvolvimento

Para o desenvolvimento deste trabalho foi utilizado a linguagem de programação Python 3.10, juntamente com as seguintes bibliotecas:

- **pandas - Versão: 2.1.1:** Biblioteca *open-source* para manipulação e análise de dados.
- **torchvision - Versão: 0.16.0:** Biblioteca *open-source* para a transformação e carregamento de dados de visão computacional em PyTorch.
- **Pillow (PIL) - Versão: 10.0.0:** Biblioteca *open-source* para processamento de imagens em Python.
- **numpy - Versão: 1.26.0:** Biblioteca *open-source* para computação numérica em Python.
- **scikit-learn - Versão: 1.3.1:** Biblioteca *open-source* para aprendizado de máquina em Python.
- **transformers (Hugging Face) - Versão: 4.35.0:** Biblioteca *open-source* para processamento de linguagem natural e visão computacional.
- **torch (PyTorch) - Versão: 2.1.0:** Biblioteca *open-source* para aprendizado profundo.
  Todo o desenvolvimento foi realizado utilizando o Google Colab

Os experimentos e desenvolvimento deste trabalho foram realizados utilizando a plataforma Google Colab, que dispõe de recursos computacionais em nuvem. O Google Colab utiliza um processador Intel(R) Xeon(R) CPU @ 2.20GHz e 12 GB de memória RAM.

3.3. Métricas de Avaliação e Função de perda

Para verificar a corretude da pesquisa, foram utilizadas certas métricas de avaliação a fim de verificar se o modelo consegue distinguir de maneira correta entre as duas classes de imagens diferentes. As métricas utilizadas foram a Área sob a curva (AUC-ROC) e F1 Score. Para calcular a função de perda durante o treinamento foi utilizado a Cross Entropy Loss.

3.3.1. Cross Entropy Loss

A Cross Entropy Loss é uma função de perda amplamente utilizada em tarefas de classificação, particularmente em problemas de classificação multiclasse. Ela mede a diferença entre a distribuição verdadeira dos rótulos (y_i) e a distribuição prevista pelo modelo, penalizando fortemente previsões incorretas. Sua fórmula leva em consideração a probabilidade prevista para a classe correta ($\widehat{y_i}$) e aplica uma penalidade logarítmica ao erro.

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{N} y_i log(\widehat{y_i})$$

No trabalho, a Cross Entropy Loss é utilizada para ajustar o modelo CLIP durante o treinamento, atualizando seus parâmetros com base na discrepância entre os rótulos verdadeiros (benigno ou maligno) e as probabilidades previstas para cada imagem. Isso permite que o modelo melhore sua capacidade de classificar corretamente as imagens de tumores bucais.

3.3.2. Área sob a curva ROC (AUC- ROC)

A Área sob a curva ou AUC-ROC mede a capacidade do modelo em distinguir entre as classes. Para um classe positiva $y = 1$ e uma classe negativa $y = 0$, a área é obtida através da integração da curva ROC como descrito na fórmula a seguir:

$$\text{AUC-ROC} = \int_0^1 TPR(FPR)dfpr$$

Onde TPR é a taxa de verdadeiros positivos (True Positive Rate), FPR é a taxa de falso positivo (False Positive Rate) e dfpr é a taxa de variação do valor de falsos positivos.

3.3.3. F1-Score

Trata-se da média harmônica entre a precisão e sensibilidade (*recall*), pode ser obtido através da fórmula:

$$F1 = 2. \frac{Precisão.Sensibilidade}{Precisão+Sensibilidade}$$

Temos ainda que:

$$Precisão = \frac{TP}{TP+FP}$$

$$Sensibilidade = \frac{TP}{TP+FN}$$

Onde TP são os resultados marcados como verdadeiro positivo (True Positive), FP são os resultados marcados como falso positivo (False Positive) e FN são os resultados que representam um falso negativo (False Negative). Um dos motivos pelo qual escolher essa métrica é a importância que a mesma insere aos valores de falso positivo e falso negativo que são extremamente importantes para o objeto do trabalho.

3.4. Modelo

Como já mencionado previamente, o modelo utilizado neste estudo é o PubMedCLIP, que é utilizado principalmente pela capacidade do seu *visual encoder* de interpretar imagens médicas com maior precisão. Todavia o modelo foi treinado a partir do modelo CLIP, que será discutido a seguir.

3.4.1 CLIP

O modelo CLIP (Contrastive Language-Image Pre-Training), desenvolvido pela OpenAI, é um framework que aprende a partir de pares texto-imagem, aplicando uma abordagem contrastiva a fim de designar descrições textuais às suas respectivas imagens relacionadas. O modelo utiliza dados multimodais para capturar as representações semânticas do texto e imagem em um espaço de características compartilhadas, onde ambas as modalidades podem ser comparadas diretamente. No processo de treinamento o modelo utiliza uma função de perda InfoCE, que maximiza a similaridade entre pares de texto-imagem correspondentes e minimiza a similaridade para os pares incorretos.

Segundo o artigo original, "CLIP demonstrates the ability to perform zero-shot image classification across various domains without requiring task-specific training" (Radford et al., 2021). Essa capacidade de generalização sem treinamento adicional em tarefas específicas diferencia o CLIP de outros modelos convencionais, permitindo sua aplicação direta em diferentes áreas. No entanto, embora o CLIP tenha sido projetado para ser generalista e de alto desempenho, ele enfrenta limitações ao lidar com dados altamente especializados, como aqueles do domínio biomédico.

Essas limitações decorrem do fato de que o CLIP foi treinado em dados da internet que, embora vastos e variados, não contêm terminologias técnicas e imagens complexas próprias de áreas como a medicina. O modelo não foi exposto suficientemente a imagens médicas específicas, como radiografias ou imagens de exames patológicos, nem a textos com jargões médicos e descrições detalhadas de diagnósticos. Como resultado, seu desempenho em tarefas biomédicas, como a interpretação de imagens clínicas, pode ser inferior. Para lidar com essas limitações, foram criadas adaptações como o PubMedCLIP, que utiliza dados médicos específicos para ajustar o modelo e torná-lo mais eficaz na identificação e interpretação de imagens no contexto médico.

3.4.2. PubMedCLIP

Devido às limitações da natureza do treinamento original do modelo CLIP, o trabalho de [Eslami et al. 2022] foi melhorar a capacidade do modelo CLIP de interpretar pares de imagem-texto no âmbito biomédico, com o intuito de utilizá-lo na área de MedVQA (Medical Visual Question Answering). Como mencionado anteriormente, utilizando os textos e

imagens da biblioteca de artigos PubMed, o novo modelo foi treinado para responder questões específicas dessa área.

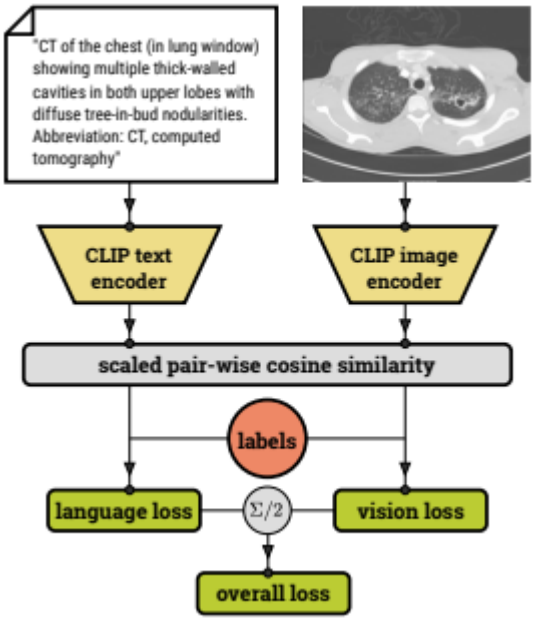

Figura 2: Pipeline do pré-treinamento do modelo PubMedCLIP. Fonte: Adaptado de [Eslami et al. 2022].

A partir desse pré-treinamento o PubMedCLIP tem seu próprio *text encoder* e *visual encoder* responsáveis por interpretar texto e imagem respectivamente, porém diferentemente do modelo CLIP, consegue apresentar resultados melhores quando comparado ao modelo original em relação ao âmbito biomédico. Os resultados são demonstrados no artigo, onde a acurácia do *visual-encoder* do novo modelo ultrapassa em aproximadamente 40% quando comparado ao modelo original. Todavia é importante notar que o *text encoder* do PubMedCLIP/CLIP não apresenta os melhores resultados quando comparados com outros modelos utilizados na área de MEDVQA [Eslami et al. 2022].

3.5. Treinamento dos modelos

Nessa etapa serão descritos os treinamentos realizados com o modelo.

3.5.1 Treinando o modelo CLIP

Em um primeiro momento foi realizado o treinamento com o modelo CLIP original para verificar a acurácia do modelo original como classificador para os nossos dados. A base de dados foi dividida entre treinamento e teste, utilizando o método *train_test_split* da

biblioteca sckitilearn, onde 80% dos dados foram designados para o treinamento e o restante para o processo de teste.

O modelo CLIP é iniciado utilizando o método *clip.load()*, onde como parâmetro utilizamos a arquitetura ViT-32, mesma arquitetura utilizada no PubMedCLIP, dessa forma tanto o modelo e o pré-processador para essa arquitetura serão carregados.

Após realizar a transformação dos dados utilizando o pré-processador do modelo, definimos o otimizador *Adam* com um *learning rate* de 0,00001, é também definido a função *nn.CrossEntropyLoss()* como a função utilizada para o cálculo da perda. Utilizando 10 épocas o modelo foi então treinado, calculando para cada época o valor de perda. Ao fim do treinamento o estado do modelo treinado foi salvo para que possa ser utilizado durante o processo de teste.

3.5.2. Treinando o modelo PubMedCLIP

Semelhante ao modelo CLIP, os dados para o treinamento desse modelo foram divididos da mesma maneira como o modelo anterior, porém para iniciarmos o modelo precisaremos utilizar o método *CLIPMODEL.from_pretrained()* e como parâmetro "flaviagiammarino/pubmed-clip-vit-base-patch32" fazendo com que o modelo seja carregado diretamente da plataforma *hugginface* com o modelo PubMedCLIP, da mesma maneira faremos o carregamento do pré-processador utilizando o mesmo método.

Utilizando o pré-processador realizaremos as transformações necessárias nas imagens e textos, utilizaremos o otimizado *Adam* com um learning rate de 0,00001 e a mesma função de perda sendo a *nn.CrossEntropyLoss()*. O que difere entre os modelos é a quantidade de épocas, por ser um modelo já pré-treinado para o ambiente biomédico, foi observado durante alguns treinamentos que não há a necessidade de um número alto de épocas para obtermos um resultado satisfatório, dessa forma para os resultados obtidos o número de épocas utilizado durante o treinamento foi 5. Assim como no modelo CLIP, em cada época foi calculado o valor de perda e ao término do treinamento, o estado do modelo também foi salvo para utilizar durante a realização dos testes.

3.6. Teste dos modelos

Nesta seção, serão detalhados os procedimentos adotados durante a avaliação dos modelos treinados.

3.6.1. Teste dos modelos

Primeiramente, para realizarmos os testes, visto que o CLIP e o PubMedCLIP são modelos multimodais, é necessário a entrada de um texto correspondente a imagem, utilizamos os rótulos (malignos e benignos) como opções de textos válidos, que nesse caso funciona como as classes que gostaríamos de distinguir.

Percorremos todas as imagens reservadas para teste e realizamos a predição utilizando o estados salvos em cada um dos modelos, onde verificamos a similaridade da imagem com o texto inserido através do valor obtido através do método *logits_per_image*.

Durante o teste realizamos o cálculo das métricas já citadas, AUC-ROC e F1-Score para avaliação dos resultados.

4. Resultados e discussões

Nesta seção os resultados obtidos através do treinamento e teste dos dois modelos serão analisados, comparados e discutidos. Além disso, serão comparados com os resultados obtidos durante outros estudos.

4.1 Resultados após o treinamento

4.1.1. Treinamento do modelo CLIP

Após o treinamento do modelo CLIP, ao avaliar os valores da métrica escolhida, nesse caso a função perda, podemos observar resultados compatíveis com o esperado. O modelo não consegue aprender bem e apenas realiza “chutes” para tentar identificar corretamente cada uma das imagens. Essa conclusão vem a partir dos valores obtidos da função perda em cada uma das épocas. Segundo Bishop (2006) a função de perda de entropia cruzada assume um valor próximo de 0.69 quando o modelo realiza previsões aleatórias em uma tarefa de classificação binária.

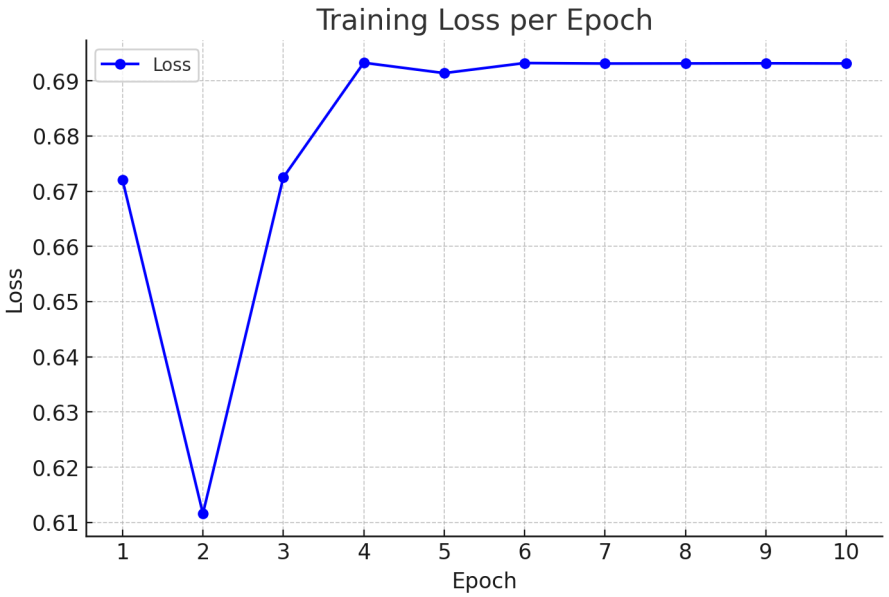

Figura 3: Gráfico do valor de perda por época do treinamento do modelo CLIP

Esses valores corroboram o pressuposto original de que o modelo CLIP não seria adequado para trabalhar com imagens médicas.

4.1.2. Treinamento do modelo PubMedCLIP

Os resultados obtidos após o treinamento do PubMedCLIP são promissores, com a função de perda apresentando valores baixos que diminuem a cada época, demonstrando que o modelo está conseguindo aprender a identificar as imagens e seus respectivos rótulos.

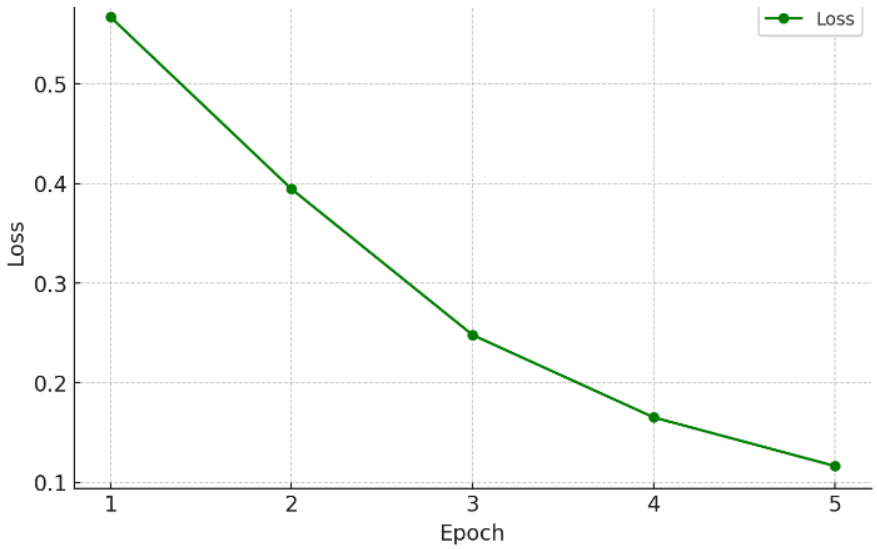

Figura 4: Gráfico do valor de perda por época do treinamento do modelo PubMedCLIP

Esses valores já demonstram resultados positivos quando comparados ao modelo CLIP original, um valor da função de perda baixo ao final do treinamento demonstra que o modelo está conseguindo prever corretamente a maior parte dos rótulos.

4.2. Resultados após os testes

Com os testes finalizados, os resultados mais uma vez corroboram os resultados trazidos pelo estudo de [Eslami et al., 2022]. O modelo pré-treinado PubMedCLIP apresentou valores muito superiores do que o modelo CLIP original.

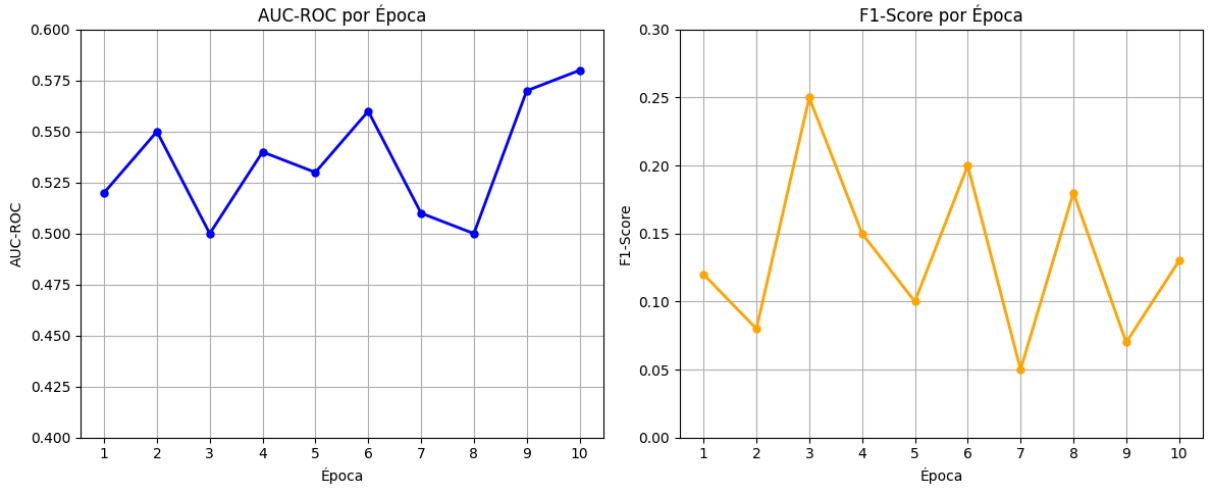

Figura 5: Gráficos dos resultados do teste do modelo CLIP

Como podemos observar, os valores para o AUC-ROC próximos a 0.5 sugerem que o modelo está realizando previsões aleatórias. Previsões essas que também resultam em valores para o F1-Score próximos a zero.

Para o modelo PubMedCLIP, mais uma vez observa-se resultados positivos, valores altos para AUC-ROC e F1-Score, o que demonstra que o modelo aprendeu a distinguir entre as duas classes corretamente e que a tarefa de classificação binária utilizando o modelo PubMedCLIP é possível.

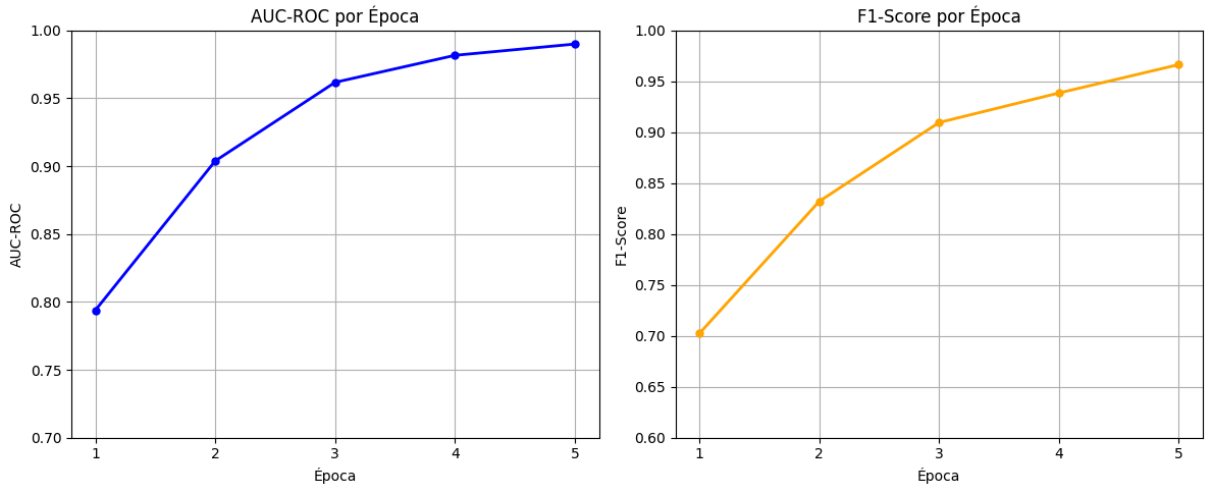

Figura 6: Gráfico dos resultados do teste do modelo PubMedCLIP

Como já mencionado, esses valores indicam resultados positivos, valores para o AUC-ROC e F1-Score próximos a 1 são desejados. Os valores finais a título de comparação para cada um foram 0.9900 para o AUC-ROC e 0.9665 para o F1-Score. Quando comparado ao trabalho realizado por [Tanriver et al. 2021], os resultados obtidos demonstram uma superioridade na escolha do PubMedCLIP, visto que o melhor resultado para o F1-Score entre todos os modelos utilizados durante o estudo foi de 0.855 durante a etapa de testes.

5. Conclusão

O objetivo principal deste trabalho era demonstrar a capacidade do modelo PubMedCLIP em classificar imagens de lesões tumorais entre malignas e benignas. Com os valores finais para o AUC-ROC e para o F1- Score sendo respectivamente 0.9900 e 0.9665, podemos considerar que comprovamos a hipótese inicial.

Apesar dos resultados obtidos serem favoráveis à utilização do *visual encoder* do modelo como uma ferramenta para classificação, isso não exclui o fato de que o modelo foi arquitetado para realizar tarefas multimodais. A escassez de dados como, prontuário do atendimento ou descrição da lesão, faz com que a utilização do modelo nessa área seja reduzido à tarefa de classificação binária de lesões.

5.1 Trabalhos Futuros

Apesar dos resultados obtidos serem satisfatórios, ainda está longe de ser suficiente para satisfazer a demanda que existe na odontologia em respeito ao uso da inteligência artificial. Com a disponibilização de mais dados, principalmente descrição de lesões e imagens de outros tipos de lesões, podemos utilizar o PubMedCLIP como uma ferramenta de MEDVQA e utilizá-lo para auxiliar em diagnósticos mais complexos.

Referências

- Schwendicke, F., Göstemeyer, G., Krois, J. (2020). "Artificial Intelligence in Dentistry: Chances and Challenges." *Journal of Dental Research*, 99(7), 769–774.
- Zou, Q., et al. (2021). "Deep Learning-based Artificial Intelligence in Dentistry." *International Journal of Oral Science*, 13(1), 1-8.
- Lee, J. H., et al. (2021). "Artificial Intelligence in Detecting Oral Cancer Using Digital Images: A Systematic Review and Meta-Analysis." *Oral Oncology*, 117, 105245.
- Li, Z., et al. (2022). "Multimodal Vision-Language Models for Disease Diagnosis: An Emerging Frontier in Medical AI." *IEEE Transactions on Medical Imaging*, 41(3), 623-633.
- INSTITUTO NACIONAL DE CÂNCER. Números de câncer. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros. Acesso em: 20 out. 2024.
- ESLAMI, S.; MEINEL, C.; DE MELO, G. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? *arXiv preprint arXiv:2112.13906*, 2022. Disponível em: https://arxiv.org/abs/2112.13906. Acesso em: 20 out. 2024.
- TANRIVER, G.; SOLUK TEKKESIN, M.; ERGEN, O. Automated Detection and Classification of Oral Lesions Using Deep Learning to Detect Oral Potentially Malignant Disorders. *Cancers*, v. 13, n. 27, 2021. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC8199603/#sec3-cancers-13-02766. Acesso em: 20 out. 2024.
- H S, Chandrashekar ; A, Geetha Kiran; S, Murali; M S, Dinesh; B R, Nanditha (2021), “Oral Images Dataset”, Mendeley Data, V2, Disponível em: https://data.mendeley.com/datasets/mhjyrn35p4/2. Acesso em: 22 out. 2024.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. Disponível em: https://arxiv.org/pdf/2103.00020. Acesso em: 23 out. 2024.
- BISHOP, Christopher M. *Pattern recognition and machine learning*. 1. ed. New York: Springer, 2006. (Information Science and Statistics). 738 p.