

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Felipe Elton Pazini Savi

Avaliação de um Grande Modelo de Linguagem Multimodal para Identificação de Diagnósticos Dermatológicos em Imagens Dermatoscópicas

Florianópolis
2026

Felipe Elton Pazini Savi

Avaliação de um Grande Modelo de Linguagem Multimodal para Identificação de Diagnósticos Dermatológicos em Imagens Dermatoscópicas

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciência da Computação.
Orientador: Prof. Aldo von Wangenheim, Prof. Dr. rer.nat.
Coorientador: Prof. Rodrigo de Paula e Silva Ribeiro, MSc.

Florianópolis
2026

Felipe Elton Pazini Savi

Avaliação de um Grande Modelo de Linguagem Multimodal para Identificação de Diagnósticos Dermatológicos em Imagens Dermatoscópicas

O presente trabalho em nível de Graduação foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Aldo von Wangenheim, Dr. Rer. nat.
Instituição Universidade Federal de Santa Catarina

Prof. Rodrigo de Paula e Silva Ribeiro, Msc.
Instituição Universidade Federal de Santa Catarina

Harley Miguel Wagner
Instituição Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Bacharel em Ciência da Computação.

_____

Profa. Lúcia Helena Martins Pacheco
Coordenadora do Curso

_____

Prof. Aldo von Wangenheim, Prof. Dr. rer.nat.
Orientador

Florianópolis, 2026.

RESUMO

O câncer de pele é o tipo de câncer mais comum globalmente. A sua detecção precoce é fundamental para um melhor prognóstico. Uma das principais formas de análise é por meio de imagens dermatoscópicas. Neste rabalho foi utilizado refinamento de prompt e injeção de informações adicionais utilizando o modelo MedGemma e um subonjunto do dataset ISIC para imagens dermatoscópicas. Como resultados, obteu-se acurácia no termo do diagnóstico de 33.86% no prompt puro e 50.51% no prompt com informações adicionais injetadas. Deste modo, este trabalho verificou como a injeção de informações extras contribuem um grande modelo de linguagem multimodal na análise de imagens dermatoscópicas.

Palavras-chave: dermatologia, dermatoscopia, descrição de imagens, llm, medgemma.

LISTA DE FIGURAS

Figura 1 – Da esquerda para a direita: Exame de dermatoscopia, imagem de aproximação e imagem panorâmica 15
Figura 2 – Imagem de dermatoscopia de um carcinoma basocelular (ISIC_0011403) 15
Figura 3 – Texto tokenizado. 17
Figura 4 – Ids dos tokens da figura 3. 17
Figura 5 – Arquitetura do transformador. 19
Figura 6 – Exemplo de relações traçadas pelo módulo *attention*. 20
Figura 7 – Arquitetura do ViT. 21
Figura 8 – Framework de (Cao et al., 2025) 22
Figura 9 – Framework de (Gajbhiye; Nandedkar; Faye, 2022) 23
Figura 10 – Framework de (Yang et al., 2025) 23
Figura 11 – Framework de (Mugdal Kush, 2024) 24
Figura 12 – Framework de (Lin et al., 2025) 25
Figura 13 – Framework de (Chakrabarti et al., 2025a) 25
Figura 14 – Processo de satização e processamento do conjunto de imagens da segunda etapa de (Evaristo, 2025) 31
Figura 15 – Arquitetura de (Chakrabarti et al., 2025b) 32
Figura 16 – Arquitetura de (Danu et al., 2023) 33
Figura 17 – Imagem de dermatoscopia do ISIC (ISIC_0011403) 35
Figura 18 – Informações disponibilizadas pelo ISIC (ISIC_0011403) 35
Figura 19 – Fluxo dos experimentos 37
Figura 20 – *Com Ground Truth*: *Output* para caso *ISIC_2382612* 38
Figura 21 – Informações no formato JSON (ISIC_001359) 38
Figura 22 – *Puro*: *General Diagnosis* 40

LISTA DE TABELAS

Tabela 1 – *String* genérica 28
Tabela 2 – *String* do IEEE 29
Tabela 3 – *String* do PubMed 29
Tabela 4 – *String* do ACM 29
Tabela 5 – *String* do ScienceDirect 30
Tabela 6 – Proporção de *outputs* válidos 38
Tabela 7 – Resultados por experimento. *Cat* e *Term* representam, respectivamente, *Diagnostic Category* e *Diagnostic Term* 40

SUMÁRIO

1 INTRODUÇÃO 8
1.0.1 Justificativa 9
1.1 OBJETIVOS 9
1.1.1 Objetivo Geral 9
1.1.2 Objetivos Específicos 9
2 METODOLOGIA 11
2.1 ETAPAS DA PESQUISA 11
3 FUNDAMENTAÇÃO TEÓRICA 14
3.1 LESÕES DE PELE 14
3.1.1 Imagens de Lesões de Pele 14
3.1.2 Câncer de Pele 15
3.1.3 Detecção e Diagnóstico de Câncer de Pele 16
3.2 GRANDES MODELOS DE LINGUAGEM 16
3.2.1 Tokenização 17
3.2.2 Incorporação Vetorial 18
3.2.3 Codificação Posicional 18
3.2.4 Transformadores 19
3.2.5 Modelos de Visão 21
3.2.6 Modelos Multimodais 22
3.3 DESCRIÇÃO DE IMAGENS MÉDICAS 22
3.4 DESCRIÇÃO DE LESÕES DE PELE 23
3.5 REVISÃO SISTEMÁTICA DA LITERATURA 25
3.5.1 Objetivos 26
3.5.2 Metodologia 26
3.5.3 Critérios de eligibilidade 27
3.5.3.1 Critérios de inclusão 27
3.5.3.2 Critérios de exclusão 28
3.5.4 Estratégia de busca 28
4 TRABALHOS CORRELATOS 31
4.1 FINE-TUNING DO MLLM LLAMA PARA A CLASSIFICAÇÃO DE LESÕES DE PELE E GERAÇÃO DE LAUDOS 31
4.2 AUTOMATED SKIN CANCER REPORT GENERATION VIA A KNOWLEDGE-DISTILLED VISION-LANGUAGE MODEL 31
4.3 GENERATION OF RADIOLOGY FINDINGS IN CHEST X-RAY BY LEVERAGING COLLABORATIVE KNOWLEDGE 32
5 DESENVOLVIMENTO 34
5.1 MATERIAIS 34

5.1.1 Tecnologias utilizadas **34**
5.1.2 medgemma-1.5-4b-it **34**
5.1.3 Conjunto de dados ISIC **34**
5.2 MÉTODOS DE AVALIAÇÃO 36
5.2.1 Relato de percepção experimental **36**
5.2.2 Experimentos **36**
5.2.3 Processamento do *output* da LLM **36**
5.2.4 Normalização dos campos **39**
5.3 RESULTADOS 39
6 CONCLUSÃO **42**
6.1 TRABALHOS FUTUROS 42
REFERÊNCIAS **43**
APÊNDICE A – LINKS **48**
APÊNDICE B – PROMPTS UTILIZADOS **49**
APÊNDICE C – ARTIGO **55**

1 INTRODUÇÃO

Nos últimos anos, a integração entre técnicas de visão computacional (computer vision) e grandes modelos de linguagem (large language models, LLMs) tem aberto novas fronteiras para o uso da inteligência artificial (IA) no diagnóstico médico a partir de imagens. A análise de imagens médicas, tradicionalmente realizada por especialistas humanos por meio de exames de imagem, desempenha papel crucial no diagnóstico, acompanhamento e tratamento de diversas doenças (Ahmed; Raja, J. Y.; Raja, M. Y. A., 2024).

No contexto da medicina diagnóstica, a aplicação de algoritmos de aprendizado profundo (deep learning) já demonstrou resultados promissores em tarefas como classificação, segmentação e detecção de lesões em imagens médicas (Souza et al., 2019). Entretanto, apesar dos avanços, muitos desses métodos ainda dependem de processamento exclusivamente visual, sem gerar uma descrição em linguagem natural, o que limita sua utilidade em laudos, explicações para médicos ou pacientes e sistemas de apoio à decisão que combinam texto e imagem.

Recentemente, surgiram abordagens multimodais: que combinam visão computacional com modelos de linguagem, permitindo que a IA interprete imagens médicas e produza descrições textuais, potencialmente semelhantes aos laudos médicos feitos por humanos. Essas abordagens buscam automatizar tarefas como identificação de características visuais relevantes (textura, forma, cor, padrão), contextualização clínica e geração de diagnóstico ou hipótese diagnóstica em linguagem natural, com aplicabilidade em especialidades como dermatologia (Lin et al., 2025).

Um exemplo recente e paradigmático desse avanço é o sistema SkinGPT-4 — um modelo de visão-linguagem (multimodal) treinado especificamente para diagnóstico dermatológico a partir de imagens de doenças de pele. O SkinGPT-4 foi desenvolvido combinando um “vision transformer” (modelo de visão) com um grande modelo de linguagem (LLM), e foi treinado em um conjunto extenso de imagens cutâneas — públicas e proprietárias — associadas a conceitos clínicos e anotações médicas (Zhou et al., 2024). Esse sistema exemplifica como LLMs adaptados à medicina podem converter informação visual em linguagem natural, gerando descrições médicas e hipóteses diagnósticas automatizadas.

Apesar dessas promessas, revisões recentes apontam que, embora em tarefas de Processamento de Linguagem Natural (NLP) os LLMs tenham alcançado desempenho satisfatório, os resultados em tarefas de processamento de imagem ainda estão longe do ideal em muitos casos. Isso ressalta a complexidade da interpretação médica de imagens e a necessidade de conjuntos de dados robustos, padronizados e bem anotados (Duan et al., 2025; Su et al., 2025).

1.0.1 Justificativa

O avanço dos modelos multimodais de inteligência artificial, capazes de integrar informações visuais e textuais, tem ampliado as possibilidades de aplicação da IA em diferentes áreas da saúde. Entre essas aplicações, destaca-se a análise de imagens médicas, na qual modelos de linguagem multimodais podem não apenas interpretar características visuais, mas também produzir respostas em linguagem natural contendo hipóteses diagnósticas e descrições clínicas. Na dermatologia, área fortemente dependente da avaliação visual de lesões cutâneas, esses modelos apresentam potencial para atuar como ferramentas de apoio à decisão clínica, auxiliando na identificação e caracterização de diferentes condições dermatológicas. Entretanto, ainda existem limitações relacionadas à confiabilidade, consistência terminológica e aderência a taxonomias diagnósticas padronizadas. Dessa forma, torna-se relevante investigar a capacidade desses modelos em identificar diagnósticos dermatológicos a partir de imagens dermatoscópicas, bem como avaliar o impacto da utilização de informações diagnósticas auxiliares no desempenho do modelo. Estudos dessa natureza contribuem para a compreensão das potencialidades e limitações dos Grandes Modelos de Linguagem multimodais aplicados à análise de imagens médicas, fornecendo evidências para futuras aplicações em sistemas de apoio ao diagnóstico.

1.1 OBJETIVOS

Nas seções abaixo, estão descritos o objetivo geral e os objetivos específicos.

1.1.1 Objetivo Geral

Avaliar a capacidade de um Grande Modelo de Linguagem multimodal na identificação de diagnósticos dermatológicos a partir de imagens dermatoscópicas, analisando seu desempenho na predição de diagnóstico geral, categoria diagnóstica e termo diagnóstico.

1.1.2 Objetivos Específicos

- Levantar e analisar conceitos relacionados a Grandes Modelos de Linguagem multimodais e suas aplicações na interpretação de imagens médicas.
- Selecionar e preparar um conjunto de imagens dermatoscópicas com informações diagnósticas de referência para a realização dos experimentos.
- Implementar experimentos utilizando o modelo MedGemma para análise automática de imagens dermatológicas em diferentes cenários de entrada.

- Comparar o desempenho do modelo em um cenário sem informações auxiliares e em um cenário com fornecimento prévio do diagnóstico geral da lesão.
- Avaliar a capacidade do modelo em predizer o diagnóstico geral, a categoria diagnóstica e o termo diagnóstico associados às lesões analisadas.
- Realizar a normalização semântica das respostas geradas pelo modelo para possibilitar a comparação com a taxonomia diagnóstica utilizada pelo conjunto de referência.
- Identificar limitações, desafios e potencialidades do uso de modelos multimodais de linguagem como ferramentas de apoio à análise dermatológica.

2 METODOLOGIA

A metodologia adotada neste Trabalho de Conclusão de Curso combina pesquisa aplicada, de caráter exploratório e descritivo, utilizando uma abordagem qualitativa e quantitativa. O objetivo é investigar modelos baseados em Grandes Modelos de Linguagem (LLMs) capazes de gerar descrições médicas a partir de imagens e avaliar sua utilidade na explicação de diagnósticos dermatológicos. Nesse sentido pode-se dizer que:

> Pesquisa Exploratória: Visa proporcionar maior familiaridade com o problema com vistas a torná-lo explícito ou a construir hipóteses. Envolvem levantamento bibliográfico; entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado; análise de exemplos que estimulem a compreensão. Assume, em geral, as formas de Pesquisa Bibliográficas e Estudos de Caso [...] realiza descrições precisas da situação e quer descobrir as relações existentes entre seus elementos [...] Recomenda-se a pesquisa exploratória quando há pouco sobre o problema a ser estudado. (Cervo; Bervian; Silva, 2009).

Do ponto de vista técnico, a pesquisa foi estruturada em três eixos principais: (i) revisão sistemática da literatura, (ii) definição e preparação do conjunto de dados e (iii) realização de experimentos computacionais com um modelo multimodal de visão e linguagem. A revisão sistemática constituiu a etapa inicial da pesquisa e teve como objetivo mapear o estado da arte da utilização de Grandes Modelos de Linguagem na descrição e interpretação de imagens médicas. Os resultados obtidos evidenciaram a crescente adoção de arquiteturas multimodais, capazes de integrar informações visuais e textuais em tarefas relacionadas ao diagnóstico e à geração de descrições clínicas. Com base nesse panorama, foi selecionado o modelo MedGemma para a etapa experimental, considerando que se trata de um modelo multimodal desenvolvido especificamente para aplicações na área médica e treinado com dados do domínio da saúde. A partir dessa definição, foram conduzidos experimentos utilizando imagens dermatoscópicas do conjunto ISIC para avaliar a capacidade do modelo em inferir informações diagnósticas a partir do conteúdo visual das lesões.

2.1 ETAPAS DA PESQUISA

A execução do trabalho foi organizada em etapas sequenciais:

1. Revisão da literatura
    - Condução de uma revisão sistemática sobre a aplicação de Grandes Modelos de Linguagem e modelos multimodais na análise e descrição de imagens médicas.
    - Identificação das principais arquiteturas utilizadas na integração entre visão computacional e linguagem natural.

- Levantamento de conjuntos de dados e estratégias de avaliação empregadas na literatura.

2. Definição e preparação do conjunto de dados
    - Seleção do conjunto de dados ISIC para realização dos experimentos.
    - Filtragem das imagens dermatoscópicas de contato não polarizadas.
    - Organização das imagens e metadados diagnósticos em formato adequado para processamento automatizado.
    - Verificação da consistência das informações diagnósticas utilizadas como referência.
3. Seleção e preparação do modelo
    - Escolha do modelo multimodal MedGemma, desenvolvido para aplicações médicas.
    - Configuração do ambiente computacional e das bibliotecas necessárias para execução do modelo.
    - Definição do protocolo de inferência e dos formatos de saída utilizados nos experimentos.
4. Realização dos experimentos
    - Execução do modelo MedGemma sobre as imagens selecionadas sem fornecimento de informações diagnósticas auxiliares.
    - Execução do modelo MedGemma com fornecimento prévio do diagnóstico geral da lesão.
    - Extração das informações de diagnóstico geral, categoria diagnóstica e termo diagnóstico geradas pelo modelo.
    - Armazenamento das respostas para posterior análise.
5. Normalização e avaliação dos resultados
    - Extração dos valores únicos produzidos pelo modelo para categorias e termos diagnósticos.
    - Construção de uma tabela de equivalência para normalização semântica das respostas.
    - Comparação dos resultados normalizados com o ground truth disponível no conjunto de dados.
    - Cálculo das métricas de acurácia para diagnóstico geral, categoria diagnóstica e termo diagnóstico.
6. Consolidação dos resultados e escrita do TCC

- Organização e interpretação dos resultados experimentais.
- Discussão dos resultados à luz da literatura revisada.
- Elaboração das conclusões, limitações do estudo e sugestões para trabalhos futuros.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são detalhados conceitos introdutórios e fundamentos teóricos necessários para a boa compreensão desta pesquisa, bem como a apresentação da revisão sistemática que originou boa parte da sustentação teórica.

3.1 LESÕES DE PELE

A pele é o maior órgão do corpo humano, com diversas funções (Jablonski, 2004):

- Proteger o corpo de agentes externos;
- Contribuir na regulação da temperatura do corpo;
- Proporcionar informações sensoriais como o tato;

Devido à exposição da pele ao ambiente, é mais comum que esse órgão sofra com doenças. As áreas afetadas são consideradas lesões de pele e podem ser usadas para diagnósticos de doenças (Sumithra; Suhil; Guru, 2015).

Vários fatores são considerados no diagnóstico de uma doença de pele. Habif (2015) recomenda que sejam considerados aspectos como o histórico da lesão, distribuição, características visuais e testes laboratoriais. Porém, o diagnóstico de uma doença de pele ainda é desafiador. Uma mesma doença de pele pode apresentar diferentes sintomas em cada instância e também pode evoluir ao longo do tempo. A análise visual é o principal componente no diagnóstico de lesões de pele. Nesse processo, a lesão é analisada pelas suas características superficiais, como a cor, textura e presença de certas estruturas.

3.1.1 Imagens de Lesões de Pele

O caráter visual da análise de lesões de pele permite que imagens sejam utilizadas em exames dermatológicos. Essas imagens podem ser agrupadas em diferentes categorias. Alguns exemplos baseados em procedimentos recomendados por (Von Wangenheim; Wagner; Nogueira, 2013) são:

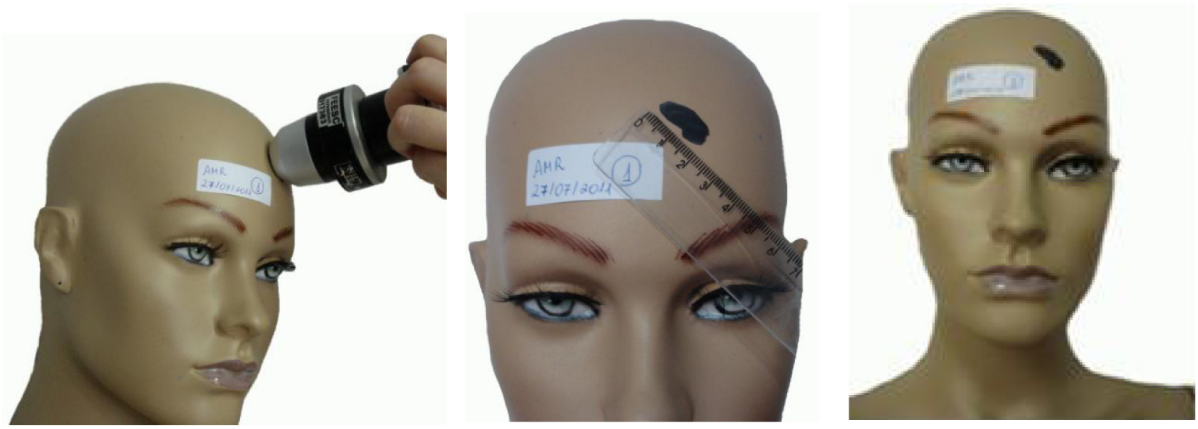
- Dermatoscopia: Estas imagens são obtidas com um equipamento especializado, o dermatoscópio. Esse método permite que um diagnóstico mais preciso seja feito, sendo melhor que o olho nu na detecção de melanomas (Bono et al., 2002);
- Foto de aproximação com régua: Esse tipo de imagem é obtida com uma fotografia feita a 30 centímetros da lesão, sem utilizar zoom. Nessas imagens, etiquetas são colocadas próximas à lesão para auxiliar na determinação do seu tamanho;

- Foto panorâmica: Nesse método são registradas imagens de regiões do corpo, como da cabeça, tronco, braços e pernas.

Estes exemplos estão ilustrados na figura 1.

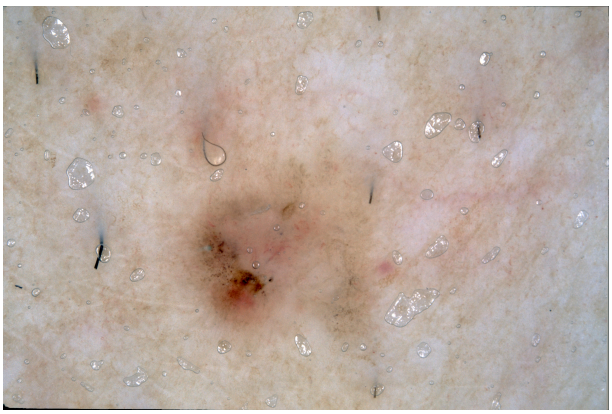
Um exemplo de imagem gerada a partir de um exame de dermatoscopia é a figura 2.

Figura 1 – Da esquerda para a direita: Exame de dermatoscopia, imagem de aproximação e imagem panorâmica



Fonte: (Von Wangenheim; Wagner; Nogueira, 2013)

Figura 2 – Imagem de dermatoscopia de um carcinoma basocelular (ISIC_0011403)



Fonte: (ISIC, 2026)

3.1.2 Câncer de Pele

O câncer é uma doença caracterizada pela multiplicação de células anormais que podem se espalhar para além do seu tecido de origem, causando tumores e

levando eventualmente à morte (Organization, 2022). O câncer de pele é o tipo mais comum da doença globalmente, com principal causa sendo a exposição prolongada à radiação ultravioleta (Craythorne Emma; Al-Niami, 2017). Em geral, essa doença afeta mais a pele clara e pode afetar a mesma pessoa mais de uma vez. Uma vez desenvolvido, há um aumento de 35% no risco de desenvolvimento de um novo câncer de pele do mesmo tipo em um período de três anos (Zink, 2014).

Os cânceres de pele são classificados em cânceres não melanoma (CPNM) e melanoma. CPNM podem ser subdivididos em carcinoma basocelular (CBC), carcinoma epidermoide (CEC), carcinoma de Merkel, entre outros. Os CPNM são os mais incidentes, correspondendo a mais de 90% dos cânceres diagnosticados, e também são os menos fatais. O subtipo CBC é o mais frequente e corresponde a mais de 75% dos casos de CPNM no Brasil (Zink, 2014). O melanoma, embora menos incidente, possui o maior grau de letalidade: cerca de 75% das mortes por câncer de pele são causadas por melanomas (Mauad, 2011).

Um exemplo de carcinoma basocelular está indicado na figura 2.

Um melhor prognóstico para esta doença está diretamente relacionado à detecção precoce para um encaminhamento adequado. Segundo (Lima Vazquez, 2015), a taxa de sobrevivência ao melanoma no Brasil é menor que a taxa global, sendo que há uma prevalência maior de casos avançados.

3.1.3 Detecção e Diagnóstico de Câncer de Pele

O câncer de pele pode ser identificado e classificado baseado nas características da lesão, incluindo o tamanho, a variação de cor, irregularidade do formato, evolução ao longo do tempo e a localização no corpo (Jones, 2020).

A análise desses sintomas é normalmente realizado visualmente por um dermatologista, sendo que técnicas como a dermatoscopia e inspeção da área da lesão podem ser realizadas no contexto da teledermatologia. Entre outros métodos não invasivos incluem-se: análise intracutânea espectrofotométrica, ultrasonografia de alta frequência e microscopia confocal de refletância. Também é possível diagnosticar a doença através de métodos invasivos como a biópsia.

Além destes métodos, também é possível utilizar Inteligência Artificial (IA) no diagnóstico. Soluções com IA utilizam imagens para tanto classificar quanto descrever a doença e podem ter uma precisão equiparável ou até maior a de dermatologistas (Jones, 2020).

3.2 GRANDES MODELOS DE LINGUAGEM

LLMs são modelos estatísticos pré-treinados que utilizam transformadores com bilhões de parâmetros, capazes de processar texto em linguagem natural.

Um grande marco no surgimento das LLMs foi o artigo (Vaswani et al., 2017), que apresentou o conceito de transformadores, tornando viável a sua implementação prática.

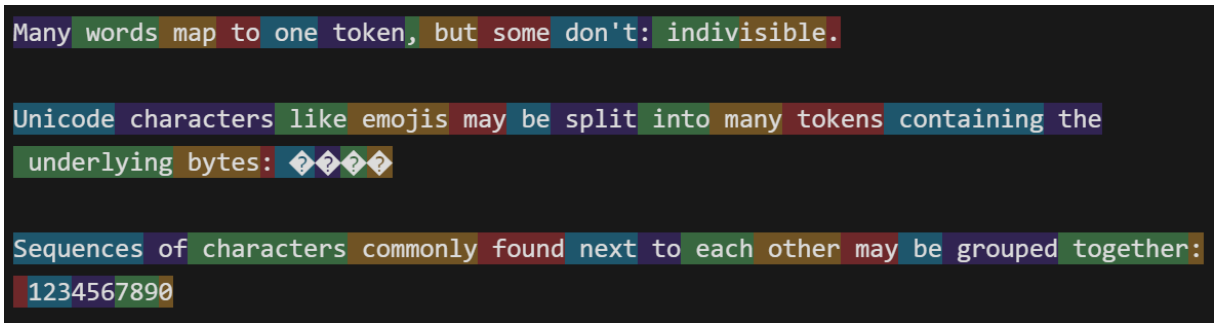
Inicialmente, o seu propósito era apenas traduzir textos. Porém, a partir de pesquisas da OpenAI no modelo GPT, demonstrou-se a capacidade do modelo de responder perguntas que não estavam presentes na base de treinamento. Os modelos posteriores, GPT-2 e GPT-3, são capazes de traduzir textos, resolver problemas básicos, interpretar texto e computar aritmética básica.

3.2.1 Tokenização

A tokenização é o processo em que textos são quebradas em constituentes menores, podendo ser palavras inteiras, ou até mesmos partes de palavras. Isso permite que o modelo aprenda novas palavras a partir da composição de palavras anteriores. Além disso, isso aumenta a eficiência do texto, reduzindo a quantidade memória, visto que palavras grandes podem ser representadas por ids menores. Na figura 3, é possível ver um exemplo de texto tokenizado.

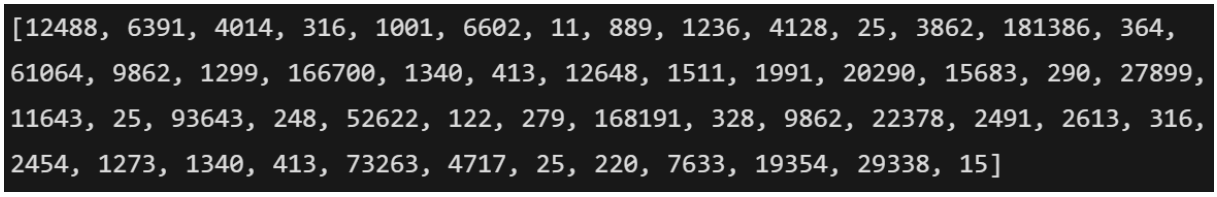
Na figura 4, é possível observar os IDs relativos a cada um dos tokens da imagem 3.

Figura 3 – Texto tokenizado.



Fonte: (OpenAI, 2025)

Figura 4 – Ids dos tokens da figura 3.



Fonte: (OpenAI, 2025)

Os tipos mais comuns de tokenização, conforme descrito em (Minaee et al., 2025), são:

- **Byte pair encoding:** O texto é quebrado considerando a frequência das palavras. Dessa forma, palavras com maior frequência são mantidas inteiras, enquanto que palavras menos frequêntes são quebradas.
- **Word piece encoding:** O algoritmo inicializa com tokens para os símbolos do alfabeto, e sucessivamente combina os tokens para formar palavars maiores, até haver tokens para todas as palavras.
- **Sentence piece encoding:** Utilizado para textos que não tenham palavras delimitadas por espaços, como por exemplo no japonês e mandarim.

3.2.2 Incorporação Vetorial

A incorporação vetorial é um vetor n-dimensional de números reais que representa dados variados, como por exemplo, palavras, tokens ou imagens. Ele permite operar dados diferentes utilizando uma única representação.

A incorporação vetorial de palavras associa uma semântica às suas direções. Portanto, se duas palavras possuem significados relacionados, elas apontam em direções semelhantes. Senão, elas são perpendiculares.

3.2.3 Codificação Posicional

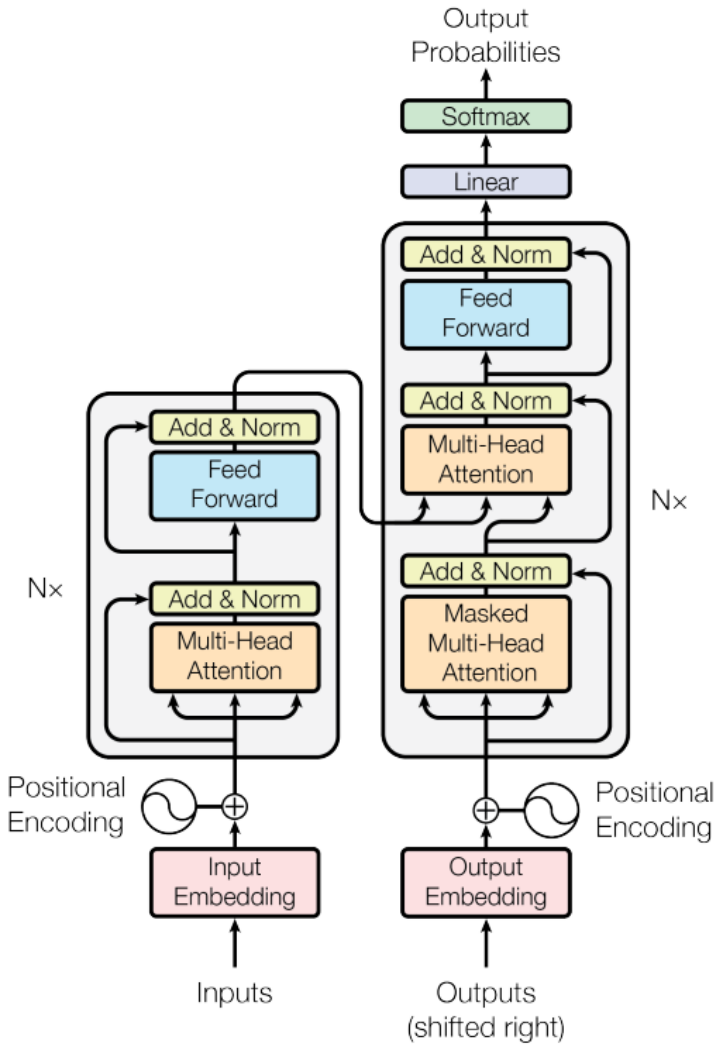
Codificação posicional é a forma de representar a posição do token no texto de entrada. A codificação original é conhecida por Absolute Positional Embedding (APE), que utiliza uma combinação de funções seno e cosseno com diferentes frequências. Esse sistema de codificação possui desvantagens significativas, como por exemplo a limitação no número de incorporações e incapacidade de lidar com distâncias relativas. Portanto, foram criadas outras formas de codificação, conforme descrito em (Minaee et al., 2025):

- **RPE:** Texto de entrada é representado por tokens conectados entre si, formando grafo dirigido conexo. Esse sistema é implementado no mecanismo de atenção do transformador. Portanto, possui a desvantagem de tornar o treinamento e a inferência mais lentos.
- **RoPE:** Aplica rotação no ângulo dos pares de vetores de *query* e *key*.
- **Vieses posicionais relativos:** Introduz vieses no nível de atenção entre *query* e *key*, penalizando os pares distantes entre si.

3.2.4 Transformadores

O transformador é responsável por receber tokens anteriores do texto e prever o próximo token. A versão moderna do transformador foi apresentada primeiramente no artigo (Vaswani et al., 2017). O transformador é composto por um codificador e um decodificador, cada um deles sendo composto por várias camadas, conforme detalhado na figura 5.

Figura 5 – Arquitetura do transformador.



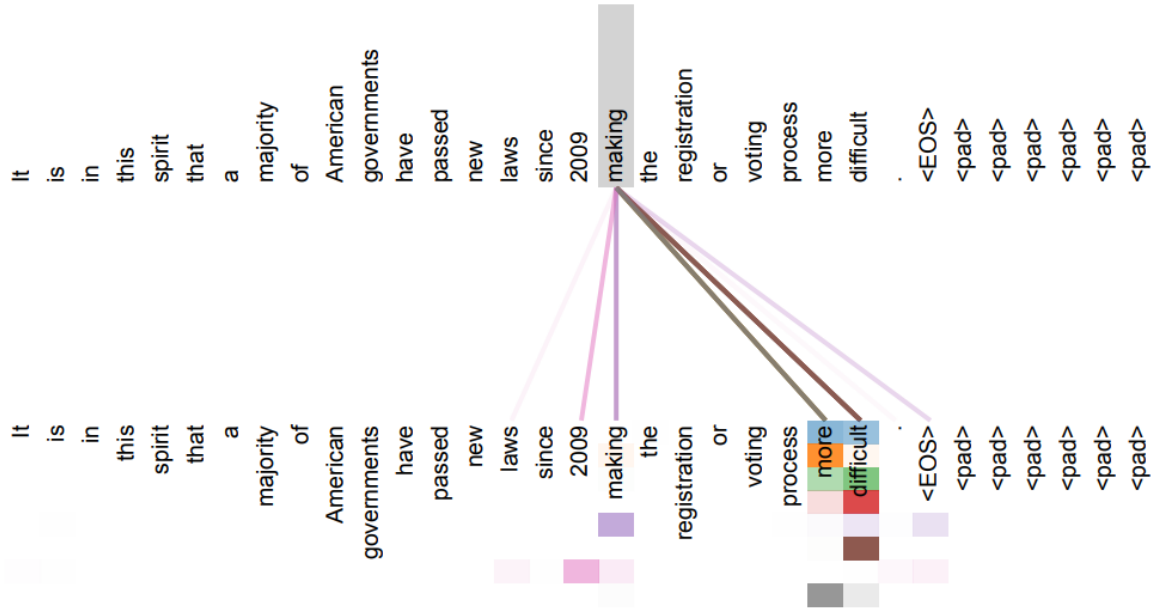
Fonte: (Vaswani et al., 2017)

No codificador, cada camada é formada por dois componentes, o *multi-head self-attention* e uma rede *feed-forward*. A saída desses componentes possui uma conexão residual e uma etapa de normalização. O decodificador possui os mesmos componentes do codificador, porém antes é aplicado um mecanismo de *multi-head*

self-attention mascarado. Além disso, a entrada do decodificador é deslocada em uma posição, para permitir que ele possa prever o próximo token.

O sistema de *multi-head self-attention* permite traçar relações entre palavras diferentes do texto. No exemplo da figura 6, é possível perceber que o conjunto de cabeças aprendeu a relação entre a palavra *making* e as palavras *more difficult*.

Figura 6 – Exemplo de relações traçadas pelo módulo *attention*.



Fonte: (Vaswani et al., 2017)

O modelo de atenção chama-se "Scaled Dot-Product Attention". Primeiramente, são calculados três vetores para cada incorporação vetorial, *query* (V_Q), *key* (V_K) e *value* (V_V). Esses vetores são agrupados na matriz E que é em seguida multiplado pelas matrizes W_Q, W_K e W_V, obtidas a partir do treinamento do modelo. Isso resulta nas matrizes Q, K e V. As equações estão descritas abaixo: $E \times W_Q = Q$ $E \times W_K = K$ $E \times W_V = V$

A segunda etapa é calcular o produto escalar dos vetores de V_Q com os vetores de V_K, por meio da operação $Q \times K^T$. O resultado indica o quanto cada vetor de V_K atende cada V_Q, mostrando se há relação entre eles.

Em seguida, a matriz resultante é dividida por $\sqrt{d_k}$, sendo d_k o número de dimensões, a fim de aumentar a estabilidade do gradiente, e os valores são normalizados por meio da função *softmax*.

No fim, é multiplicado por V, aplicando o mecanismo de ação à matriz efetiva.

Isso é descrito pela fórmula: $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

Os vetores gerados por cada cabeça de atenção são concatenados e multiplicados pela matriz W_O, formando a saída da subcamada de atenção:

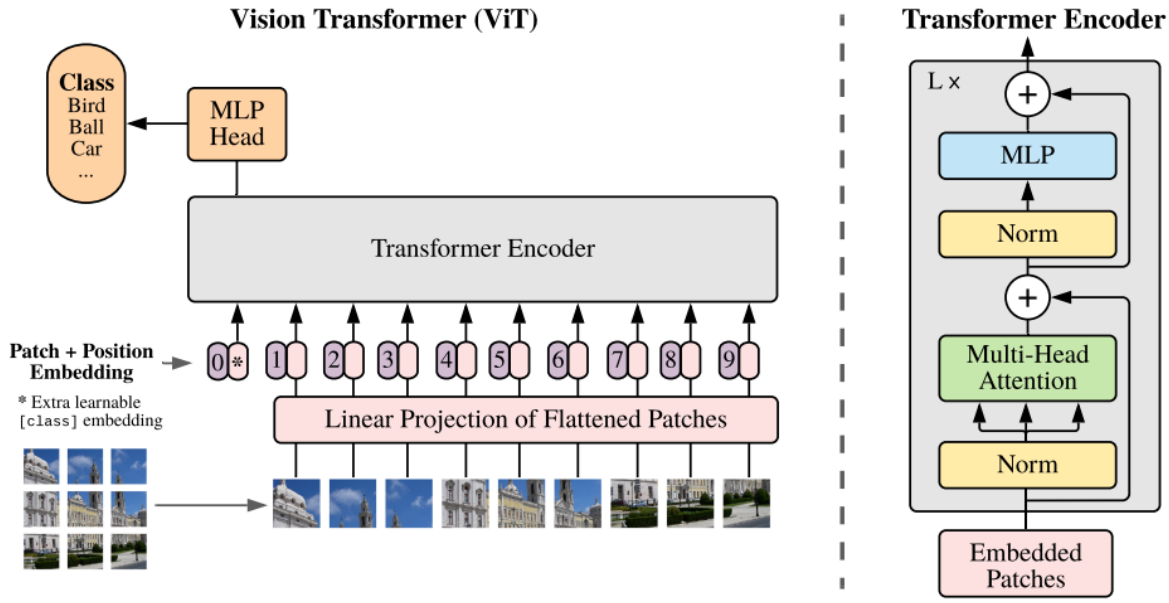
$MultiHead(Q, K, V) = Concat(head_1, ..., head_k)W_O$

3.2.5 Modelos de Visão

Modelos de visão são modelos que operam em imagens como entrada. No contexto de LLMs, um marco no seu desenvolvimento foi o trabalho "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"de Dosovitskiy et al. (2021), que introduziu os transformadores visuais.

A arquitetura do ViT está descrita na figura 7.

Figura 7 – Arquitetura do ViT.



Fonte: (Dosovitskiy et al., 2021)

Os ViTs funcionam primeiramente dividindo a imagem em seções de menor resolução, que são posteriormente planificadas e incorporadas. Cada uma das incorporações geradas recebe uma codificação posicional.

Essas incorporações são utilizadas para treinar o transformador.

O funcionamento é semelhante ao descrito no artigo (Vaswani et al., 2017), diferenciando na aplicação da normalização antes do mecanismo *multi-head attention*.

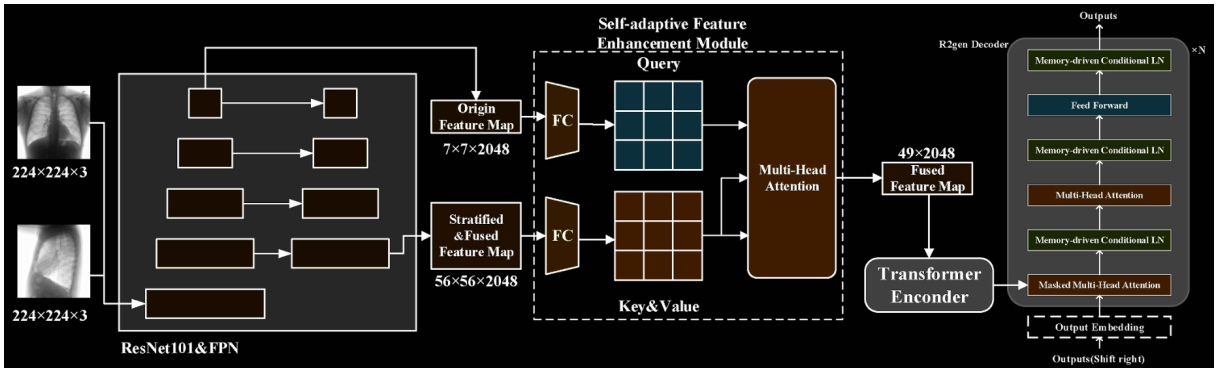
3.2.6 Modelos Multimodais

3.3 DESCRIÇÃO DE IMAGENS MÉDICAS

Antes do surgimento dos transformadores, já era possível classificar imagens utilizando redes neurais convolucionals (CNNs). Porém, após os transformadores, tornou-se possível adicionar descrições às imagens em linguagem natural, e até mesmo gerar relatórios médicos.

Um dos estudos que realiza isso é (Cao et al., 2025), que propõe um framework para gerar descrições de imagens de radiografia de tórax, o *Stratified Adaptive Enhanced Net* (SAENet). O framework está representado na figura 8. Inicialmente, a partir do ResNet101, é criada uma rede *Feature Pyramid Network*, para extrair tanto *features* de alto nível quanto de baixo nível. Em seguida, os dados passam por um sistema de *multi-head attention*, para aprimorar as *features*. Por fim, é utilizado um componente de geração de relatório médico, com *encoder* baseado no modelo de (Vaswani et al., 2017) e *decoder* baseado no modelo de (Chen et al., 2020). Testando nos *datasets* IU X-ray e Peir Gross, foi obtido um *BLEU score* de 0.500.

Figura 8 – Framework de (Cao et al., 2025)

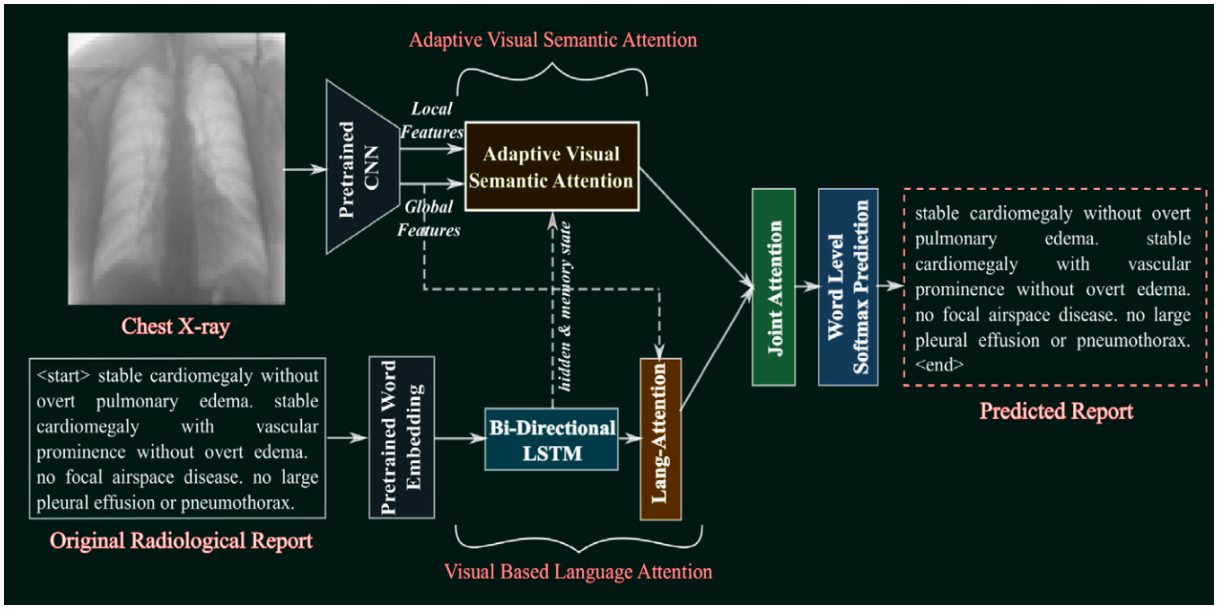


Fonte: (Cao et al., 2025)

Em (Gajbhiye; Nandedkar; Faye, 2022), é proposto um sistema de *Adaptive Multilevel Multi-Attention (AMLMA)* para gerar relatórios de radiografia de tórax. Na figura 9, é possível perceber que o *encoder* de imagem utiliza um *residual attention block* para extrair *features* tanto locais quanto globais. Testando no dataset IU-CXR, foi obtido *BLEU score* de até 0.471.

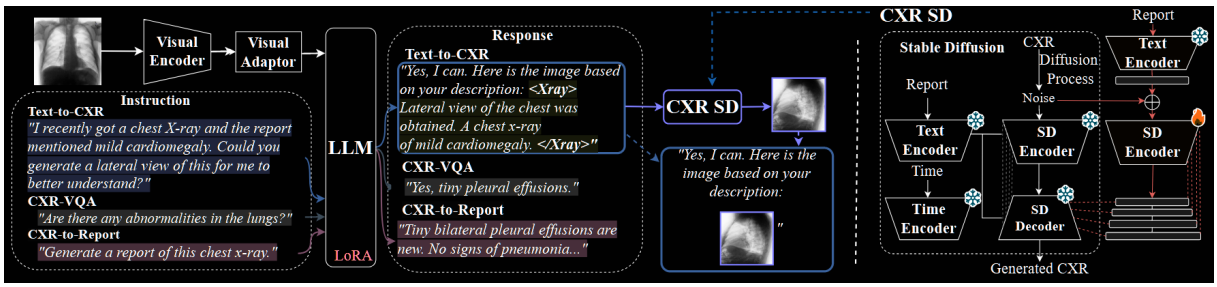
Em (Yang et al., 2025), é proposto um framework que permite gerar relatórios de radiografia de tórax, representado na figura 10. É utilizado o framework de LLM mPLUG-Owl. Para realizar o *fine tuning* do modelo, foi utilizado *Low-Rank Adaptation* (LoRA). Antes de ser processada pela LLM, a imagem é processada por um *visual encoder CLIP*.

Figura 9 – Framework de (Gajbhiye; Nandedkar; Faye, 2022)



Fonte: (Gajbhiye; Nandedkar; Faye, 2022)

Figura 10 – Framework de (Yang et al., 2025)



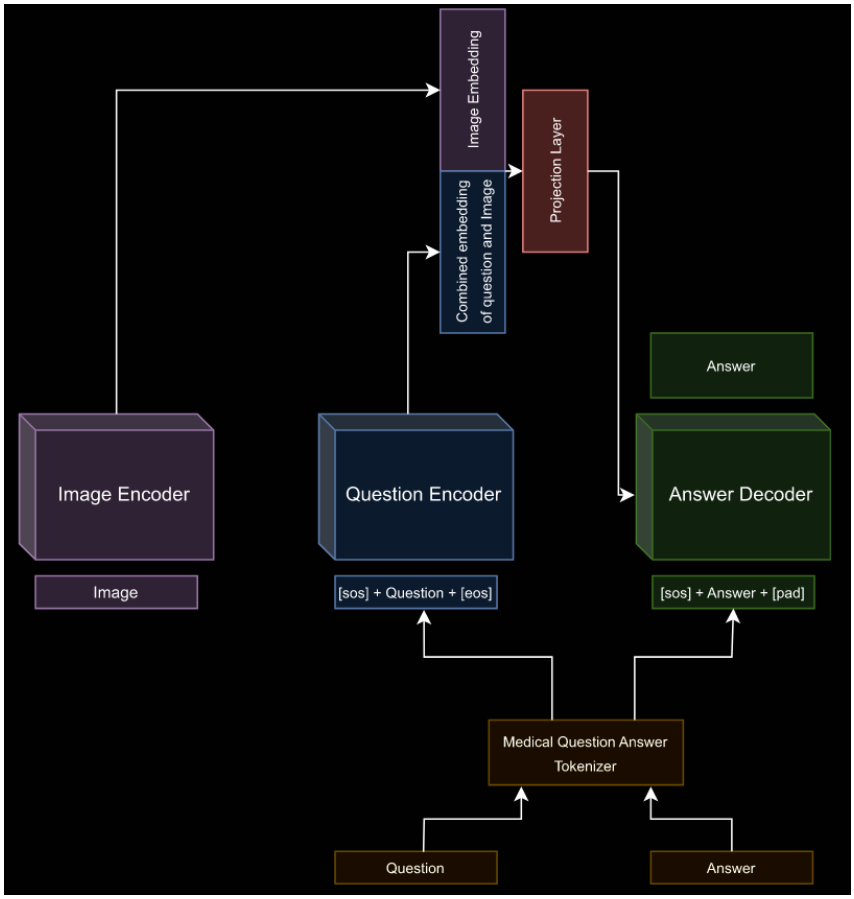
Fonte: (Yang et al., 2025)

Em (Mugdal Kush, 2024), é apresentado um sistema de *visual question answering* (VQA) para imagens médicas. O fluxo está descrito na figura (Mugdal Kush, 2024), inspirado no BLIP. A imagem e o texto são codificados separadamente utilizando, respectivamente, mecanismos de *self attention* e *multi head attention*. Em seguida, passam por um *projection layer* e pelo decodificador, gerando a resposta do VQA. Testando a arquitetura nos datasets ImageCLEF 2019 e VQA-RAD, o *BLEU score* foi 0.41.

3.4 DESCRIÇÃO DE LESÕES DE PELE

Em (Lin et al., 2025), é apresentado um framework para geração de imagens dermatológicas e relatório. Para este último, conforme é possível observar na figura

Figura 11 – Framework de (Mugdal Kush, 2024)



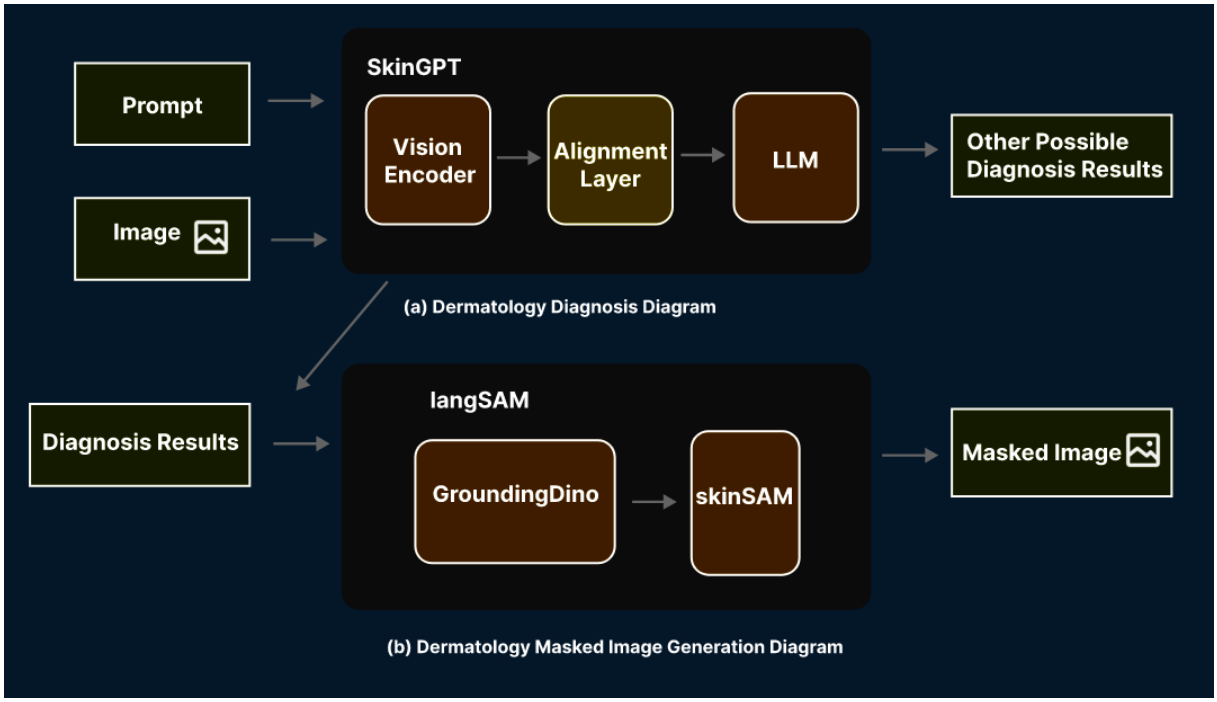
Fonte: (Mugdal Kush, 2024)

12, foi utilizado o SkinGPT-4, que é formado por três módulos, o visual encoder, o projection layer e o advanced large language decoder. O visual encoder é composto de Vision Transformer e Q-Transformer, que incorporam a imagem. O projection layer alinha o texto com a imagem. Já o large language decoder utiliza Vicuna e Llama.

Em (Doshi, 2025), é realizado o retrieval de textos a partir de imagens dermatológicas. O modelo utilizado foi o CLIP, com fine-tuning utilizando o dataset SkinCAP.

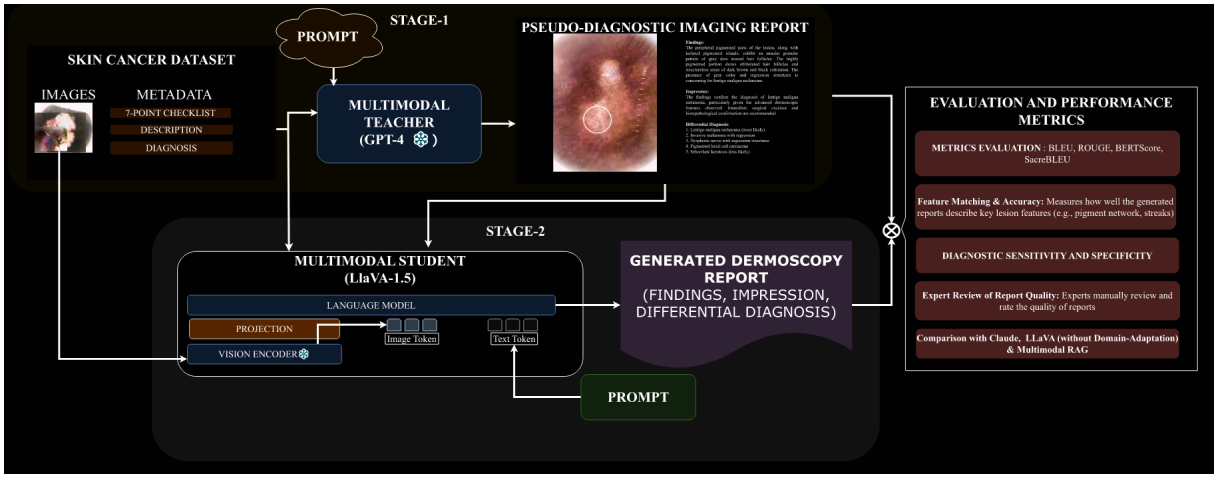
Em (Chakrabarti et al., 2025a), é apresentado um framework de geração de relatório. Conforme é possível observar na figura 13, é utilizado uma configuração professor-estudante, em que o modelo de visão proprietário *GPT-4V* age como professor, enquanto que o modelo *open-source LLaVA-v1.5-13B* age como estudante. No estágio 1, o modelo professor gera relatórios estruturados a partir de imagens dermatológicas. No estágio 2, o modelo estudante realiza *fine-tuning* a partir das imagens dermatológicas e os relatórios gerados pelo modelo professor.

Figura 12 – Framework de (Lin et al., 2025)



Fonte: (Lin et al., 2025)

Figura 13 – Framework de (Chakrabarti et al., 2025a)



Fonte: (Chakrabarti et al., 2025a)

3.5 REVISÃO SISTEMÁTICA DA LITERATURA

Para fortalecer a fundamentação teórica e explorar as tecnologias e métodos utilizados atualmente para geração de descrição de imagens médicas, foi realizada uma revisão sistemática da literatura.

3.5.1 Objetivos

Essa revisão sistemática tem como objetivo identificar, descrever e classificar pesquisas existentes de abordagens computacionais que produzem saída textual a partir de imagens. Mais especificamente, o objetivo da revisão é:

- Mapear os métodos e modelos usados para adicionar legendas a imagens médicas, geração de relatórios, sistemas de *visual question answering* (VQA) e explicações textuais;
- Examinar as modalidades de imagens médicas consideradas, com foco em particular nas imagens dermatológicas e de teledermatologia;
- Resumir as estratégias de avaliação, *datasets*, e métricas de performance usadas nesses estudos;
- Identificar lacunas e oportunidades no desenvolvimento e validação clínica de sistemas que explicam ou justificam diagnósticos baseados em imagens por meio da linguagem.

3.5.2 Metodologia

Esta revisão foi projetada e conduzida como uma revisão de escopo, adaptando as diretrizes propostas por (Kitchenham; Charters, 2007) e as diretrizes PRISMA (*Preferred Reporting Items for System- atic Reviews and Meta-Analyses*) para transparência na busca (Moher et al., 2009). A elegibilidade dos estudos foi guiada pelo framework População–Conceito–Contexto (PCC), recomendado para revisões de escopo.

- População (P): Imagens médicas de qualquer modalidade, incluindo fotografias clínicas, imagens dermatoscópicas, radiologia, patologia ou outras fontes de imagens de diagnóstico.
- Conceito (C): Métodos e modelos computacionais que geram saídas em linguagem natural a partir de imagens — como geração de legenda de imagens, geração de relatórios clínicos, *visual question answering* (VQA) ou raciocínios textuais que explicam ou justificam decisões baseadas em imagens.
- Contexto (C): Ambientes de pesquisa biomédica e clínica, com foco específico em dermatologia e teledermatologia, mas também abrangendo outros domínios de imagem médica onde técnicas de geração de texto são avaliadas ou comparadas.

Esse framework está sendo utilizado para definir os critérios de inclusão e exclusão, desenvolver a estratégia de busca e estruturar o processo de extração de dados. O objetivo é mapear a literatura científica que aborda métodos computacionais

que geram descrições ou explicações textuais a partir de imagens médicas, com ênfase em aplicações em dermatologia e teledermatologia.

Estudos relevantes foram identificados por meio de buscas abrangentes em grandes bases de dados bibliográficas, incluindo PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, sem restrição quanto ao desenho do estudo. As estratégias de busca combinaram termos de vocabulário controlado (por exemplo, MeSH) e palavras-chave em texto livre relacionadas a imagem médica, geração de texto (por exemplo, "image captioning", "report generation", "vision question answering") e inteligência artificial (por exemplo, "deep learning", "large anguage model", "vision-language model").

Os registros obtidos foram exportados, mesclados e deduplicados antes da triagem. Dois revisores estão avaliando independentemente títulos e resumos de acordo com critérios de inclusão e exclusão predefinidos, com discrepâncias resolvidas por consenso. A triagem do texto completo será realizada em registros elegíveis, e os dados serão extraídos para uma forma estruturada, incluindo tipo de estudo, modalidade de imagem, tarefa textual, arquitetura de modelo, métricas de avaliação e relevância específica para dermatologia. Os resultados serão resumidos descritivamente para destacar tendências, conjuntos de dados, características metodológicas e lacunas na pesquisa atual.

3.5.3 Critérios de eligibilidade

3.5.3.1 Critérios de inclusão

- Artigo apresenta ou avalia um método, modelo ou abordagem que gera descrições textuais a partir de imagens médicas.
- Estudos envolvendo imagens médicas de qualquer modalidade, com interesse particular em dermatologia e teledermatologia.
- O método produz saída em linguagem natural, em vez de apenas rótulos de classificação ou mapas de calor visuais.
- Métodos baseados em aprendizado de máquina, aprendizado profundo, grandes modelos de linguagem (LLMs), modelos de visão-linguagem (VLMs) ou outras técnicas de inteligência artificial multimodal.
- Artigos de pesquisa original, artigos sobre conjuntos de dados, relatórios técnicos ou pré-impressões descrevendo um método ou conjunto de dados implementado.
- Artigos escritos em inglês.
- Estudos publicados no intervalo de 2022-2025.

3.5.3.2 Critérios de exclusão

- Estudos que apenas realizam classificação, detecção ou segmentação de imagens sem qualquer saída textual ou linguística.
- Artigos editoriais, comentários, peças de opinião, revisões ou trabalhos puramente teóricos sem um sistema ou conjunto de dados implementado.
- Artigos que se concentram em dados não visuais.
- Artigos que focam em pergunta e resposta sem geração de relatório estruturado ou uma explicação textual aprofundada.
- Estudos nos quais a descrição textual é produzida manualmente ou não gerada computacionalmente.
- Estudos pedagógicos ou de interface que não descrevem ou avaliam um método computacional para geração de texto.
- Texto não totalmente disponível.

3.5.4 Estratégia de busca

Foi desenvolvida uma *string* genérica de busca, que foi adaptada para cada base de dados especificamente.

Generic String
(("medical image*"OR "dermatology"OR "dermoscopy"OR "skin lesion*"OR "teledermatology"OR "telemedicine") AND ("image caption*"OR "report generation"OR "image-to-text"OR "visual question answering"OR "VQA"OR "textual explanation"OR "rationale"OR "vision-language"OR "multimodal"OR "description") AND ("machine learning"OR "deep learning"OR "artificial intelligence"OR "large language model"OR "vision-language model"))

Tabela 1 – *String* genérica

IEEEXplore	
String	(("medical image*"OR "dermatology"OR "dermoscopy"OR "skin lesion*") AND ("image caption*"OR "report generation"OR "image-to-text"OR "visual question answering"OR "VQA"OR "textual explanation"OR "vision-language"OR "multimodal"OR "image description*") AND ("machine learning"OR "deep learning"OR "artificial intelligence"OR "large language model"OR "vision-language model"))
Filters	Range: 2022 to 2025 Language: English Publication Type: Conferencs, Journals
Results	787

Tabela 2 – *String* do IEEE

PubMed	
String	(("Image Interpretation, Computer-Assisted"[MeSH] OR "Natural Language Processing"[MeSH] OR caption*[tiab] OR "report generation"[tiab] OR "image-to-text"[tiab] OR "vision-language"[tiab] OR "vision language"[tiab] OR "VQA"[tiab] OR (explain*[tiab] AND (text*[tiab] OR rationale*[tiab])))) AND (dermatology[MeSH Terms] OR dermatolog*[tiab] OR dermoscop*[tiab] OR "skin lesion*"[tiab] OR teledermatolog*[tiab] OR teledermatology[tiab] OR telemedicine[MeSH Terms]) AND ("machine learning"[tiab] OR "deep learning"[tiab] OR "large language model*"[tiab] OR "vision-language model*"[tiab] OR "multimodal"[tiab] OR "foundation model*"[tiab]) NOT (editorial[pt] OR comment[pt])
Filters	Range: 2022/1/1 to 2025/12/31 Language: English
Results	96

Tabela 3 – *String* do PubMed

ACM	
String	(("image captioning"OR "report generation"OR "visual question answering"OR VQA OR "textual rationale"OR "explainable"NEAR/2 (text OR language) OR "vision-language"OR multimodal) AND (dermatology OR dermoscopy OR "skin lesion"OR teledermatology OR "medical imaging"))
Filters	Search Type: Abstract Year: 2022-2025 Content Type: Research Article Bases: ACM Guide to Computing Literatura and ACM Full-Text Collection
Results	109

Tabela 4 – *String* do ACM

ScienceDirect	
String	(("medical image"OR "dermatology") AND ("report generation"OR "visual question answering"OR "textual explanation") AND ("deep learning"OR "large language model"OR "vision-language model"))
Filters	Search Type: Title, abstract or author-specified keywords Article Type: Research Article Year: 2022-2025
Results	36

Tabela 5 – *String* do ScienceDirect

4 TRABALHOS CORRELATOS

Neste capítulo, são apresentados trabalhos correlatos que realizam a classificação de imagens médicas a partir de LLMs.

4.1 FINE-TUNING DO MLLM LLAMA PARA A CLASSIFICAÇÃO DE LESÕES DE PELE E GERAÇÃO DE LAUDOS

Este trabalho realiza o *fine-tuning* do LLaMA 3.2 11B utilizando as técnicas QLoRA e LoRA para classificar imagens de lesão de pele, testado em duas etapas.

Na primeira etapa, o modelo foi treinado e testado sobre o conjunto de dados HAM10000, que possui 13354 imagens de dermatoscopia. Essas imagens foram classificadas em 7 categorias de lesões de pele distintas. O modelo obteve uma acurácia de 86.2%.

Na segunda etapa, o modelo foi treinado e testado sobre o conjunto de imagens de aproximação do Sistema Integrado Catarinense de Telemedicina e Telessaúde. Essas imagens, por sua vez, foram classificadas entre 34 classes distintas. O modelo obteve uma acurácia de 45.2%. O processo de satização e processamento do conjunto de imagens da segunda etapa é representado na figura 14

4.2 AUTOMATED SKIN CANCER REPORT GENERATION VIA A KNOWLEDGE-DISTILLED VISION-LANGUAGE MODEL

O artigo propõe um modelo multimodal de visão-linguagem treinado por meio de um *framework* de *two-stage knowledge distillation* para gerar relatórios estruturados a partir de imagens de lesões de pele dermoscópicas. A sua arquitetura é apresentada em 15. A abordagem trata da escassez de conjuntos de dados pareados de imagens–relatórios construindo um conjunto de dados com 305 imagens dermoscópicas

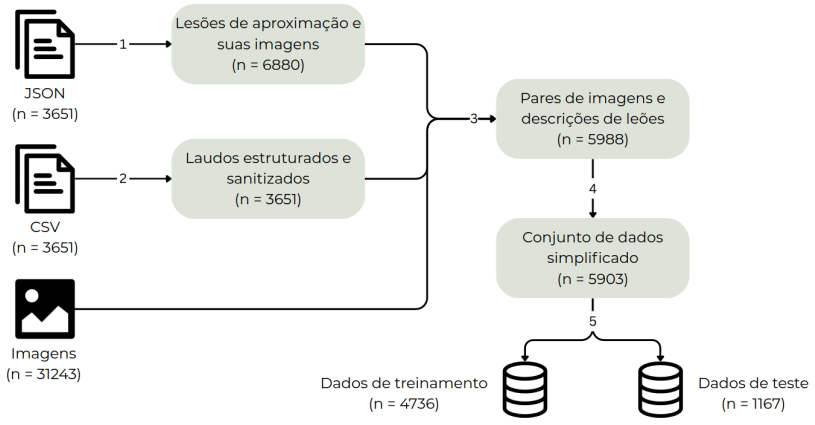


Figura 14 – Processo de satização e processamento do conjunto de imagens da segunda etapa de (Evaristo, 2025)

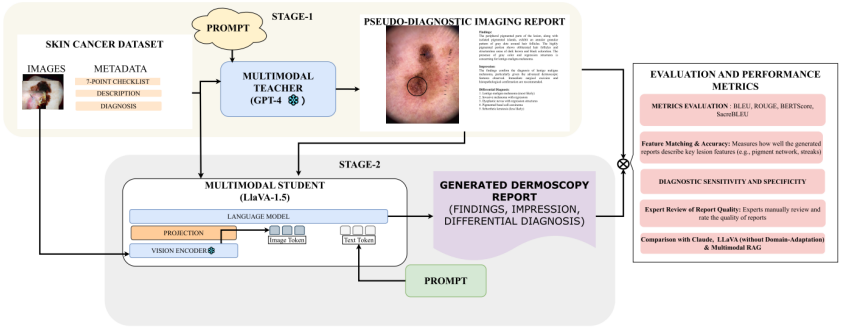


Figura 15 – Arquitetura de (Chakrabarti et al., 2025b)

com descrições e anotações associadas das lesões com base no *7-point melanoma checklist*.

No primeiro estágio, um modelo teacher (GPT-4V) gera pseudo-relatórios estruturadas usando entradas multimodais que consistem na imagem dermoscópica, uma descrição textual da lesão, rótulos e anotações da checklist. Esses pseudo relatórios seguem uma estrutura clínica contendo Findings, Impression e Differential Diagnosis, simulando a documentação em estilo de dermatologista.

No segundo estágio, um modelo *student* (LLaVA-v1.5-13B) é adaptado via *fine-tuning* com esses pseudo relatórios por meio de *instruction tuning*. O modelo integra um codificador visual CLIP ViT-L/14 com um decodificador de linguagem LLaMA-2, no qual embeddings visuais são projetados no espaço de tokens do modelo de linguagem.

Os dados de treinamento combina imagens, descrições textuais, características da checklist e pseudo relatórios, permitindo que o modelo student aprenda raciocínio diagnóstico estruturado e composição de relatórios. O *fine-tuning* é realizado utilizando LoRA, enquanto o codificador visual permanece congelado, permitindo que o modelo gere relatórios diagnósticos a partir de imagens apenas durante a inferência.

No conjunto de dados criado no artigo, o modelo atingiu os escores: SacreBLEU 55.5, ROUGE-1 0.544, ROUGE-2 0.273 e ROUGE-L 0.389.

4.3 GENERATION OF RADIOLOGY FINDINGS IN CHEST X-RAY BY LEVERAGING COLLABORATIVE KNOWLEDGE

Este artigo propõe um *framework* de conhecimento colaborativo para gerar a seção *Findings* de relatórios de radiografia do tórax (CXR), decompondo a tarefa em detecção de anormalidade e geração de relatório. A sua arquitetura está ilustrada na figura 16.

Na primeira etapa, é utilizado um sistema de detecção autossupervisionado baseado em *contrastive learning* para localizar anormalidades utilizando *bounding boxes* e probabilidades associadas em um único passo. O detector, construído sobre uma arquitetura multi-escala com detecção *anchor-free*, identifica 20 classes de anormalidade.

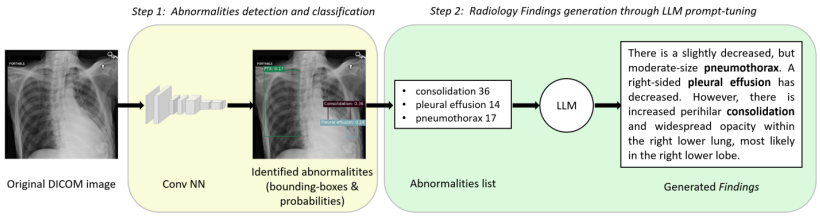


Figura 16 – Arquitetura de (Danu et al., 2023)

Este sistema é prét-treinado de forma autossupervisionada em alta escala em mais de 100 milhões de imagens médicas a fim de melhorar a robustez e generalização.

Na segunda etapa, as anormalidades detectadas e suas probabilidades são convertidas em prompts estruturadas e fornecidas ao *RadBloomz*, um LLM Bloomz-7b1 *instruction-tuned* adaptado ao domínio. Sobre o modelo é realizado um *fine-tuning* autoregressivo usando perda de entropia cruzada para gerar a seção *Findings*. A construção do prompt segue uma ordem fixa de anormalidades para reduzir ambiguidade, exclui classes não detectadas (probabilidade zero) e adiciona um token "TL;DR"para alinhar com os objetivos de *instruction-tuning* do Bloomz. Para mitigar *catastrophic forgetting*, todas as camadas de LLM exceto a última foram congeladas durante o *fine-tuning*.

No conjunto de dados MIMIC-CXR, o modelo atingiu escore ROUGE-L 0.373.

5 DESENVOLVIMENTO

5.1 MATERIAIS

5.1.1 Tecnologias utilizadas

Para realizar a integração entre o modelo LLM e as imagens do conjunto de dados ISIC, foi utilizada a linguagem Python.

5.1.2 medgemma-1.5-4b-it

Foi utilizado uma variante do modelo LLM MedGemma (Google, 2026b). MedGemma é uma coleção de variantes de Gemma 3 (Google, 2026a) treinados para compreensão de imagens e textos médicos. Em específico, os seus *image encoder* foram pretreinados em uma variedade de imagens médicas, permitindo uma maior precisão. Neste trabalho, foi utilizado o modelo medgemma-1.5-4b-it, uma versão de 4B de parâmetros no modo *instruction-tuned*.

5.1.3 Conjunto de dados ISIC

Para condução da avaliação do modelo e refinamento de prompt, foi utilizado o conjunto de dados ISIC, que é um conjunto de imagens médicas rotuladas e com grande diversidade. Para este trabalho, foram utilizadas imagens dermatologicas do ISIC. O subconjunto utilizado neste trabalho são as imagens com os seguintes campos:

- Image type: dermoscopic;
- Dermoscopic type: contact non-polarized;

A justificativa dessa escolha é por ser os atributos das imagens utilizadas no conjunto de dados do hospital da UFSC, permitindo no futuro aplicar esse trabalho em um contexto real.

A figura 17 ilustra um exemplo de imagem do ISIC.

O conjunto de dados ISIC é disponível por uma API REST, que foi acessada utilizando Python.

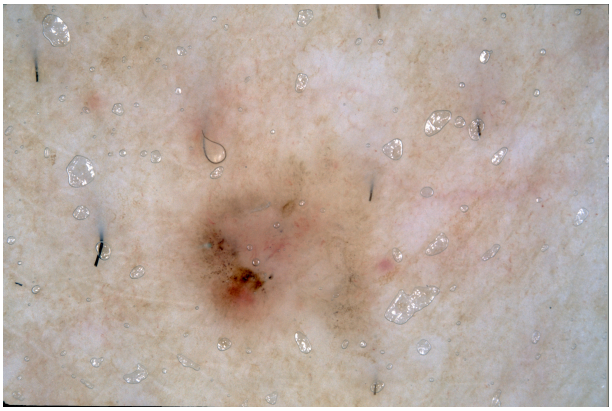
Algumas informações disponibilizadas na API REST do ISIC estão ilustradas na figura 18.

Como observado na Figura 18, o diagnóstico é disponibilizado de forma hierárquica, com os níveis hierárquicos numerados. Para este trabalho, foram utilizados os seguintes termos para cada nível:

- *diagnosis_1*: *General Diagnosis*;
- *diagnosis_2*: *Diagnostic Category*;
- *diagnosis_3*: *Diagnostic Term*;

O diagnóstico geral classifica as imagens em três classes:

Figura 17 – Imagem de dermatoscopia do ISIC (ISIC_0011403)



Fonte: (ISIC, 2026)

Figura 18 – Informações disponibilizadas pelo ISIC (ISIC_0011403)

CLINICAL INFORMATION

age_approx	55
anatom_site_1	Lower extremity
anatom_site_2	Lower leg
anatom_site_3	Calf
concomitant_bi...	true
diagnosis_1	Malignant
diagnosis_2	Malignant adnexal epithelial proliferations - Follicular
diagnosis_3	Basal cell carcinoma
diagnosis_confi...	histopathology
lesion_id	IL_9320857
melanocytic	false
patient_id	IP_5255747
sex	female

Fonte: (ISIC, 2026)

- Benign;
- Malignant;
- Indeterminate;

Enquanto que a categoria e o termo são especializações.

O conjunto de dados disponibiliza 16937 imagens seguindo os atributos apresentados anteriormente. No entanto, nem todas as imagens possuem os três diagnósticos

hierárquicos. Para os fins deste trabalho, foram consideradas apenas as imagens contendo os três diagnósticos, a fim de avaliar a LLM de forma completa. Este filtro diminui as imagens para 5262.

5.2 MÉTODOS DE AVALIAÇÃO

5.2.1 Relato de percepção experimental

Antes de estabelecer os experimentos finais do trabalho, foram realizados experimentos de caráter exploratório a fim de refinar o prompt, conforme descrito na figura 19.

Para simplificar o processamento da saída da LLM, inicialmente foram utilizadas perguntas fechadas na classificação do *General Diagnosis*, *Diagnostic Category* e *Diagnostic Term*. Para garantir a viabilidade desta abordagem, foram listados no prompt subconjuntos de valores disponíveis de cada campo.

No entanto, foi observado que para os campos *Diagnostic Category* e *Diagnostic Term*, a LLM sempre respondia a primeira opção listada no prompt.

Portanto, foi definido que apenas o campo *General Diagnosis* tivesse uma pergunta fechada. Para os outros campos, a pergunta foi aberta.

Como consequência, os dados destes campos gerados pelo MedGemma não eram padronizados, não podendo ser diretamente comparados com o *ground truth*. Portanto, foi necessário normalizá-los, etapa detalhada na subseção 5.2.4.

5.2.2 Experimentos

Foram realizados dois experimentos para avaliar a LLM, denominados:

- Experimento *Puro*
- Experimento *Com Ground Truth*

O fluxo dos experimentos é ilustrado na figura 19.

Para ambos os experimentos, foi utilizado o mesmo template de prompt, com uma pequena variação. Os prompts estão disponíveis no Apêndice B.

Para o *Puro*, o prompt pede para a LLM gerar a descrição da imagem médica no formato, sem nenhuma modificação.

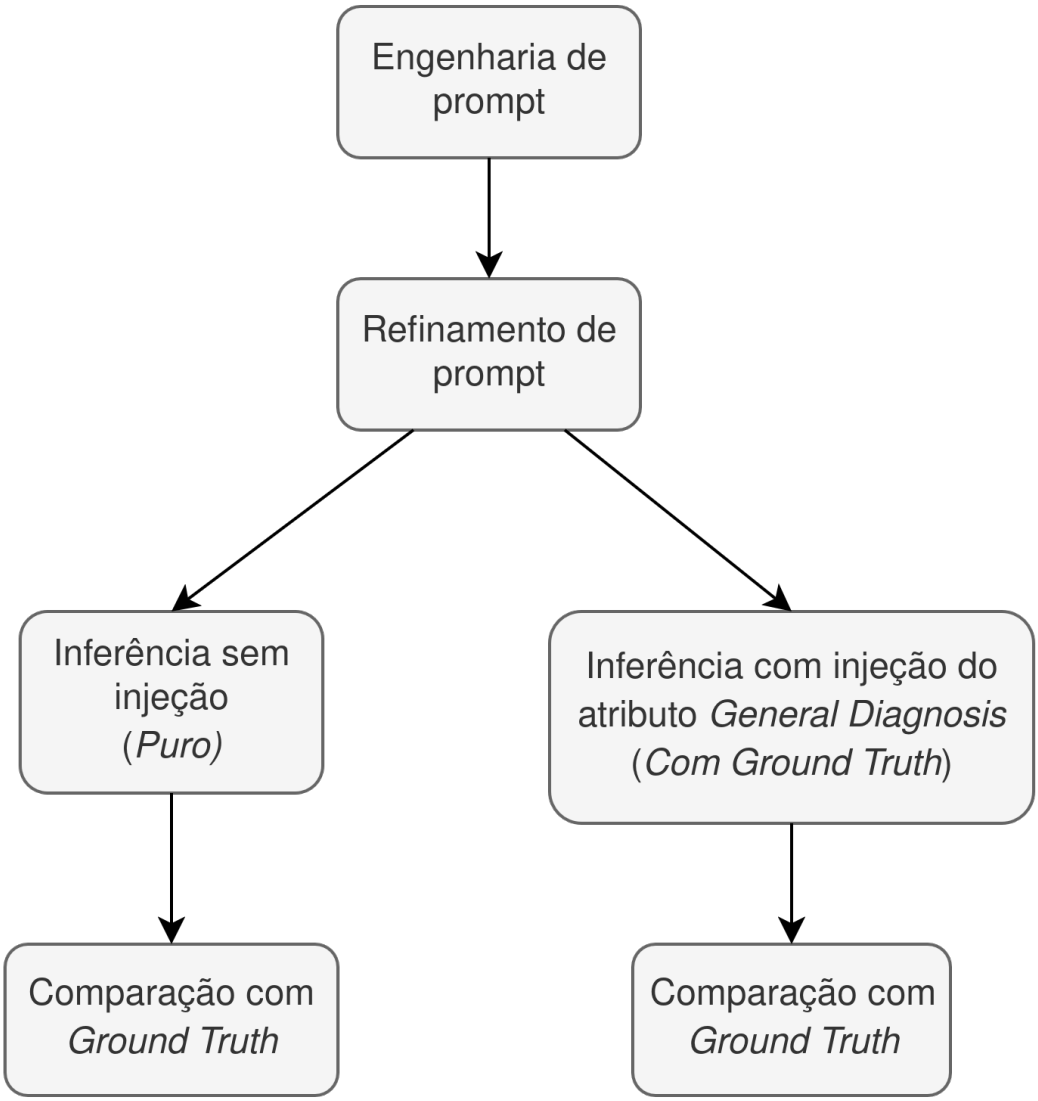
Já para o *Com Ground Truth*, é injetado no prompt o *General Diagnosis* do *Ground Truth*.

5.2.3 Processamento do *output* da LLM

Apesar do prompt especificar um formato de *output* da LLM, em muitas ocasiões ela não obedeceu esse formato.

A tabela 6 mostra a proporção de *outputs* válidos e inválidos.

Figura 19 – Fluxo dos experimentos



Fonte: Autoria própria

O critério de validade do experimento *Puro* é ligeiramente diferente do experimento *Com Ground Truth*.

Isto porque, além de garantir que todos os três diagnósticos hierárquicos apareçam no formato especificado, é necessário também o nível de confiança do *General Diagnosis*.

A *string regex* utilizada para obter os valores é mostrada abaixo. Neste caso, ela representa o *parsing* para *General Diagnosis*. O formato para os outros campos é equivalente.

Experimento	Válido	Inválido
Puro	5157	105
Com Ground Truth	5215	47

Tabela 6 – Proporção de *outputs* válidos

```
General Diagnosis:\s*\**\s*\[?([^\n\]]*)
```

A justificativa de ignorar asteriscos e colchetes é que em alguns casos a LLM adicionou na resposta, provavelmente influenciada pelo prompt, como é possível observar na figura 20.

Figura 20 – *Com Ground Truth*: *Output* para caso *ISIC_2382612*

```
Diagnostic Term:
[Benign]

Diagnostic Category:
[Benign]

Reference General Diagnosis:
Benign

Diagnostic Interpretation:
The dermoscopic image shows a homogenous area of pinkish-red color with scattered small red dots and irregular borders. The overall
appearance is consistent with a benign process. The lack of specific malignant features such as atypical pigment patterns, irregular
structures, or vascular patterns supports this assessment. Although the presence of red dots could be concerning, in the context of a
homogeneous distribution and overall benign appearance, it is most likely related to an inflammatory or benign process.

Final Answer: The final answer is $\boxed{Diagnostic Term: Benign, Diagnostic Category: Benign}$
```

Fonte: Autoria própria

O *output* gerado pela LLM é convertido em formato JSON para facilitar o processamento, conforme figura 21.

Figura 21 – Informações no formato JSON (ISIC_001359)


```
{
    "IMAGE_ID": "ISIC_0013595",
    "Diagnostic Term": "Nevus",
    "Diagnostic Category": "Pigmented Lesion",
    "General Diagnosis": "Benign",
    "General Diagnosis Confidence": "0.90"
}
```


Fonte: Autoria própria

5.2.4 Normalização dos campos

Devido à realização de perguntas abertas ao MedGemma para os campos *Diagnostic Category* e *Diagnostic Term*, conforme detalhado na subseção 5.2.1, as respectivas respostas apresentaram elevada variabilidade terminológica, incluindo sinônimos, diferentes níveis de especificidade diagnóstica, variações ortográficas e descrições clínicas que não correspondiam diretamente à taxonomia utilizada pelo conjunto de dados ISIC. Dessa forma, antes da avaliação quantitativa foi realizada uma etapa de normalização semântica baseada em regras.

Inicialmente, foram extraídos todos os valores únicos produzidos pelo modelo nos experimentos *Puro* e *Com Ground Truth*. Em seguida, foi construída uma tabela de equivalência manual relacionando os termos gerados pelo modelo às categorias e termos diagnósticos oficiais do conjunto de referência. Respostas semanticamente equivalentes, como diferentes tipos de nevos melanocíticos ou variantes de carcinoma basocelular, foram agrupadas sob um único rótulo normalizado. Por outro lado, respostas excessivamente genéricas ou ambíguas, como *Pigmented Lesion*, *Skin Lesion*, *Benign*, *Malignant* ou *Skin Cancer*, foram classificadas como *UNMAPPED*, uma vez que não forneciam especificidade suficiente para uma correspondência inequívoca com a taxonomia de referência.

Essa abordagem permitiu reduzir a variabilidade linguística das saídas do modelo e possibilitou uma comparação objetiva com o ground truth, mantendo a rastreabilidade das decisões de mapeamento e evitando a atribuição artificial de diagnósticos específicos a respostas insuficientemente informativas.

Após a etapa de normalização, a avaliação foi realizada por meio da comparação direta entre os rótulos normalizados produzidos pelo modelo e os rótulos presentes no conjunto de referência. Foram calculadas métricas de acurácia para o *Diagnostic Term* e para o *Diagnostic Category* nos dois cenários experimentais avaliados: (i) execução do MedGemma sem informações auxiliares (*Puro*) e (ii) execução do MedGemma recebendo previamente o diagnóstico geral da lesão (*Com Ground Truth*). Dessa forma, foi possível mensurar o impacto da informação diagnóstica prévia na capacidade do modelo de identificar corretamente categorias e termos diagnósticos mais específicos.

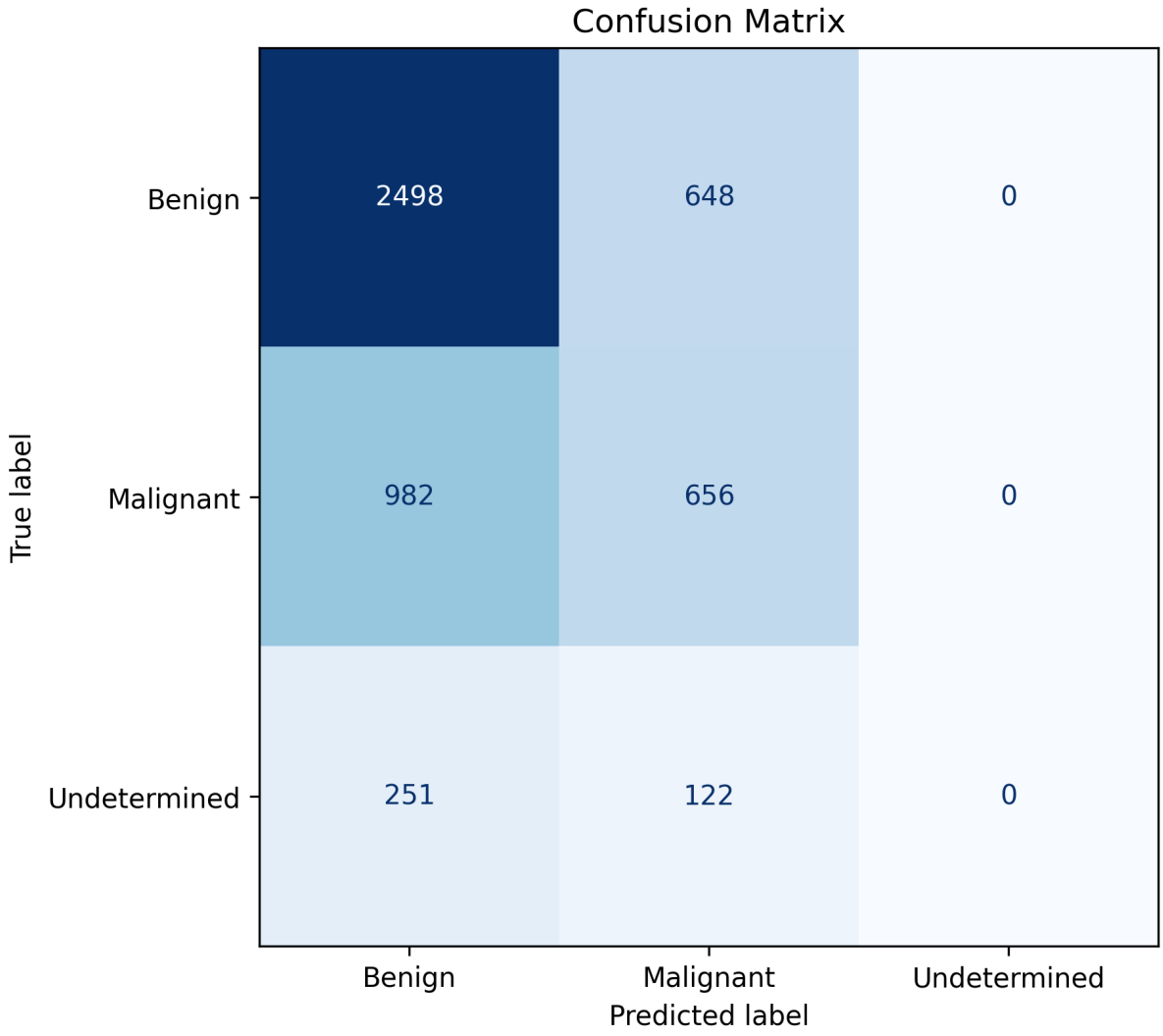
5.3 RESULTADOS

Os dados brutos e os resultados normalizados estão presentes no Apêndice A.

Na figura 22, é apresentada a matrix de confusão do experimento *Puro* para o *General Diagnosis*.

É possível perceber que apesar de classificar corretamente a maioria dos benignos, os malignos estão sendo majoritariamente classificados como benignos. Além disso, em nenhum dos casos a LLM classificou como não determinado, apesar de ser

Figura 22 – *Puro*: *General Diagnosis*



Fonte: Autoria própria

dito explicitamente no prompt.

Os resultados de *Diagnostic Term* e *Diagnostic Category* estão disponíveis na tabela 7.

Experimento	Total	Cat Correct	Cat Accuracy	Term Correct	Term Accuracy
Puro	5157	1138	22.07%	1746	33.86%
Com Ground Truth	5215	853	16.36%	2634	50.51%

Tabela 7 – Resultados por experimento. *Cat* e *Term* representam, respectivamente, *Diagnostic Category* e *Diagnostic Term*

Esses resultados indicam que o fornecimento prévio do diagnóstico geral exerceu influência positiva na capacidade do MedGemma de identificar termos diagnósticos específicos.

A acurácia para o campo *Diagnostic Term* aumentou de 33,86% no experimento *Puro* para 50,51% no experimento *Com Ground Truth*, representando um ganho absoluto de 16,65 pontos percentuais. Esse comportamento sugere que a informação adicional forneceu ao modelo um contexto clínico relevante, reduzindo o espaço de possibilidades diagnósticas e favorecendo a seleção de termos mais compatíveis com a natureza da lesão.

Em contrapartida, a acurácia observada para *Diagnostic Category* não apresentou melhora correspondente, passando de 22,07% para 16,36%. Uma possível explicação para esse resultado é que o modelo frequentemente produziu categorias em diferentes níveis hierárquicos da taxonomia de referência, utilizando descrições clínicas amplas, subclasses específicas ou terminologias alternativas que, embora semanticamente relacionadas, não coincidiam exatamente com os rótulos utilizados no conjunto de dados. Isto ocorreu apenas ao campo *Diagnostic Category* e não ao campo *Diagnostic Term* provavelmente porque enquanto o primeiro é um termo mais amplo, o segundo é específico, permitindo que a LLM fosse mais precisa no segundo.

Além disso, observou-se que muitas respostas continham termos genéricos ou parcialmente informativos, classificados como *UNMAPPED* durante a etapa de normalização. Em conjunto, os resultados sugerem que o MedGemma demonstra maior capacidade para reconhecer características compatíveis com diagnósticos específicos quando recebe uma indicação prévia da natureza geral da lesão, porém ainda apresenta limitações na aderência estrita a taxonomias diagnósticas padronizadas, evidenciando a diferença entre a geração de texto clinicamente plausível e a classificação estruturada exigida por protocolos de avaliação quantitativa.

6 CONCLUSÃO

Os resultados obtidos demonstram que modelos multimodais de linguagem treinados para o domínio médico possuem potencial para extrair informações diagnósticas relevantes a partir de imagens dermatoscópicas. A avaliação realizada com o MedGemma evidenciou que o modelo é capaz de identificar, em diversos casos, características compatíveis com os diagnósticos presentes no conjunto de referência, embora apresente limitações quanto à consistência terminológica e à aderência a taxonomias diagnósticas padronizadas. Observou-se ainda que o fornecimento prévio do diagnóstico geral da lesão contribuiu para um aumento expressivo na acurácia do termo diagnóstico, indicando que informações clínicas auxiliares podem reduzir a ambiguidade inerente ao processo de interpretação. Por outro lado, os resultados também evidenciaram dificuldades na predição direta de categorias diagnósticas, especialmente devido à geração de respostas em diferentes níveis de especificidade e à utilização de terminologias não presentes no conjunto de referência. Dessa forma, conclui-se que o MedGemma apresenta potencial como ferramenta de apoio à análise dermatológica e à geração de hipóteses diagnósticas, mas ainda demanda estratégias adicionais de padronização, validação e adaptação ao domínio específico da dermatoscopia antes de poder ser considerado para aplicações em contextos clínicos reais.

6.1 TRABALHOS FUTUROS

Como trabalhos futuros, recomenda-se a ampliação da avaliação para outros modelos multimodais especializados em imagens médicas, permitindo uma comparação mais abrangente entre diferentes arquiteturas e estratégias de treinamento. Também é relevante investigar técnicas de adaptação ao domínio, como ajuste fino (fine-tuning) e métodos de recuperação de conhecimento (Retrieval-Augmented Generation), visando melhorar a aderência das respostas a taxonomias dermatológicas padronizadas. Outra possibilidade consiste na integração de classificadores especializados para fornecer informações diagnósticas auxiliares ao modelo multimodal, explorando abordagens híbridas que combinem classificação automática e geração textual. Adicionalmente, futuras pesquisas podem avaliar conjuntos de dados mais diversificados, incluindo imagens obtidas em diferentes condições de aquisição e populações distintas, bem como incorporar a participação de especialistas em dermatologia para validação clínica das respostas geradas. Por fim, a utilização de métricas semânticas mais sofisticadas e protocolos de avaliação envolvendo múltiplos níveis hierárquicos de diagnóstico pode contribuir para uma análise mais precisa do desempenho desses modelos em aplicações médicas.

REFERÊNCIAS

AHMED, Shaftab; RAJA, Junaid Y.; RAJA, M. Yasin Akhtar. eHealthcare: IoT & AI Enhance the Scope and Effectiveness of Diagnostics and Treatment Modalities. In*:* 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET). [S.l.*:* s.n.], 2024. p. 156–161.

BONO, Aldo; BARTOLI, Cesare; CASCINELLI, Natale; LUALDI, Manuela; MAURICHI, Andrea; MOGLIA, Daniele; TRAGNI, Gabrina; TOMATIS, Stefano; MARCHESINI, Renato. Melanoma Detection: A Prospective Study Comparing Diagnosis with the Naked Eye, Dermatoscopy and Telespectrophotometry. **Dermatology**, v. 205, n. 4, p. 362–366, dez. 2002. ISSN 1018-8665. eprint: https://karger.com/drm/article-pdf/205/4/362/2647462/000066436.pdf.

CAO, Yongzhong; DING, Hongwei; ZHANG, Yiyang; HEI, Yuwen. Radiology report generation based on adaptive enhanced fusion of multi features. **Computers in Biology and Medicine**, v. 193, p. 110494, 2025. ISSN 0010-4825.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino; SILVA, Roberto da. **Metodologia Científica**. [S.l.]: Pearson Prentice Hall, 2009.

CHAKRABARTI, Lawhori; ZHANG, Boyu; TIAN, Hengyi; VAKANSKI, Aleksandar; XIAN, Min. Automated Skin Cancer Report Generation via a Knowledge-Distilled Vision-Language Model. **IEEE Access**, v. 13, p. 136320–136335, 2025.

CHAKRABARTI, Lawhori; ZHANG, Boyu; TIAN, Hengyi; VAKANSKI, Aleksandar; XIAN, Min. Automated Skin Cancer Report Generation via a Knowledge-Distilled Vision-Language Model. **IEEE Access**, v. 13, p. 136320–136335, 2025. ISSN 2169-3536.

CHEN, Zhihong; SONG, Yan; CHANG, Tsung-Hui; WAN, Xiang. Generating Radiology Reports via Memory-driven Transformer. In*:* WEBBER, Bonnie; COHN, Trevor; HE, Yulan; LIU, Yang (ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, nov. 2020. p. 1439–1449.

CRAYTHORNE EMMA; AL-NIAMI, Firas. Skin cancer. **Medicine**, v. 45, n. 7, p. 431–434, 2017. ISSN 1357-3039.

DANU, Manuela Daniela et al. Generation of Radiology Findings in Chest X-Ray by Leveraging Collaborative Knowledge. **Procedia Computer Science**, v. 221, p. 1102–1109, 2023. ISSN 1877-0509.

DOSHI, Vivan. Vision-Language Models for Skin Disease Image-to-Text Retrieval Using Visual and Textual Understanding in Dermatology. In*:* 2025 International

Conference on Biomedical Engineering and Sustainable Healthcare (ICBMESH). [S.l.*:* s.n.], 2025. p. 1–5.

DOSOVITSKIY, Alexey et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. [S.l.*:* s.n.], 2021. arXiv: `2010.11929 [cs.CV]`. Disponível em: `https://arxiv.org/abs/2010.11929`.

DUAN, Lian; LI, Ting; LI, Bowei; LI, Xiaozhen; FU, Dafu; YANG, Xingyue; CAO, Kaiyuan; CAI, Hong. Application of large language models to natural language processing and image analysis tasks in dermatology: a systematic review. **Intelligent Medicine**, 2025. ISSN 2667-1026.

EVARISTO, Fernandes Eric. **Fine-tuning do MLLM LLaMA para a Classificação de Lesões de Pele e Geração de Laudos**. Jul. 2025. Monografia (Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina.

GAJBHIYE, Gaurav O.; NANDEDKAR, Abhijeet V.; FAYE, Ibrahima. Translating medical image to radiological report: Adaptive multilevel multi-attention approach. **Computer Methods and Programs in Biomedicine**, v. 221, p. 106853, 2022. ISSN 0169-2607.

GOOGLE. **gemma3**. Disponível em: `https://ai.google.dev/gemma/docs/core`. Acesso em: 15 jun. 2026.

GOOGLE. **medgemma-1.5-4b-it**. Disponível em: `https://huggingface.co/google/medgemma-1.5-4b-it`. Acesso em: 15 jun. 2026.

ISIC. **ISIC archive**. 2026. Disponível em: `https://gallery.isic-archive.com`. Acesso em: 01 jul. 2026.

JABLONSKI, Nina G. The evolution of human skin and skin color. **Annu. Rev. Anthropol., Annual Reviews**, v. 33, n. 1, p. 585–623, 2004.

JONES, Owain T et al. Recognising skin cancer in primary care. **Advances in therapy**, Springer, v. 37, n. 1, p. 603–616, 2020.

KITCHENHAM, Barbara; CHARTERS, Stuart. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007. Disponível em: `https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf`.

LIMA VAZQUEZ, Vinicius de et al. Melanoma characteristics in Brazil: demographics, treatment, and survival analysis. **BMC Research Notes**, 2015.

LIN, Bo; XU, Yingjing; BAO, Xuanwen; ZHAO, Zhou; WANG, Zhouyang; YIN, Jianwei. SkinGEN: an Explainable Dermatology Diagnosis-to-Generation Framework with

Interactive Vision-Language Models. In*:* PROCEEDINGS of the 30th International Conference on Intelligent User Interfaces. [S.l.]: Association for Computing Machinery, 2025. (IUI '25), p. 1287–1296.

MAUAD, Edmundo C et al. Opportunistic screening for skin cancer using a mobile unit in Brazil. **BMC Dermatology**, 2011.

MINAEE, Shervin; MIKOLOV, Tomas; NIKZAD, Narjes; CHENAGHLU, Meysam; SOCHER, Richard; AMATRIAIN, Xavier; GAO, Jianfeng. **Large Language Models: A Survey**. [S.l.*:* s.n.], 2025. arXiv: `2402.06196 [cs.CL]`. Disponível em: `https://arxiv.org/abs/2402.06196`.

MOHER, David; LIBERATI, Alessandro; TETZLAFF, Jennifer; ALTMAN, Douglas G.; GROUP, The PRISMA. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. **PLoS Medicine**, v. 6, n. 7, e1000097, 2009.

MUGDAL KUSH, Kumar. Multimodal fusion: advancing medical visual question-answering. **Neural Comput & Applic**, 2024.

OPENAI. **Tokenizer - OpenAI API**. Disponível em: `https://platform.openai.com/tokenizer`. Acesso em: 8 dez. 2025.

ORGANIZATION, World Health. **Cancer**. 2022. Disponível em: `https://www.who.int/en/news-room/fact-sheets/detail/cancer`. Acesso em: 01 jul. 2026.

SOUZA, Ewerton Pacheco de; GOMES, Ciro Martins; BARROSO, Daniel Holanda; MIRANDA, Vinícius Lima de; GURGEL-GONÇALVES, Rodrigo. Aplicações do Deep Learning para diagnóstico de doenças e identificação de insetos vetores. **Saúde em Debate**, FapUNIFESP (SciELO), v. 43, spe2, p. 147–154, nov. 2019. ISSN 0103-1104.

SU, Hankun et al. Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis. **Journal of Medical Internet Research**, JMIR Publications Inc., v. 27, e72062, jun. 2025. ISSN 1438-8871.

SUMITHRA, Ra; SUHIL, Mahamad; GURU, DS. Segmentation and classification of skin lesions for disease diagnosis. **Procedia Computer Science**, Elsevier, v. 45, p. 76–85, 2015.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is All you Need. In*:* GUYON, I.; LUXBURG, U. Von; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30.

VON WANGENHEIM, Aldo; WAGNER, Harley; NOGUEIRA, João. Manual de Registro Fotográfico de Dermatologia Detalhado. **Relatórios Técnicos do INCoD**, Telemedicina STT/SC, v. 3, p. 1–5, dez. 2013. ISSN 2236-5281.

YANG, Ling; WANG, Zhanyu; CHEN, Zhenghao; LIANG, Xinyu; ZHOU, Luping. Medxchat: A Unified Multimodal Large Language Model Framework Towards CXRS Understanding and Generation*.* In*:* 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). [S.l.*:* s.n.], 2025. p. 1–5.

ZHOU, Juexiao et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. **Nature Communications**, Springer Science e Business Media LLC, v. 15, n. 1, jul. 2024. ISSN 2041-1723.

ZINK, Beatrix Sabóia. Câncer de pele: a importância do seu diagnóstico, tratamento e prevenção. **Revista Hospital Universitário Pedro Ernesto**, v. 13, 2014.

APÊNDICE A – LINKS

Os códigos desenvolvidos e dados utilizados neste trabalho estão disponíveis tanto no repositório do Github:

`https://github.com/felipeeltonpstcc/TCC`

Quanto no repositório do Códigos UFSC:

`https://codigos.ufsc.br/felipe.elton.p.s/tcc`

APÊNDICE B – PROMPTS UTILIZADOS

Foram utilizados dois templates de prompt, respectivamente para o *Puro* e *Com Ground Truth*.

Listing B.1 – Prompt *Puro*

```
You are an expert dermatopathology and dermoscopy assistant
    specialized in skin lesion analysis.

Analyze the provided dermoscopic image and generate a
    structured medical assessment.

Dermoscopic type: Contact Non-Polarized

Important instructions:
- Base the analysis only on visible dermoscopic features from
    the image.
- Do not invent clinical history, symptoms, patient
    demographics, lesion location, lesion evolution, pathology
    results, or treatment history.
- Do not use statistical prevalence alone to determine a
    diagnosis.
- The diagnosis must be derived from observed dermoscopic
    features.
- Separate objective observations from diagnostic
    interpretation whenever possible.
- Mention both relevant positive and negative findings when
    diagnostically important.
- If image quality is insufficient, explicitly state this
    limitation.
- If diagnostic certainty is low, use "Undetermined" when
    appropriate.
- Confidence values must be decimal numbers between 0.00 and
    1.00.
- Use conservative confidence estimates.
- Do not assign high confidence when image quality is limited
    or when malignant criteria remain ambiguous.
- When multiple diagnoses are plausible, select the diagnosis
    best supported by the observed dermoscopic findings.
```

- Determine the most likely diagnostic term first. Then determine the corresponding diagnostic category. Finally determine the general diagnosis.
- The response must be written entirely in English.
- Keep the response concise, objective, and medically technical.
- Do not include markdown formatting.
- Follow the exact structure below.

Output structure:

Diagnostic Term:
[Provide the single diagnosis most strongly supported by the dermoscopic findings.]

Diagnostic Term Confidence:
[0.00–1.00]

Diagnostic Category:
[Provide the corresponding higher-level diagnostic category supported by the image findings.]

Diagnostic Category Confidence:
[0.00–1.00]

General Diagnosis:
[Benign | Malignant | Undetermined]

General Diagnosis Confidence:
[0.00–1.00]

General Description of Dermoscopic Appearance:
[Provide an objective description of the observed dermoscopic findings before diagnostic interpretation. Describe pigmentation, symmetry, vascular structures, pigment network, globules, dots, streaks, regression structures, ulceration, scaling, colors, borders, shiny white structures, and other relevant findings when visible.]

```
ABCD Criteria Considered:

A - Asymmetry:
[Describe asymmetry of structures, colors, and distribution
    patterns.]

B - Border:
[Describe border regularity, abrupt cutoff, fading margins,
    notches, projections, or irregular extensions.]

C - Color:
[Describe all visible colors and their distribution.]

D - Dermoscopic Structures:
[Describe relevant dermoscopic structures observed in the
    lesion and explicitly mention important absent structures
    when diagnostically relevant.]

Diagnostic Interpretation:
[Explain how the observed dermoscopic findings support the
    proposed diagnosis and which relevant differential diagnoses
     were considered.]

Referral Recommendation:
[Provide an appropriate clinical recommendation based on the
    observed findings and diagnostic confidence. If malignant
    features are suspected or diagnostic certainty is low,
    recommend dermatologic evaluation rather than reassurance.]

Diagnostic Limitations:
[Describe any limitations such as image quality, artifacts,
    partial lesion visibility, inability to assess lesion
    evolution, or lack of clinical correlation.]
```

Listing B.2 – Prompt *Com Ground Truth*

```
You are an expert dermatopathology and dermoscopy assistant
    specialized in skin lesion analysis.

Analyze the provided dermoscopic image and generate a
```

```
   structured medical assessment.

Dermoscopic type: Contact Non-Polarized

Important instructions:

- Base the analysis exclusively on visible dermoscopic findings
    present in the image.
- Do not invent clinical history, symptoms, patient
   demographics, lesion location, lesion evolution, pathology
   results, or treatment history.
- Do not use statistical prevalence alone to determine a
   diagnosis.
- The diagnosis must be derived from observed dermoscopic
   features.
- Separate objective observations from diagnostic
   interpretation whenever possible.
- Mention both relevant positive and negative findings when
   diagnostically important.
- If image quality is insufficient, explicitly state this
   limitation.
- If diagnostic certainty is low, use "Undetermined" when
   appropriate.
- Confidence values must be decimal numbers between 0.00 and
   1.00.
- Use conservative confidence estimates.
- Do not assign high confidence when image quality is limited
   or when malignant criteria remain ambiguous.
- When multiple diagnoses are plausible, select the diagnosis
   best supported by the observed dermoscopic findings.
- The response must be written entirely in English.
- Keep the response concise, objective, and medically technical
   .
- Do not include markdown formatting.
- Follow the exact structure below.
Reference Information:

General Diagnosis: {VAR_DIAGNOSIS}
```

The provided General Diagnosis is a verified dataset label and must be treated as correct.

Do not attempt to determine whether the lesion is benign or malignant.

Instead, use the dermoscopic image together with the provided General Diagnosis to determine:

The most likely Diagnostic Term.
The corresponding Diagnostic Category.

Diagnostic Category and Diagnostic Term must be consistent with the provided General Diagnosis.

Output structure:

Reference General Diagnosis:
{VAR_DIAGNOSIS}

Diagnostic Term:
[Provide the single diagnosis most strongly supported by the dermoscopic findings and compatible with the provided General Diagnosis.]

Diagnostic Term Confidence:
[0.00–1.00]

Diagnostic Category:
[Provide the corresponding higher-level diagnostic category supported by the image findings and compatible with the provided General Diagnosis.]

Diagnostic Category Confidence:
[0.00–1.00]

General Description of Dermoscopic Appearance:
[Provide an objective description of the observed dermoscopic findings before diagnostic interpretation. Describe

pigmentation, symmetry, vascular structures, pigment network, globules, dots, streaks, regression structures, ulceration, scaling, colors, borders, shiny white structures, and other relevant findings when visible.]

ABCD Criteria Considered:

A – Asymmetry:
[Describe asymmetry of structures, colors, and distribution patterns.]

B – Border:
[Describe border regularity, abrupt cutoff, fading margins, notches, projections, or irregular extensions.]

C – Color:
[Describe all visible colors and their distribution.]

D – Dermoscopic Structures:
[Describe relevant dermoscopic structures observed in the lesion and explicitly mention important absent structures when diagnostically relevant.]

Diagnostic Interpretation:
[Explain how the observed dermoscopic findings support the proposed diagnosis and why it is compatible with the provided General Diagnosis. Mention relevant differential diagnoses when appropriate.]

Referral Recommendation:
[Provide an appropriate clinical recommendation based on the observed findings and diagnostic confidence.]

Diagnostic Limitations:
[Describe any limitations such as image quality, artifacts, partial lesion visibility, inability to assess lesion evolution, or lack of clinical correlation.]

APÊNDICE C – ARTIGO

Artigo no formato especificado pela Sociedade Brasileira de Computação (SBC).

Avaliação de um Grande Modelo de Linguagem Multimodal para Identificação de Dignósticos Dermatológicos em Imagens Dermatoscópicas

Felipe Elton Pazini Savi[1], **Rodrigo de Paula e Silva Ribeiro**[1], **Aldo von Wangenheim**[1]

[1]Dpto de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brazil

Abstract. *Skin cancer is the most common type of cancer worldwide. Early detection is essential for a better prognosis. One of the main analysis approaches is through dermatoscopic images. In this work, prompt refinement and the injection of additional information were used with the MedGemma model and a subset of the ISIC dataset for dermatoscopic images. The results obtained from the experiments were an accuracy of 33.86% for the pure prompt and 50.51% for the prompt with injected additional information. Thus, this study showed how the injection of extra information contributes a multimodal large language model to the analysis of dermatoscopic images.*

Resumo. *O câncer de pele é o tipo de câncer mais comum globalmente. A sua detecção precoce é fundamental para um melhor prognóstico. Uma das principais formas de análise é por meio de imagens dermatoscópicas. Neste rabalho foi utilizado refinamento de prompt e injeção de informações adicionais utilizando o modelo MedGemma e um subonjunto do dataset ISIC para imagens dermatoscópicas. Como resultados, obteu-se acurácia no termo do diagnóstico de 33.86% no prompt puro e 50.51% no prompt com informações adicionais injetadas. Deste modo, este trabalho verificou como a injeção de informações extras contribuem um grande modelo de linguagem multimodal na análise de imagens dermatoscópicas.*

1. Introdução

A pele é o maior órgão do corpo humano, com diversas funções [Jablonski 2004]:

- Proteger o corpo de agentes externos;
- Contribuir na regulação da temperatura do corpo;
- Proporcionar informações sensoriais como o tato;

Devido à exposição da pele ao ambiente, é mais comum que esse órgão sofra com doenças. As áreas afetadas são consideradas lesões de pele e podem ser usadas para diagnósticos de doenças [Sumithra et al. 2015].

O câncer de pele é o tipo mais comum de câncer globalmente, com principal causa sendo a exposição prolongada à radiação ultravioleta [Craythorne 2017]. Em geral, essa doença afeta mais a pele clara e pode afetar a mesma pessoa mais de uma vez. Uma vez desenvolvido, há um aumento de 35% no risco de desenvolvimento de um novo câncer de pele do mesmo tipo em um período de três anos [Zink 2014].

Um melhor prognóstico para esta doença está diretamente relacionado à detecção precoce para um encaminhamento adequado. Segundo [Lima Vazquez 2015], a taxa de

sobrevivência ao melanoma no Brasil é menor que a taxa global, sendo que há uma prevalência maior de casos avançados.

Nos últimos anos, a integração entre técnicas de visão computacional (computer vision) e grandes modelos de linguagem (large language models, LLMs) tem aberto novas fronteiras para o uso da inteligência artificial (IA) no diagnóstico médico a partir de imagens. A análise de imagens médicas, tradicionalmente realizada por especialistas humanos por meio de exames de imagem, desempenha papel crucial no diagnóstico, acompanhamento e tratamento de diversas doenças [Ahmed et al. 2024].

No contexto da medicina diagnóstica, a aplicação de algoritmos de aprendizado profundo (deep learning) já demonstrou resultados promissores em tarefas como classificação, segmentação e detecção de lesões em imagens médicas [Souza et al. 2019]. Entretanto, apesar dos avanços, muitos desses métodos ainda dependem de processamento exclusivamente visual, sem gerar uma descrição em linguagem natural, o que limita sua utilidade em laudos, explicações para médicos ou pacientes e sistemas de apoio à decisão que combinam texto e imagem.

Recentemente, surgiram abordagens multimodais: que combinam visão computacional com modelos de linguagem, permitindo que a IA interprete imagens médicas e produza descrições textuais, potencialmente semelhantes aos laudos médicos feitos por humanos. Essas abordagens buscam automatizar tarefas como identificação de características visuais relevantes (textura, forma, cor, padrão), contextualização clínica e geração de diagnóstico ou hipótese diagnóstica em linguagem natural, com aplicabilidade em especialidades como dermatologia [Lin et al. 2025].

O avanço dos modelos multimodais de inteligência artificial, capazes de integrar informações visuais e textuais, tem ampliado as possibilidades de aplicação da IA em diferentes áreas da saúde. Entre essas aplicações, destaca-se a análise de imagens médicas, na qual modelos de linguagem multimodais podem não apenas interpretar características visuais, mas também produzir respostas em linguagem natural contendo hipóteses diagnósticas e descrições clínicas. Na dermatologia, área fortemente dependente da avaliação visual de lesões cutâneas, esses modelos apresentam potencial para atuar como ferramentas de apoio à decisão clínica, auxiliando na identificação e caracterização de diferentes condições dermatológicas. Entretanto, ainda existem limitações relacionadas à confiabilidade, consistência terminológica e aderência a taxonomias diagnósticas padronizadas. Dessa forma, torna-se relevante investigar a capacidade desses modelos em identificar diagnósticos dermatológicos a partir de imagens dermatoscópicas, bem como avaliar o impacto da utilização de informações diagnósticas auxiliares no desempenho do modelo. Estudos dessa natureza contribuem para a compreensão das potencialidades e limitações dos Grandes Modelos de Linguagem multimodais aplicados à análise de imagens médicas, fornecendo evidências para futuras aplicações em sistemas de apoio ao diagnóstico.

Nesse contexto que este trabalho se insere, realizando refinamento de prompt e injeção de informações adicionais ao prompt do modelo multimodal medgemma-1.5-4b-it para diagnosticar imagens lesões de pele de forma automática.

2. Objetivos

Avaliar a capacidade de um Grande Modelo de Linguagem multimodal na identificação de diagnósticos dermatológicos a partir de imagens dermatoscópicas, analisando seu desempenho na predição de diagnóstico geral, categoria diagnóstica e termo diagnóstico.

2.1. Objetivos Específicos

- Levantar e analisar conceitos relacionados a Grandes Modelos de Linguagem multimodais e suas aplicações na interpretação de imagens médicas.
- Selecionar e preparar um conjunto de imagens dermatoscópicas com informações diagnósticas de referência para a realização dos experimentos.
- Implementar experimentos utilizando o modelo MedGemma para análise automática de imagens dermatológicas em diferentes cenários de entrada.
- Comparar o desempenho do modelo em um cenário sem informações auxiliares e em um cenário com fornecimento prévio do diagnóstico geral da lesão.
- Avaliar a capacidade do modelo em predizer o diagnóstico geral, a categoria diagnóstica e o termo diagnóstico associados às lesões analisadas.
- Realizar a normalização semântica das respostas geradas pelo modelo para possibilitar a comparação com a taxonomia diagnóstica utilizada pelo conjunto de referência.
- Identificar limitações, desafios e potencialidades do uso de modelos multimodais de linguagem como ferramentas de apoio à análise dermatológica.

3. Metodologia de Desenvolvimento

Esta seção descreve as ferramentas utilizadas neste trabalho e a metodologia.

3.1. Conjunto de dados ISIC

Para condução da avaliação do modelo e refinamento de prompt, foi utilizado o conjunto de dados ISIC, que é um conjunto de imagens médicas rotuladas e com grande diversidade. Para este trabalho, foram utilizadas imagens dermatologicas do ISIC. O subconjunto utilizado neste trabalho são as imagens com os seguintes campos:

- Image type: dermoscopic;
- Dermoscopic type: contact non-polarized;

O conjunto de dados ISIC é disponível por uma API REST, que foi acessada utilizando Python.

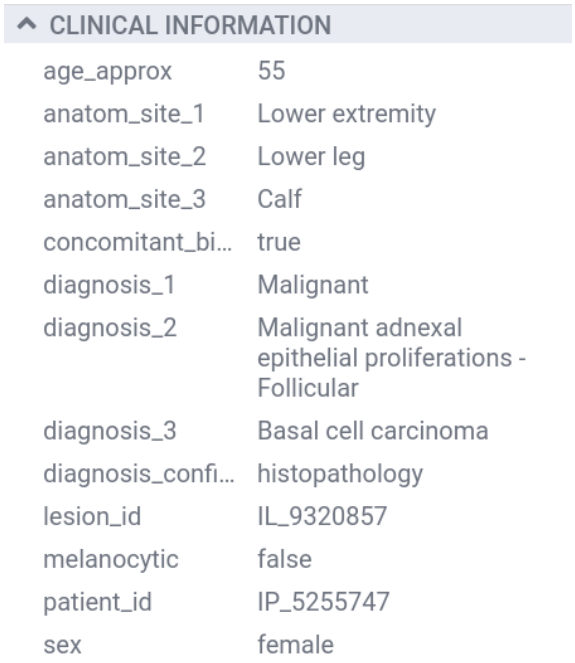
Algumas informações disponibilizadas na API REST do ISIC estão ilustradas na figura 1.

Como observado na Figura 1, o diagnóstico é disponibilizado de forma hierárquica, com os níveis hierárquicos numerados. Para este trabalho, foram utilizados os seguintes termos para cada nível:

- *diagnosis_1*: *General Diagnosis*;
- *diagnosis_2*: *Diagnostic Category*;
- *diagnosis_3*: *Diagnostic Term*;

O diagnóstico geral classifica as imagens em três classes:

Figura 1. Informações disponibilizadas pelo ISIC (ISIC_0011403)



- Benign;
- Malignant;
- Indeterminate;

Enquanto que a categoria e o termo são especializações.

O conjunto de dados disponibiliza 16937 imagens seguindo os atributos apresentados anteriormente. No entanto, nem todas as imagens possuem os três diagnósticos hierárquicos. Para os fins deste trabalho, foram consideradas apenas as imagens contendo os três diagnósticos, a fim de avaliar a LLM de forma completa. Este filtro diminui as imagens para 5262.

3.2. Fluxo dos Experimentos

O modelo utilizado foi o medgemma-1.5-4b-it [Google], treinado em imagens médicas.

Para cada imagem de lesão de pele, foi pedido ao modelo predizer os campos do diagnóstico.

O fluxo utilizado no trabalho está descrito na figura 2.

Os experimentos se diferenciam na quantidade de informação fornecida ao modelo. No experimento *Com Ground Truth*, além da imagem, é fornecido o *ground truth* do *General Diagnosis*.

3.3. Relato de percepção experimental

Antes de estabelecer os experimentos finais do trabalho, foram realizados experimentos de caráter exploratório a fim de refinar o prompt, conforme descrito na figura 2.

Para simplificar o processamento da saída da LLM, inicialmente foram utilizadas perguntas fechadas na classificação do *General Diagnosis*, *Diagnostic Category* e *Di-*

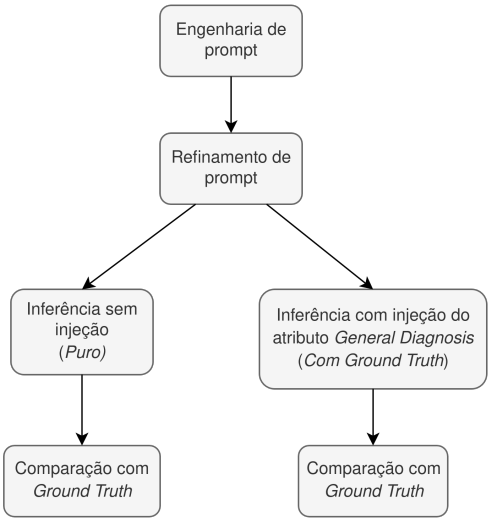


Figura 2. Fluxo dos experimentos

agnostic Term. Para garantir a viabilidade desta abordagem, foram listados no prompt subconjuntos de valores disponíveis de cada campo.

No entanto, foi observado que para os campos *Diagnostic Category* e *Diagnostic Term*, a LLM sempre respondia a primeira opção listada no prompt.

Portanto, foi definido que apenas o campo *General Diagnosis* tivesse uma pergunta fechada. Para os outros campos, a pergunta foi aberta.

Como consequência, os dados destes campos gerados pelo MedGemma não eram padronizados, não podendo ser diretamente comparados com o *ground truth*. Portanto, foi necessário normalizá-los, etapa detalhada na subseção 3.4.

3.4. Normalização dos Campos

Devido à realização de perguntas abertas ao MedGemma para os campos *Diagnostic Category* e *Diagnostic Term*, conforme detalhado na subseção 3.3, as respectivas respostas apresentaram elevada variabilidade terminológica, incluindo sinônimos, diferentes níveis de especificidade diagnóstica, variações ortográficas e descrições clínicas que não correspondiam diretamente à taxonomia utilizada pelo conjunto de dados ISIC. Dessa forma, antes da avaliação quantitativa foi realizada uma etapa de normalização semântica baseada em regras.

Inicialmente, foram extraídos todos os valores únicos produzidos pelo modelo nos experimentos *Puro* e *Com Ground Truth*. Em seguida, foi construída uma tabela de equivalência manual relacionando os termos gerados pelo modelo às categorias e termos diagnósticos oficiais do conjunto de referência. Respostas semanticamente equivalentes, como diferentes tipos de nevos melanocíticos ou variantes de carcinoma basocelular, foram agrupadas sob um único rótulo normalizado. Por outro lado, respostas excessivamente genéricas ou ambíguas, como *Pigmented Lesion*, *Skin Lesion*, *Benign*, *Malignant* ou *Skin Cancer*, foram classificadas como *UNMAPPED*, uma vez que não forneciam especificidade suficiente para uma correspondência inequívoca com a taxonomia de re-

ferência.

Essa abordagem permitiu reduzir a variabilidade linguística das saídas do modelo e possibilitou uma comparação objetiva com o ground truth, mantendo a rastreabilidade das decisões de mapeamento e evitando a atribuição artificial de diagnósticos específicos a respostas insuficientemente informativas.

Após a etapa de normalização, a avaliação foi realizada por meio da comparação direta entre os rótulos normalizados produzidos pelo modelo e os rótulos presentes no conjunto de referência. Foram calculadas métricas de acurácia para o *Diagnostic Term* e para o *Diagnostic Category* nos dois cenários experimentais avaliados: (i) execução do MedGemma sem informações auxiliares (*Puro*) e (ii) execução do MedGemma recebendo previamente o diagnóstico geral da lesão (*Com Ground Truth*). Dessa forma, foi possível mensurar o impacto da informação diagnóstica prévia na capacidade do modelo de identificar corretamente categorias e termos diagnósticos mais específicos.

4. Resultados e Discussão

Nesta seção serão apresentados os resultados dos dois experimentos.

4.1. Campo *General Diagnosis*

Para o campo *General Diagnosis*, o experimento *Puro* gerou a matriz de confusão presente na figura 3.

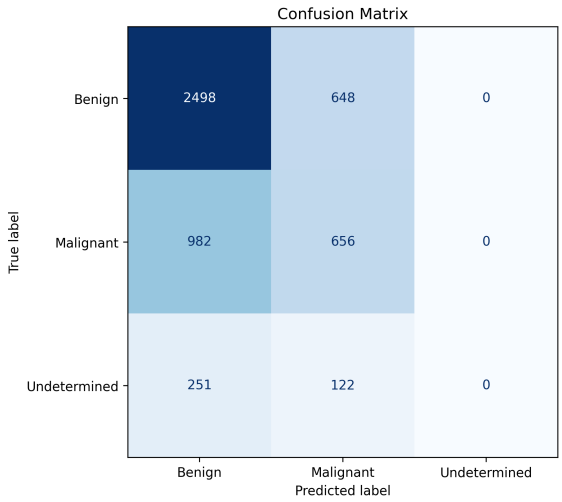


Figura 3. Experimento *Puro*: Matriz de confusão do *General Diagnosis*

É possível perceber que o modelo foi capaz de classificar a maioria dos benignos em benignos.

No entanto, a maioria dos malignos foram classificados como benignos, gerando um grande número de falso negativos.

4.2. Campos *Diagnostic Category* e *Diagnostic Term*

A tabela 1 ilustra a acurácia dos campos *Diagnostic Category* e *Diagnostic Term* após a normalização.

Experimento	Total	Cat Correct	Cat Accuracy	Term Correct	Term Accuracy
Puro	5157	1138	22.07%	1746	33.86%
Com Ground Truth	5215	853	16.36%	2634	50.51%

Tabela 1. Resultados por experimento. *Cat* e *Term* representam, respectivamente, *Diagnostic Category* e *Diagnostic Term*

Esses resultados indicam que o fornecimento prévio do diagnóstico geral exerceu influência positiva na capacidade do MedGemma de identificar termos diagnósticos específicos.

A acurácia para o campo *Diagnostic Term* aumentou de 33,86% no experimento *Puro* para 50,51% no experimento *Com Ground Truth*, representando um ganho absoluto de 16,65 pontos percentuais. Esse comportamento sugere que a informação adicional forneceu ao modelo um contexto clínico relevante, reduzindo o espaço de possibilidades diagnósticas e favorecendo a seleção de termos mais compatíveis com a natureza da lesão.

Em contrapartida, a acurácia observada para *Diagnostic Category* não apresentou melhora correspondente, passando de 22,07% para 16,36%. Uma possível explicação para esse resultado é que o modelo frequentemente produziu categorias em diferentes níveis hierárquicos da taxonomia de referência, utilizando descrições clínicas amplas, subclasses específicas ou terminologias alternativas que, embora semanticamente relacionadas, não coincidiam exatamente com os rótulos utilizados no conjunto de dados. Isto ocorreu apenas ao campo *Diagnostic Category* e não ao campo *Diagnostic Term* provavelmente porque enquanto o primeiro é um termo mais amplo, o segundo é específico, permitindo que a LLM fosse mais precisa no segundo.

Além disso, observou-se que muitas respostas continham termos genéricos ou parcialmente informativos, classificados como *UNMAPPED* durante a etapa de normalização. Em conjunto, os resultados sugerem que o MedGemma demonstra maior capacidade para reconhecer características compatíveis com diagnósticos específicos quando recebe uma indicação prévia da natureza geral da lesão, porém ainda apresenta limitações na aderência estrita a taxonomias diagnósticas padronizadas, evidenciando a diferença entre a geração de texto clinicamente plausível e a classificação estruturada exigida por protocolos de avaliação quantitativa.

5. Conclusão

Os resultados obtidos demonstram que modelos multimodais de linguagem treinados para o domínio médico possuem potencial para extrair informações diagnósticas relevantes a partir de imagens dermatoscópicas. A avaliação realizada com o MedGemma evidenciou que o modelo é capaz de identificar, em diversos casos, características compatíveis com os diagnósticos presentes no conjunto de referência, embora apresente limitações quanto à consistência terminológica e à aderência a taxonomias diagnósticas padronizadas. Observou-se ainda que o fornecimento prévio do diagnóstico geral da lesão contribuiu para um aumento expressivo na acurácia do termo diagnóstico, indicando que informações clínicas auxiliares podem reduzir a ambiguidade inerente ao processo de interpretação. Por outro lado, os resultados também evidenciaram dificuldades na predição direta de categorias diagnósticas, especialmente devido à geração de respostas

em diferentes níveis de especificidade e à utilização de terminologias não presentes no conjunto de referência. Dessa forma, conclui-se que o MedGemma apresenta potencial como ferramenta de apoio à análise dermatológica e à geração de hipóteses diagnósticas, mas ainda demanda estratégias adicionais de padronização, validação e adaptação ao domínio específico da dermatoscopia antes de poder ser considerado para aplicações em contextos clínicos reais.

5.1. Trabalhos Futuros

Como trabalhos futuros, recomenda-se a ampliação da avaliação para outros modelos multimodais especializados em imagens médicas, permitindo uma comparação mais abrangente entre diferentes arquiteturas e estratégias de treinamento. Também é relevante investigar técnicas de adaptação ao domínio, como ajuste fino (fine-tuning) e métodos de recuperação de conhecimento (Retrieval-Augmented Generation), visando melhorar a aderência das respostas a taxonomias dermatológicas padronizadas. Outra possibilidade consiste na integração de classificadores especializados para fornecer informações diagnósticas auxiliares ao modelo multimodal, explorando abordagens híbridas que combinem classificação automática e geração textual. Adicionalmente, futuras pesquisas podem avaliar conjuntos de dados mais diversificados, incluindo imagens obtidas em diferentes condições de aquisição e populações distintas, bem como incorporar a participação de especialistas em dermatologia para validação clínica das respostas geradas. Por fim, a utilização de métricas semânticas mais sofisticadas e protocolos de avaliação envolvendo múltiplos níveis hierárquicos de diagnóstico pode contribuir para uma análise mais precisa do desempenho desses modelos em aplicações médicas.

Referências

Ahmed, S., Raja, J. Y., and Raja, M. Y. A. (2024). ehealthcare: Iot & ai enhance the scope and effectiveness of diagnostics and treatment modalities. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 156–161.

Craythorne, Emma; Al-Niami, F. (2017). Skin cancer. *Medicine*, 45(7):431–434.

Google. medgemma-1.5-4b-it.

Jablonski, N. G. (2004). The evolution of human skin and skin color. *Annu. Rev. Anthropol., Annual Reviews*, 33(1):585–623.

Lima Vazquez, V. d. e. a. (2015). Melanoma characteristics in brazil: demographics, treatment, and survival analysis. *BMC Research Notes*.

Lin, B., Xu, Y., Bao, X., Zhao, Z., Wang, Z., and Yin, J. (2025). Skingen: an explainable dermatology diagnosis-to-generation framework with interactive vision-language models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1287–1296, New York, NY, USA. Association for Computing Machinery.

Souza, E. P. d., Gomes, C. M., Barroso, D. H., Miranda, V. L. d., and Gurgel-Gonçalves, R. (2019). Aplicações do deep learning para diagnóstico de doenças e identificação de insetos vetores. *Saúde em Debate*, 43(spe2):147–154.

Sumithra, R., Suhil, M., and Guru, D. (2015). Segmentation and classification of skin lesions for disease diagnosis. *Procedia Computer Science*, 45:76–85.

Zink, B. S. (2014). Câncer de pele: a importância do seu diagnóstico, tratamento e prevenção. *Revista Hospital Universitário Pedro Ernesto*, 13.