



Universidade Estadual de Campinas  
Instituto de Computação - IC

***Visions Transforms e geração de imagens sintéticas  
com GPT-4 na classificação de doenças em folhas de  
mandioca***

Suellen Sena da Silva (177261)

Prof. Dr. Christian Esteve Rothenberg

Campinas  
2023

# **Conteúdo**

<b>1</b>	<b>Introdução</b>	<b>2</b>
<b>2</b>	<b>Metodologia</b>	<b>2</b>
<b>3</b>	<b>Resultados</b>	<b>5</b>
<b>4</b>	<b>Conclusão</b>	<b>9</b>

# 1 Introdução

A mandioca, conhecida cientificamente como *Manihot esculenta*, é uma cultura vital em muitas partes do mundo, especialmente em regiões tropicais, onde serve como uma importante fonte de carboidratos na dieta humana [Nassar e Ortiz 2007]. No entanto, a produtividade e a qualidade da mandioca são frequentemente comprometidas por várias doenças bacterianas e vírais, como mancha bacteriana (*bacterial blight*), doença das estrias marrons (*brown streak disease*) e doença do mosaico (*mosaic disease*) [Casinga et al. 2021]. A identificação e o diagnóstico precoces dessas doenças são cruciais para implementar medidas de controle e minimizar perdas significativas de colheita e consequentemente, na economia mundial [Parmar, Sturm e Hensel 2017].

A mancha bacteriana, causada pela bactéria *Xanthomonas axonopodis* *pv. manihotis*, é notória por suas lesões nas folhas, tornando-as murchas e podendo levar à morte da planta. Essa doença prospera em condições de alta umidade e temperatura e pode ser disseminada de diversas formas, incluindo ferramentas agrícolas e insetos. Por outro lado, a doença das estrias marrons, é provocada por um vírus transmitido por moscas-brancas. Caracterizada por manchas marrons nos tubérculos e folhas, a doença resulta em perdas consideráveis, principalmente nos tubérculos. Já a doença do mosaico, também causada por vírus, manifesta-se através de padrões de mosaico nas folhas, comprometendo a fotossíntese e o crescimento geral da planta.

Este projeto propõe um método inovador utilizando *Vision Transformers* (*ViTs*) para a identificação e diagnóstico de doenças em folhas de mandioca. A técnica emergente no campo do aprendizado profundo e processamento de imagens, oferece uma abordagem promissora devido à sua capacidade de capturar características complexas e padrões em imagens. Ao contrário dos métodos convencionais baseados em CNN (*Convolutional Neural Networks*), os ViTs trabalham com uma série de transformações que permitem uma melhor generalização e interpretação das características visuais das folhas de mandioca [Khan et al. 2022], tornando-os particularmente adequados para identificar padrões sutis associados a infecções específicas.

Além disso, a geração de imagens sintéticas surge como uma necessidade neste projeto, dada a limitação de dados reais disponíveis para treinamento para determinados tipos de doenças que acometem a planta. Por isso, propomos a utilização de técnicas avançadas de geração de imagens para criar representações artificiais, que auxiliarão a complementar o conjunto de dados existente e serão utilizadas para teste [Goodfellow et al. 2014]. Dessa forma, é possível avaliar duas diferentes técnicas em inteligência artificial para resolver problemas reais.

# 2 Metodologia

As ViTs representam uma mudança importante no campo do processamento de imagens, diferenciando-se significativamente dos tradicionais métodos de aprendizagem profundo. A arquitetura *transformers*, composta por duas partes principais, o decodificador e o codificador, representa um avanço significativo no campo do processamento de linguagem natural e, mais recentemente, no processamento de imagens. Dessa forma, lidando com tradução de máquina, o codificador recebe dados de entrada, como uma frase, e produz uma representação

intermediária. Já o decodificador, de forma inovadora, decodifica essa representação passo a passo para gerar a saída. Entender a seção do codificador é fundamental para os ViTs.

Inicialmente, os dados de entrada são incorporados em um vetor. A camada de incorporação ajuda a obter uma representação vetorial aprendida para cada palavra. Em seguida, uma codificação posicional é injetada nos *embeddings* de entrada, pois o *transformer* não comprehende a ordem da sequência que está sendo passada como entrada.

O componente de atenção multi-cabeças é onde as coisas diferem. Ele consiste em três vetores aprendíveis: *Query*, *Key* e *Value*. A motivação para isso vem da recuperação de informações, onde você pesquisa (*query*) e o mecanismo de busca compara sua *query* com uma chave e responde com um valor. Os vetores *Q* e *K* passam por uma multiplicação de matriz produto escalar para produzir uma matriz de pontuação, que indica o quanto uma palavra deve atender a cada outra palavra.

A matriz de pontuação é então escalonada de acordo com as dimensões dos vetores *Q* e *K* para garantir gradientes mais estáveis. Em seguida, essa matriz é processada por uma função *softmax* para transformar as pontuações de atenção em probabilidades. Depois, a matriz resultante com probabilidades é multiplicada com o vetor de valor.

Por fim, os vetores de saída *QK* e *V* são concatenados e alimentados em uma camada linear para processamento adicional. A auto-atenção é realizada para cada palavra na sequência, e os vetores de valor de saída são concatenados e adicionados à conexão residual vinda da camada de entrada. Em seguida, a representação resultante é passada para uma *LayerNorm* para normalização.

Além disso, a saída passa por uma rede *feed-forward* pontual para obter uma representação ainda mais rica, e os resultados são novamente normalizados por *LayerNorm* com resíduos adicionados da camada anterior.

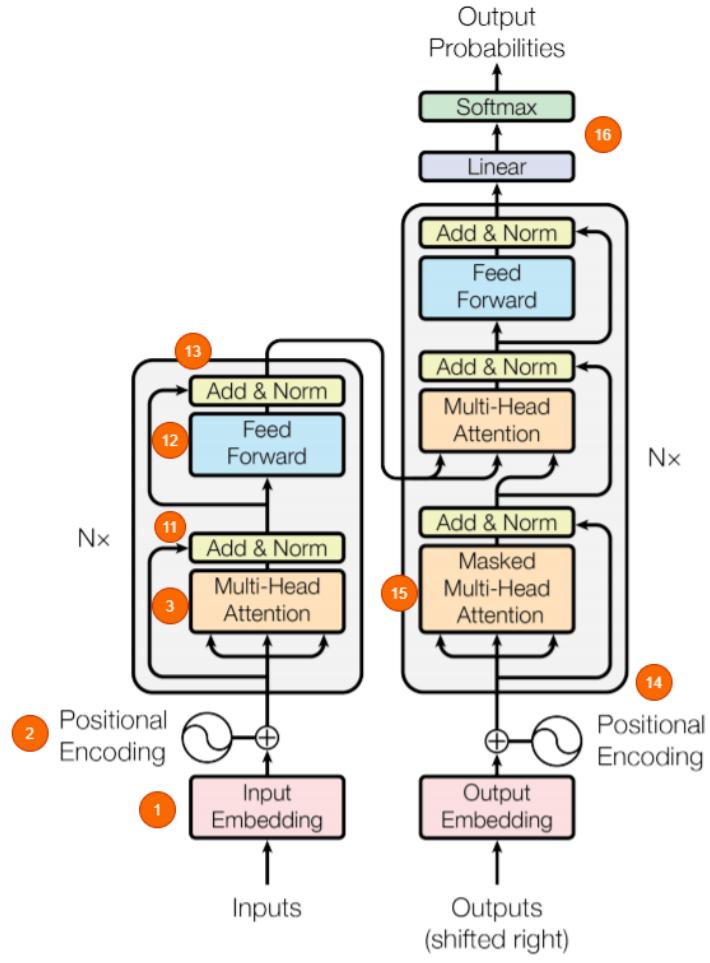


Figura 1: Arquitetura de um *transformer*.

Embora aplicar *transformers* em imagens sempre pareceu ser um desafio, pois ao contrário de palavras ou frases, imagens contêm muito mais informações na forma de pixels, e seria extremamente difícil, mesmo com hardware atual, atender a cada pixel da imagem, os pesquisadores do *Google* propuseram uma abordagem diferente [Dosovitskiy et al. 2020], que pode ser o próximo grande passo na visão computacional, principalmente por discutir que a dependência de CNNs pode não ser mais necessária.

Nos ViTs, apenas a parte do codificador do *transformer* é utilizada, mas a diferença está em como as imagens são alimentadas na rede. As imagens são divididas em *patches* de tamanho fixo, que são então desenrolados (achatados) e enviados para processamento adicional na rede. Cada *patch* da imagem é primeiro desenrolado em um grande vetor e multiplicado com uma matriz de *embedding*, também aprendível. Esses *patches* incorporados são combinados com o vetor de *embedding* posicional e alimentados no *transformers*. Daqui em diante, tudo é igual a um *transformer* padrão, com a única diferença sendo que, em vez de um decodificador, a saída do codificador é passada diretamente para uma rede neural *feed-forward* para obter a saída de classificação.

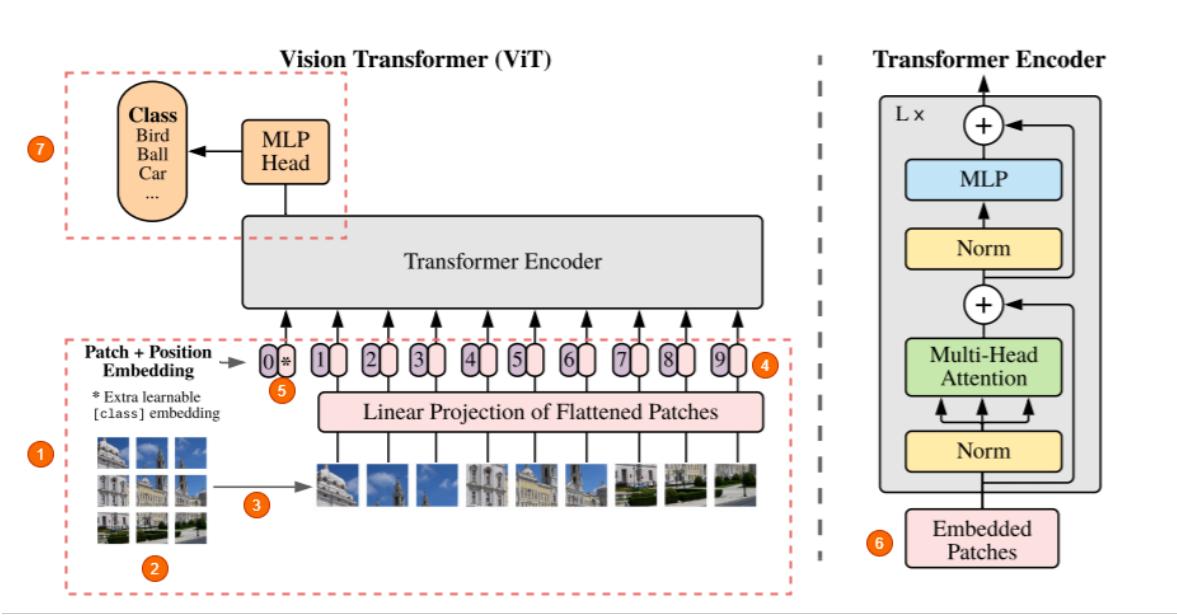
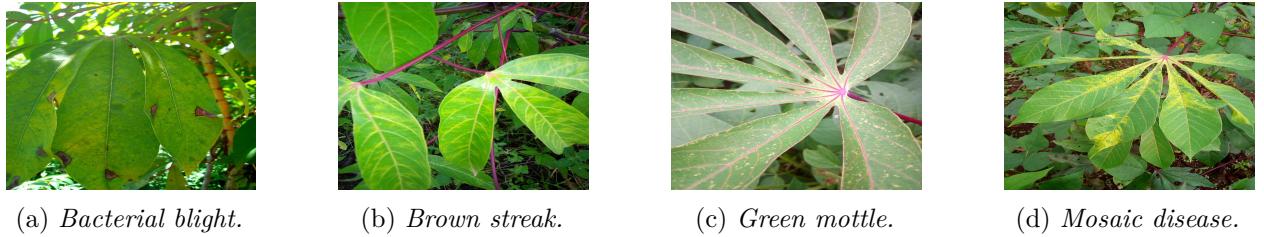


Figura 2: Arquitetura ViT.

### 3 Resultados

Para realizar esse projeto, foi explorado um extenso banco de imagens disponível no *Kaggle*, composto por aproximadamente 21.400 fotografias reais, como mostrado na Figura 3, para o treinamento de modelos de aprendizado profundo na identificação de doenças em folhas de mandioca. Dessas imagens, 20% foram selecionadas para compor o conjunto de dados de validação.



(a) *Bacterial blight*. (b) *Brown streak*. (c) *Green mottle*. (d) *Mosaic disease*.

Figura 3: Imagens do banco *Kaggle*.

Contudo, devido à presença de algumas imagens corrompidas, o número final de imagens efetivamente utilizadas para o treinamento foi reduzido para 18.850. É importante mencionar também que o conjunto de dados apresenta um desequilíbrio em algumas classes de doenças, conforme ilustrado na Figura 4. Adicionalmente, todos os códigos de desenvolvimento e imagens geradas usados para o treinamento dos modelos estão disponíveis no repositório [Cassava leaf disease classification](#), proporcionando transparência e facilitando a reprodução ou a extensão de nosso trabalho por outros pesquisadores e desenvolvedores interessados na área.

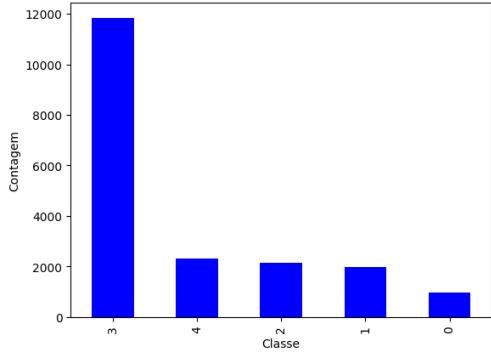


Figura 4: Distribuição de imagens por classe.

Além de utilizar os conjuntos de dados disponíveis no *Kaggle*, foi empregado o *ChatGPT-4* como uma ferramenta adicional para receber descrições detalhadas das características específicas de cada doença que afeta a mandioca. Estas descrições foram usadas como *prompts* para instruir o *GPT-4* a gerar imagens o mais realistas possível. O objetivo dessa abordagem era avaliar a capacidade dos ViTs de generalizar a partir dessas imagens e testar a eficácia do *DALL-E* na produção de visuais de alta qualidade. Portanto, solicitou-se ruídos nas imagens geradas pelo *DALL-E*, que incluem a adição de elementos como mãos, plantas saudáveis aparecendo de fundo, variações na iluminação entre as imagens e outros detalhes que podem ocorrer em um ambiente real. Para cada classe, foram geradas cinco imagens diferentes, cada uma com suas próprias características únicas. Isso ajudou a tornar o conjunto de dados mais diversificado e a testar a capacidade do modelo de reconhecer doenças em situações diversas e complexas.



(a) *Bacterial blight*. (b) *Brown streak*. (c) *Green mottle*. (d) *Mosaic disease*.

Figura 5: Imagens artificiais geradas pelo *DALL-E*.

Embora seja evidente que as imagens artificiais geradas ainda não reproduzem perfeitamente as características da realidade, é importante reconhecer o avanço significativo nas pesquisas envolvendo imagens criadas por inteligência artificial generativa. Esse campo está se mostrando extremamente valioso, especialmente em situações onde os dados disponíveis são escassos, como no caso de doenças raras. O desenvolvimento contínuo dessas tecnologias não apenas aprimora a qualidade e a veracidade das imagens geradas, mas também abre novas possibilidades para o estudo e compreensão de condições menos comuns, contribuindo assim para avanços significativos em áreas que tradicionalmente sofrem com a limitação de dados.

Os resultados obtidos com a rede ViT foram notavelmente positivos, destacando a eficácia deste modelo em nosso projeto. A rede atingiu aproximadamente 96% de acurácia no conjunto de treinamento e 85% no conjunto de validação, e isso com apenas 50 épocas de treinamento, conforme a Figura 6.

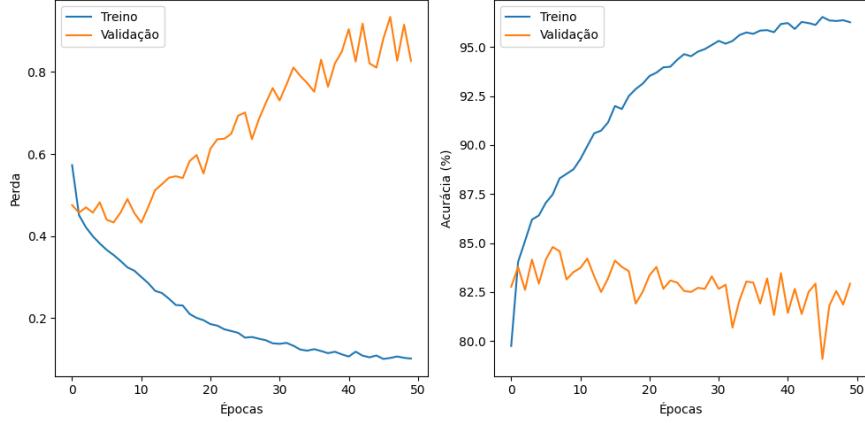


Figura 6: Perda e acurácia para os conjuntos de treino e validação. Arquitetura ViT.

No entanto, ao observar a matriz de confusão apresentada na Figura 7, verifica-se que para determinadas classes a rede não conseguiu realizar a discriminação com eficiência, principalmente para classe 0 (*bacterial blight*). Para as demais classes, o resultado foi superior, destacando-se a classe 3 (*green mottle*), que consequentemente representava a maior parte do banco de dados, tal fato auxilia na generalização de aprendizado da rede.

Ao analisar a matriz de confusão apresentada na Figura 7, pode-se observar que a rede teve dificuldades em discriminar eficazmente algumas classes, especialmente a classe 0 (*bacterial blight*). No entanto, para as demais classes, o desempenho foi melhor, com destaque para a classe 3 (*green mottle*), que representava a maior parte do conjunto de dados. Isso sugere que a predominância da classe 3 no conjunto de dados pode ter contribuído para um aprendizado mais eficaz por parte da rede neural, enquanto outras classes apresentaram desafios adicionais de discriminação.



Figura 7: Matriz de confusão para o conjunto de validação.

Finalmente, a rede *baseline* foi avaliada no conjunto de teste, composto exclusivamente por imagens geradas artificialmente. É importante notar que essas imagens são animadas e ainda não alcançam o nível de realismo absoluto, uma vez que não houve treinamento do modelo considerando imagens sintéticas, apenas imagens reais. Dadas essas considerações, os resultados não atenderam plenamente às expectativas, conforme evidenciado na Figura 8. Ainda, não foram obtidos acertos para as classes 0, 1 e 2, sugerindo que a rede está descalibrada e tende a classificar a maioria das imagens como pertencentes à classe 3. Por outro lado, para a classe 4, que consiste em imagens de folhas saudáveis em diferentes contextos, o modelo obteve um desempenho superior, o que pode ser atribuído à maior facilidade de distinção nessa classe.

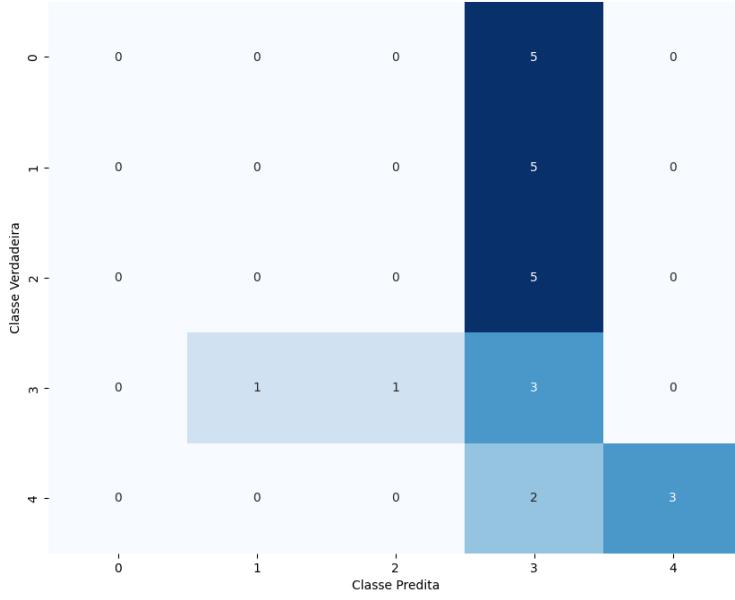


Figura 8: Matriz de confusão para o conjunto de imagens sintéticas.

## 4 Conclusão

Os resultados apresentados na seção anterior são significativos, considerando a diversidade e a complexidade das imagens usadas no treinamento. Muitas dessas imagens apresentam características bastante distintas e ruídos externos que fazem parte do ambiente não controlado da imagem. Portanto, é evidente a capacidade do ViT de capturar detalhes sutis e generalizar informações com alta precisão, reforçando sua adequação para tarefas complexas de classificação de imagens e destacando seu potencial no campo da visão computacional.

No entanto, ao utilizar imagens artificiais, os resultados não atingiram o nível esperado. Isso porque, o fato dos testes serem baseados em uma rede *baseline* dificulta sua capacidade de generalização. Além disso, não foram incorporadas imagens artificiais no treinamento, o que poderia auxiliar a rede a lidar com problemas específicos dessa natureza. Vale ressaltar que a utilização de imagens sintéticas como um método de *data augmentation*, especialmente para as classes minoritárias, poderia ter sido uma estratégia benéfica para melhorar o desempenho da rede em cenários desafiadores [Shorten e Khoshgoftaar 2019].

Por fim, é importante destacar que, mesmo com a qualidade ótima das imagens, o fato de elas não representarem imagens reais pode ter contribuído negativamente, levando em consideração o poder de atenção aos detalhes das ViTs. Contudo, este trabalho tem como objetivo fomentar discussões sobre a qualidade da geração de imagens sintéticas e o uso de redes originalmente projetadas para processar texto, que agora estão apresentando resultados promissores em tarefas de processamento de imagens.

## Referências

- [Casinga et al. 2021]CASINGA, C. M. et al. Expansion of the cassava brown streak disease epidemic in eastern democratic republic of congo. *Plant Disease*, Am Phytopath Society, v. 105, n. 8, p. 2177–2188, 2021.
- [Dosovitskiy et al. 2020]DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [Goodfellow et al. 2014]GOODFELLOW, I. et al. Generative adversarial nets. *Advances in neural information processing systems*, v. 27, 2014.
- [Khan et al. 2022]KHAN, S. et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022.
- [Nassar e Ortiz 2007]NASSAR, N.; ORTIZ, R. Cassava improvement: challenges and impacts. *The Journal of Agricultural Science*, Cambridge University Press, v. 145, n. 2, p. 163–171, 2007.
- [Parmar, Sturm e Hensel 2017]PARMAR, A.; STURM, B.; HENSEL, O. Crops that feed the world: Production and improvement of cassava for food, feed, and industrial uses. *Food Security*, Springer, v. 9, p. 907–927, 2017.
- [Shorten e Khoshgoftaar 2019]SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.